Amélioration et développement de méthodes de sélection du nombre de composantes et de prédicteurs significatifs pour une régression PLS et certaines de ses extensions à l'aide du bootstrap

Jérémy Magnanensi

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTÉ

EA3430, Progression tumorale et microenvironnement. Approches translationnelles et épidémiologie.

Institut de Recherche Mathématique Avancée, UMR 7501, Labex IRMIA

THÈSE présentée par :

Jérémy MAGNANENSI

soutenue le : **18 Décembre 2015**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Biostatistique

Amélioration et développement de méthodes de sélection du nombre de composantes et de prédicteurs significatifs pour une régression PLS et certaines de ses extensions à l'aide du bootstrap.

THÈSE dirigée par :

Pr MEYER Nicolas	PUPH, Faculté de Médecine, EA3430, université de Strasbourg

RAPPORTEURS :

Pr SABATIER Robert	PU, Faculté de Pharmacie, université de Montpellier
Dr HANAFI Mohamed	IR, ONIRIS Nantes

AUTRES MEMBRES DU JURY :

Dr BERTRAND Frédéric	MCU, IRMA, UMR 7501, université de Strasbourg
Dr GUENOT Dominique	DR, EA3430, université de Strasbourg
Pr VINZI Vincenzo Esposito	PU, ESSEC Business School, Paris

Jury

Le jury présidant à la soutenance de cette thèse est composé des personnes suivantes :

- **Pr. Nicolas Meyer**, directeur de thèse
- **Dr. Frédéric Bertrand**, co-encadrant de thèse et examinateur
- **Pr. Robert Sabatier**, rapporteur externe
- **Dr. Mohamed Hanafi**, rapporteur externe
- **D.R. Dominique Guenot**, examinatrice interne
- **Pr. Vincenzo Esposito Vinzi**, examinateur externe

Ce jury est complété par la présence, en qualité de membre invité, de la personne suivante :

- **Dr. Myriam Maumy-Bertrand**, membre invité

Liste des participations à divers événements

Dans le cadre de cette thèse, j'ai pu assister et participer à plusieurs événements.

- GDR Statistique et Santé - 24 et 25 Juin 2013 - Paris
- Deuxièmes Rencontres de R - 27 et 28 Juin 2013 - Lyon
- 8$^{\text{th}}$ *International Conference on Partial Least Squares and Related Methods* - 26 au 28 Mai 2014 - Paris

  Communication sous forme d'un poster intitulé « *A new bootstrap based stopping criterion in PLS components construction* ».
- Journée Labex IRMIA - 5 Juin 2014 - Strasbourg

  Communication orale autour de la détermination du nombre de composantes en PLS.
- 10$^{\text{ème}}$ Semaine Études Mathématiques Entreprises - 23 au 27 Juin 2014 - Strasbourg

  Communication orale sur la problématique « Analyse et classification de courbes de charge électrique ».
- 8$^{\text{th}}$ *International Conference on Computational and Methodological Statistics* - 12 au 14 Décembre 2015 - Londres, UK

  Communication sous forme d'un poster, en cours d'élaboration.

Remerciements

L'écriture des remerciements est le moment qui, à travers la recherche des personnes ayant participé de près ou de loin à l'élaboration de ce travail, de quelque sorte que ce soit, me permet l'espace d'un instant de me remémorer les événements ayant constitué ces trois dernières années de ma vie. Cela me permet notamment de réaliser à quel point l'élaboration et l'aboutissement de ce travail leur est en grande partie dû. Chaque personne que je m'apprête à remercier ici m'a, à sa façon, apporté une aide précieuse si ce n'est essentielle. Je tiens d'avance à m'excuser auprès de toutes les personnes qui, lorsque je les reverrai, me feront réaliser leur absence au sein de ces quelques lignes. Sachez que ces oublis ne représentent en aucun cas un acte volontaire, mais bien la preuve qu'une recherche ponctuelle dans ses souvenirs ne saurait être d'une exhaustivité totale, bien que ces souvenirs soient, et resteront, inoubliables. Je vous remercie ainsi d'avance, vous les « oubliés »de ces quelques lignes, d'avoir été présents, d'avoir été simplement vous-mêmes.

Je tiens tout d'abord à remercier mon directeur de thèse, le Professeur Nicolas Meyer, pour m'avoir permis de vivre cette aventure qu'est le doctorat. Provenant de milieux différents, il aura, dès le départ, su me faire confiance afin de mener à bien ce travail. Le fait que nous ne nous connaissions pas avant n'a jamais été un frein, bien au contraire. Ses conseils et remarques ont toujours été d'une réelle pertinence et m'ont constamment permis d'améliorer ma production. Travailler sous sa direction aura été un plaisir et, encore une fois, merci d'avoir permis à ce travail de thèse de voir le jour.

En poursuivant avec les personnes m'ayant fait confiance depuis le début, et même avant, je ne peux que grandement remercier le Docteur Frédéric Bertrand, mon co-encadrant de thèse. Il aura lui aussi permis l'élaboration de ce projet, en créant le lien avec mon directeur de thèse et en surmontant les difficultés de départ, afin d'aboutir au lancement de mon doctorat. Tout au long de ces trois années, son calme, sa qualité pédagogique, ses indications et conseils mathématiques égrainés tout au long de réunions extrêmement intéressantes et enrichissantes auront eu un impact significatif sur ce travail. Merci de m'avoir permis de travailler dans de telles conditions, d'avoir toujours tenu compte des mes idées, de m'avoir constamment motivé à les développer, d'avoir été attentif, compréhensif et d'avoir tenu compte des différents éléments de ma vie, notamment personnelle, qui se sont déroulés durant ces trois années. Cette ouverture d'esprit et cette confiance mutuelle auront été l'un des points forts de ce travail.

Je tiens également à remercier Myriam Maumy-Bertrand pour son implication dans ces

travaux. Ses conseils et son sens de l'organisation sans faille m'auront grandement aidé à structurer le temps dont je disposais afin de mener à bien cette entreprise. Son écoute, sa protection et son soutien, tout au long de ces trois années, représentent pour moi des éléments remarquables, éléments qui m'auront permis de vivre au mieux le développement de ma thèse. Un grand merci pour tout cela.

Je tiens ensuite à remercier l'ensemble de l'équipe d'accueil EA3430 et plus particulièrement le Docteur Dominique Guenot, directrice de l'équipe et examinatrice interne de cette thèse, pour son intérêt envers mon travail ainsi que ses conseils et avis biologiques sur les résultats que j'ai obtenu. Intégrer cette équipe essentiellement constituée de biologistes et de médecins m'a permis d'entrevoir un univers que je ne connaissais pas avant, renforçant d'autant plus mon intérêt pour cette notion d'interdisciplinarité.

Je remercie naturellement les autres membres de mon jury de thèse, le Professeur Robert Sabatier et le Docteur Mohamed Hanafi, tous deux rapporteurs externes, ainsi que le Professeur Vincenzo Esposito Vinzi, examinateur externe, d'avoir bien voulu consacrer de leur temps que je sais précieux afin de juger ce travail. Je les remercie également pour l'intérêt qu'ils ont porté à mon sujet de thèse ainsi que pour leurs futures remarques et suggestions, dont je sais d'avance qu'elles seront constructives et pertinentes.

Je tiens également à remercier l'ensemble du personnel de l'UFR de Mathématique de Strasbourg et de l'IRMA pour leur aide précieuse. Je pense tout particulièrement à Alexis Palaticky, Alain Sartout et Alexandre Ancel pour leurs aides précieuses en informatique.

Je tiens également à remercier tout particulièrement le Professeur Thomas Delzant, responsable du Labex IRMIA, de m'avoir fait confiance en acceptant de consacrer une partie de ce Labex pour le financement de ma thèse.

Je tiens ensuite à remercier mes deux collègues doctorant avec qui j'ai partagé mon bureau, Amaury et Florian. Merci à eux d'avoir égaillé mes journées, d'avoir su intégrer des moments plus légers, dont je tairai ici les contenus, mais qui participent, d'après moi, à l'équilibre nécessaire au travail de thèse.

Je remercie également l'ensemble des doctorants de l'UFR de mathématiques pour leurs sollicitations quand était venu le temps d'aller se vider l'esprit autour d'un ou plusieurs verres mais aussi pour leurs conseils mathématiques et aides pratiques qui m'auront grandement aidé tout au long de ces trois années. Je remercie ainsi tout particulièrement Stéphane, Simon, Ranine, Mohamad, Guillaume et Vincent, notamment pour ses récits internationaux passionnant.

Enfin et pour terminer avec le milieu professionnel, je tenais à remercier Nicolas, post-doctorant en Statistique, pour ses conseils et son aide, notamment lors de la rédaction de ma thèse. Un grand merci également à Théo, lui aussi post-doctorant en Statistique, mais également ami de longue date, pour son soutien et son aide notamment à travers ses indications

sur la langue anglaise lors de la rédaction de mes articles.

Il est maintenant temps de remercier l'ensemble des personnes qui n'ont pas forcément partagé mon quotidien mais qui, à travers leurs présences et leurs soutiens, m'ont permis de mener à bien cette expérience.

Mes premières pensées reviennent plus que jamais à mes parents. D'une certaine façon, la finalisation de ce doctorat représente l'aboutissement de ma vie scolaire et étudiante et s'il y a bien deux personnes à qui je dois l'ensemble de cette première partie de ma vie, ce sont eux. Il me parait donc inconcevable de ne pas leur rendre hommage, ce travail étant, en quelque sorte, au moins autant le leur que le mien. Ma mère m'a apporté et inculqué toute l'humilité, à mes yeux, nécessaire à la réussite. Elle m'a transmis la notion de respect à avoir envers les personnes nous entourant, afin d'apprendre au mieux de leurs expériences et de leur savoir. Elle m'a sensibilisé depuis mon plus jeune âge à la notion d'ouverture d'esprit, notion menant à la curiosité nécessaire pour continuer à avancer, apprendre, s'interroger et finalement résoudre les questionnements d'une vie. Mon père, dans le rôle qui lui incombe et qu'il a rempli avec succès à mes yeux, m'a inculqué la notion de l'ambition. Il m'a ainsi appris à ne jamais renoncer, à toujours croire en soi, à être persuadé que tout est possible à celui qui s'en donne les moyens. Il représente à ce titre un exemple pour moi. Lui s'est chargé de m'apporter le nécessaire pour me permettre d'aller au bout de mes ambitions. Enfin, tous deux ont su me transmettre la valeur du travail. Ils m'ont constamment et indirectement fait réaliser la chance qu'était la mienne de pouvoir étudier, voyager, rencontrer, partager et je leur dois tout cela. Finalement, peu importe le lieu, le moment, le contexte, j'ai pu constamment ressentir leur présence, leur amour et leur soutien sans faille dans les bons comme mauvais passages de ma vie. Je pense et j'espère les avoir rendu fiers, fiers du travail exemplaire que eux effectuent depuis maintenant 28 ans. Maman, Papa, un merci ne suffira pas mais c'est le moins que je puisse faire pour le moment, alors merci pour tout.

Mes pensés vont désormais vers deux autres personnes que je ne pourrai jamais suffisamment remercier pour tout ce qu'ils ont fait pour moi.

La première, ma grande sœur, a toujours été présente pour moi. Elle a su me conseiller, me rassurer et me réconforter quand j'en ai eu besoin. Elle a aussi parfaitement su remplir son rôle quand il s'agissait de me remettre sur le « droit »chemin si je puis dire ainsi. Elle est pour moi quelqu'un chez qui j'ai pu me « réfugier »sans avoir peur d'un quelconque jugement, à chaque fois que j'en ai ressentis le besoin. Sans elle, sans sa présence, sans ses conseils, ce travail n'aurait certainement jamais vu le jour. Nat, je n'en dirais pas plus, tu sais le reste, alors simplement, merci d'être la grande sœur dont tout petit frère rêve, merci pour tout.

La deuxième personne, mon grand frère, a également toujours été un exemple pour moi. Sa joie de vivre, son enthousiasme et sa générosité font de lui une personne unique. Lui aussi a toujours été présent quand j'en ai eu besoin et ses conseils ont, pour moi, toujours été pertinents et essentiels. Sa curiosité, son perfectionnisme et sa capacité à ne retenir que les bons côtés m'ont fortement influencé et impressionné. D'une manière différente mais équivalente, sans lui, je n'en serai certainement pas là où j'en suis aujourd'hui. Merci pour tout Pat.

Le moment est maintenant venu de remercier l'ensemble de mes amis. Merci à vous pour tous ces moments de détente, ces sorties, ces voyages, ces instants de ma vie que vous avez rendu uniques par votre présence. Vous m'avez permis de me vider la tête et l'esprit durant ces trois années et plus encore, m'apportant ainsi le courage et l'enthousiasme nécessaire à cette réussite. Merci aussi à vous, ceux qui ont pris le temps de m'écouter, de me réconforter, de m'aider et de me conseiller dans mes moments de doute. Je n'oublierai jamais tout cela. J'ai une pensé toute particulière pour Joe et Claire, merci tout simplement d'être qui vous êtes et d'avoir toujours été là. Un grand merci à Michel et Nico, votre amitié et votre hospitalité pour m'accueillir sur vos canapés respectifs ont été des points déterminants dans la réussite de ce doctorat. Merci à vous Fanny et Zahir pour votre joie de vivre lors de nos répétitions et soirées. Un grand merci à mes anciens collègues de la faculté de mathématique, je pense tout particulièrement à Thibaut, Thomas et Cyril. Un grand merci à Sharzad pour le temps qu'elle a bien voulu prendre afin de me fournir ses indications précieuses sur le domaine de la génétique. Finalement, merci à tous, Dju, Axel, Jean, Thomas, Anne-So, Malenka, Julie, Yank, Nelly, Vincent, Aimie, Olivier, Roxane, Symon et tous les autres, vous m'avez tous, d'une façon ou d'une autre, permis d'accomplir tout cela.

Je finirai par remercier la personne qui partage ma vie depuis cinq ans maintenant, et qui m'a donc accompagné tout au long de cette expérience. Et forcé d'admettre que c'est en pensant à elle que je me rends le plus compte du chemin parcouru durant ces trois dernières années. Tu m'as fais grandir sur beaucoup d'aspects, tu m'as appris énormément de choses, à tes côtés la vie m'a paru plus facile, plus belle, plus intéressante. Ta soif de découverte, de nouveaux défis, ton envie de réaliser chacun de tes rêves m'a toujours impressionné et motivé. Si tout n'a pas toujours été facile entre nous durant ces trois dernières années, à tes côtés rien ne me parait impossible. Merci pour tous les magnifiques moments que l'on a partagé, merci pour ton soutien, merci de m'avoir permis de m'évader du quotidien quand j'en ai eu besoin et de m'avoir fait rêver aussi souvent. Ce travail te revient en grande partie, merci Marion.

Table des matières

Première partie

Introduction

Chapitre 1

Problématiques biologiques et statistiques

Le fort développement des sciences informatiques ainsi que des technologies de stockage de données a permis une évolution significative des méthodes et procédés dans de multiples domaines. Les milieux médicaux et biologiques ont notamment considérablement évolué à travers le développement de techniques sophistiquées, permettant la récolte de très grandes quantités de données (Schena *et al.*, 1995). L'étude du génome humain est un exemple de référence et est devenu un domaine de recherche extrêmement dynamique, que ce soit par l'étude de mutations géniques ou encore par la capture et l'analyse de l'expression génique.

1.1 Le cas de l'analyse génomique

« La génomique est la science qui étudie la structure, le contenu et l'évolution des génomes. L'objectif premier était le séquençage en masse des séquences nucléotidiques. L'évolution technologique aidant, le domaine d'étude s'est élargi pour ne plus se limiter à la détermination des séquences mais y inclure également l'analyse de l'expression et de la fonction aussi bien des gènes que des protéines. Parallèlement, la bio-informatique et les études biologiques assistées par ordinateur sont de plus en plus souvent intégrées à ces analyses. »(Gibson and Muse, 2004, p.1)

1.1.1 Historique

L'étude de la génomique est devenue un domaine de recherche en forte expansion à la fin du 20ème siècle. Ainsi, une entreprise scientifique majeure financée sur fond publique a vu le jour en 1993, le Projet Génome Humain (PGH) (Collins and Galas, 1993). Ce projet sera considéré comme une réussite tant les innovations qui ont suivi ont été nombreuses et fondamentales, entraînant de ce fait la génomique à se développer en tant que discipline à part entière (Burris *et al.*, 1998). Ce projet entraînera la publication officielle d'une première ébauche de séquences du génome humain en 2001 (Venter *et al.*, 2001). Hormis le séquençage du génome à proprement parler, une deuxième partie de la génomique a été développée. Cette seconde branche consiste en l'analyse du transcriptome fournissant ainsi des données

sur l'expression des gènes à l'échelle d'un génome entier. Dans la pratique, cela a donné naissance à la technologie des puces à ADN ou *microarrays*.

1.1.2 Les *microarrays*

L'idée de la puce ADN n'est pas aussi récente que le PGH puisque l'idée et la méthodologie associée ont été introduites par Chang (1983). Il aura cependant fallu attendre l'article de Schena *et al.* (1995) proposant une méthode simple permettant de suivre simultanément le niveau relatif d'expression de milliers de gènes, pour observer un développement significatif de cette technique. Ainsi, plusieurs technologies pour l'analyse simultanée de l'expression de milliers de gènes ont vu le jour depuis. Citons-en deux particulièrement important, les micro-alignements d'ADN complémentaire (ADNc) et les micro-alignements d'oligonucléotides. Deux techniques, liées à deux types de puce, permettent globalement leur analyse, la *CGH-array* (*Comparative Genomic Hybridization-array*) et la *SNP-array* (*Single Nucleotide Polymorphism-array*).

La technique de la *CGH* a été initialement mise au point par Kallioniemi *et al.* (1992). Il s'agit d'une méthode de cytogénétique permettant d'analyser les variations du nombre de copies des gènes dans l'ADN entre deux individus, l'un étant l'ADN du patient ciblé, couplé à un fluorochorme rouge (rhodamine) et l'autre l'ADN du sujet contrôle couplé à un fluorochorme vert (fluoresceine).

L'ADN du patient ciblé est mélangé avec l'ADN contrôle et hybridé sur des lames sur lesquelles sont étalés des chromosomes en métaphase, étape du cycle cellulaire qui permet de bien identifier chaque paire de chromosomes. L'ADN marqué en vert et celui marqué en rouge s'hybrident de façon compétitive avec l'ADN en métaphase :

- si, dans le tissu issu du patient ciblé, le nombre de copies d'un gène est augmenté, la sonde tumorale s'hybridera préférentiellement et le segment de chromosome apparaitra en rouge.
- si le nombre de copies est diminué, la sonde de tissu sain s'hybridera préférentiellement et le segment de chromosome apparaitra en vert.
- si le nombre de copies n'est pas modifié, les deux sondes s'hybrideront en quantité équivalente et le segment de chromosome apparaitra en orange.

La fluorescence est ensuite détectée par un microscope à épifluorescence et quantifiée par analyse d'image quantitative.

La résolution, c'est à dire la marge d'erreur, de telles techniques sur bande chromosomique reste de 10 à 15 Mb (Millions de paires de bases) voir de 3 à 5 Mb dans les meilleures conditions et selon les régions du génome (Andrieux, 2008). Solinas-Toldo *et al.* (1997) et Pinkel *et al.* (1998) proposent une adaptation de la technique d'hybridation génomique, la puce de *CGH* comparative (*CGH array*). Cette nouvelle technique est basée sur l'adaptation de la méthodologie à des lames de verre constituées de nombreux puits et dans lesquels sont fixés les fragments d'ADN génomique humain connus (sonde). Il s'agit de la *CGH* sur réseau d'ADN ou *CGH-array*. La *CGH-array* permet ainsi une approche hautement résolutive, *i.e.*

de 1Mb à quelques kilobases, puisque la taille des sondes est de l'ordre de 100 à 200 kb voir moins suivant le type de puces à ADN utilisé (Andrieux, 2008). Les intensités de fluorescence sont alors captées par un scanner laser puis traitées numériquement, retournant ainsi une base de données contenant le rapport d'intensité de fluorescence entre les deux ADN pour chaque puits. Ce processus est résumé sur la Figure 1.1 extraite du livre de Gibson and Muse (2004).

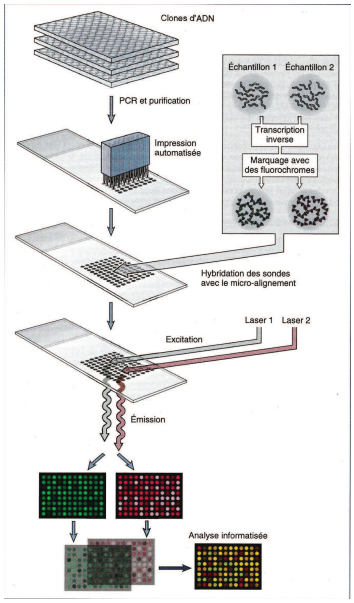


FIGURE 1.1 – Processus de fonctionnement et d'analyse d'une CGH-array.

Précisons enfin que la CGH-array peut être liée à des micro-alignements d'ADNc, comme décrit ci-dessus, ou à du micro-alignement d'oligonucléotides.

La *SNP* array est une technique liée à du micro-alignement d'oligonucléotides et se différencie de la *CGH-array* par la comparaison de l'ADN du patient à une référence informatique faisant ainsi office d'unité d'hybridation. Dans ce cadre de micro-alignement d'oligonucléotides, chaque gène est découpé en plusieurs probe sets, que l'on peut définir comme de courts fragments d'ADN constitués d'une série d'oligonucléotides de séquences différentes, correspondant chacun à une région spécifique du gène en question. Un probe set est alors créé en modifiant un seul nucléotide au milieu de la séquence de sorte à créer une séquence quasi-similaire. En effet, dû à la petite taille des probe sets, ceci est réalisé afin de contrôler les possibilités d'hybridation croisée avec des petites séquences similaires dans des transcrits autres que celui que l'on veut tester ; le probe set modifié ne devant pas s'hybrider en toute rigueur. La moyenne de la différence de signal entre celui lié au probe set original et celui lié au probe set modifié est alors retourné comme niveau d'expression liée.

1.1.3 L'allélotypage

L'allélotypage est une technique issue de la biochimie moléculaire, consistant en la recherche d'anomalies ciblant les microsatellites. Les microsatellites sont des zones particulières

de l'ADN, définis comme une répétition continue de motifs composés de 1 à 6 nucléotides d'où leur nom de *Simple Tandem Repeat* (*STR*) ou encore *Simple Sequence Repeat* (*SSR*) (Li *et al.*, 2002b). La longueur de ces séquences, c'est-à-dire le nombre de répétitions, varie selon l'espèce, l'individu mais également d'un allèle à l'autre chez un même individu, voire d'une cellule à l'autre du fait d'erreurs au cours de la réplication de l'ADN. Ainsi, les deux allèles d'un microsatellite se différencient non pas par la séquence des bases mais bien par leur nombre de répétitions.

Ces séquences n'ont en soi pas de fonction particulière s'agissant de régions non codantes. Cependant, leur répartition sur l'ensemble du génome et surtout à proximité de gènes, permet de connaître le statut normal ou altéré des gènes localisés à proximité de ces microsatellites (Naidoo and Chetty, 1998). A l'heure actuelle, la séquence de la quasi-totalité des microsatellites est connue et les gènes à proximité, identifiés. Par conséquent, connaître le statut d'un microsatellite permet de connaitre le statut d'un ou des gènes proches.

Les cellules cancéreuses étant liées à des remaniements chromosomiques importants, l'utilisation et l'analyse des microsatellites sur l'ensemble du génome permettent donc de cartographier simultanément ces altérations. Les données que nous avons pu avoir à traiter proviennent donc de cette application de l'analyse des déséquilibres alléliques des microsatellites.

Plus concrètement, une *Polymerase Chain Reaction* (*PCR*), technique de biologie moléculaire développée par Mullis *et al.* (1986), est effectuée en premier lieu. La *PCR* est une technique d'amplification *in vitro* d'une séquence d'ADN qui est effectuée en procédant à des répétitions de cycles de température précise permettant successivement de dénaturer l'ADN en le séparant en deux brins avant d'hybrider les amorces introduites avec les simples brins. Ainsi, à partir d'une très faible quantité d'ADN et d'amorces spécifiques constituées d'oligonucléotides de synthèse, on peut obtenir une grande quantité de matériel génétique exploitable en répétant un certain nombre de fois l'opération. Une fois le matériel génétique obtenu, les fragments amplifiés sont analysés par électrophorèse. Le principe est alors le suivant : on effectue ce procédé sur de l'ADN supposé sain et sur l'ADN à tester provenant du tissu cancéreux. Lors de la migration des amplifiats d'ADN, quatre pics apparaissent, deux liés à l'ADN supposé sain et servant de référence, et deux liés au tissu cancéreux. Les hauteurs des deux pics représentent la quantité (proportionnelle) d'amplifiats liée à chacun des deux allèles du microsatellite. En effet, puisque le nombre de répétitions de nucléotides des deux allèles du microsatellite n'est pas identique (sauf dans le cas de microsatellite homozygote), cette réaction de polymérisation en chaîne va permettre d'amplifier deux fragments de taille différente, chaque pic correspondant à un allèle. Pour chaque ADN, les rapports entre les hauteurs des deux pics sont déterminés. Le rapport lié au tissu sain est normalement compris entre 0,8 et 1. Si les deux rapports sont à peu près égaux, il n'y a pas de déséquilibre allélique. En revanche, si le rapport lié au tissu cancéreux est nettement inférieur à 0,8, on peut conclure à une perte ou à un gain de segment chromosomique (Skotheim *et al.*, 2001). Les biologistes fixent alors un seuil de binarisation en dessous duquel le déséquilibre sera considéré comme étant vérifié à un certain risque près. Cette binarisation, dans le cadre d'une régression logistique sur des petits échantillons, va permettre de renforcer la robustesse du modèle (Tufféry, 2010, p.84). Les données résultant de telles études et formant notre matrice de prédicteurs $\mathbf{X}$ sont donc binaires.

1.2 Propriétés structurelles des bases de données

Nous venons d'aborder de façon pratique et relativement sommaire les mécanismes et procédés liés aux puces à ADN et à l'allélotypage. Ces avancées technologiques et scientifiques, bien que fondamentales, se heurtent très souvent à une capacité restreinte d'analyse statistique des résultats, diminuant significativement la portée et les conséquences de telles avancées.

« Alors que l'analyse par micro-alignements a de fortes chances de contribuer à la compréhension d'autres maladies [...], les progrès nécessiteront des procédures biostatistiques plus sophistiquées pour détecter de faibles signaux dont la variation d'expression est importante parce qu'il s'agit, soit de petits changements d'expression de régulateurs clés, soit de changements importants ne se produisant que dans une petite proportion de cellules dans un échantillon de tissu complexe. »(Gibson and Muse, 2004, p.173)

En effet, les bases de données issues de telles études sont liées à des caractéristiques dont il faut tenir compte lors de leur traitement statistique. Afin de détailler quelque peu ces caractéristiques ainsi que leurs conséquences, commençons par introduire quelques notations.

Soient $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ le vecteur réponse observé, avec $(.)^T$ représentant la transposée. Soit $\mathbf{X} = (\mathbf{x}_1, \ldots \mathbf{x}_p) \in \mathcal{M}_{n,p}(\mathbb{R})$ la matrice des p régresseurs.

Une première caractéristique repose sur les dimensions de la matrice $\mathbf{X}$. En effet, nous nous retrouvons rapidement confronté à la problématique de la grande dimension, à savoir un cadre où $p \gg n$. Or, lorsque $p > n$, nous obtenons :

$$\begin{aligned} &\mathrm{Rg}\left(\mathbf{X}^T\mathbf{X}\right) = \mathrm{Rg}\left(\mathbf{X}\right) \leqslant n < p, \textit{ avec } \mathbf{X}^T\mathbf{X} \in \mathcal{M}_p(\mathbb{R}) \\ \Rightarrow &det\left(\mathbf{X}^T\mathbf{X}\right) = 0 \end{aligned}$$

Dans ce cas, la matrice carrée $\mathbf{X}^T\mathbf{X} \in \mathcal{M}_p(\mathbb{R})$ n'étant pas de rang plein, l'utilisation de procédés de modélisation usuels telle que la régression linéaire couplée à l'estimation des paramètres par Moindres Carrés Ordinaires (MCO) n'est plus envisageable. En effet, la matrice $\mathbf{X}^T\mathbf{X}$ n'étant pas inversible, l'utilisation d'un inverse généralisé de cette matrice noté $\left(\mathbf{X}^T\mathbf{X}\right)^-$ entraîne le fait qu'il n'y ait plus unicité des estimations $\widehat{\beta}^{MCO} = \left(\mathbf{X}^T\mathbf{X}\right)^- \mathbf{X}^T\mathbf{y}$ (Hastie *et al.*, 2009, p.46).

Cette situation est bien connue et d'une importance croissante, spécialement dans le domaine de la génomique (Hastie *et al.*, 2009, ch.18). Cette problématique, dans le cadre de ce travail de thèse, est essentiellement liée aux bases de données issues de puces à ADN. En effet, que ce soit pour des raisons logistiques, humaines, financières ou éthiques, le nombre de sujets n pouvant être inclus dans une étude est souvent limité. Cependant, les techniques de séquençage génique liées aux puces à ADN et décrites dans la partie 1.1.2 permettent de récolter des données sur plusieurs milliers de probe sets simultanément, retournant ainsi des bases de données avec un nombre de prédicteurs bien supérieur au nombre de sujets inclus dans les études. Dans le cas d'études d'allélotypage, la matrice $\mathbf{X}$ obtenue peut également être liée à ce type de caractéristique bien que la différence entre le nombre de régresseurs et le nombre de sujets soit globalement moins importante que dans le cas d'études menées sur

une puce à ADN.

Une seconde caractéristique réside dans les fortes corrélations (en valeur absolue) présentes entre certains prédicteurs. Dans le cas des puces à ADN, une première raison est naturellement liée au problème de la grande dimension des bases de données, caractéristique que nous venons de détailler. En effet, comme indiqué à l'instant, le nombre de prédicteurs est supérieur au nombre de sujet. Ainsi, p vecteurs dans un espace de dimension n avec $n < p$ implique mathématiquement l'existence de corrélations non-nulles entre certains prédicteurs. Cependant, même en admettant que nous puissions inclure un nombre suffisant de sujets de telle sorte que $n > p$, il existe des raisons biologiques à la corrélation entre probe sets. En effet, s'agissant d'expressions géniques, il est biologiquement connu que le niveau d'expression d'un gène peut moduler les niveaux d'expression d'autres gènes. Cette régulation des gènes entre eux induit un phénomène de « *clustering* ». Il est par ailleurs connu que les gènes desdits clusters, qui présentent donc une forte structure de corrélation, partagent des fonctions biologiques. La modélisation de tels réseaux de gènes est ainsi devenu un domaine de recherche à part entière (Basso *et al.*, 2005).

Mathématiquement, les cas de fortes corrélations linéaires entre les prédicteurs sont problématiques. En effet, la présence de liens forts entre les régresseurs implique directement des valeurs propres liées à $\mathbf{X}^T\mathbf{X}$ proches de zéro. La matrice de variance-covariance de l'estimateur des paramètres β^{MCO} dépendant de $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ (Hastie *et al.*, 2009, p.47), on se retrouve alors en face d'importants problèmes de conditionnement liés à la forte variabilité de ces estimateurs, entraînant des incohérences liées aux signes des estimations et l'absence de significativité des régresseurs (Tenenhaus, 1998, p.79).

Enfin une dernière caractéristique est observée dans le cadre de l'allélotypage, celle des données manquantes. En effet, dans le cas de microsatellites dits homozygotes, les deux allèles possèdent le même nombre de répétitions de la séquence de base. Dans ce cas, un seul pic est visible après l'amplification par PCR, ce pic représentant alors une superposition des deux pics originels (1.1.3). Ainsi, si dans le tissu considéré il y a une perte ou un gain de la portion chromosomique étudiée, la PCR ne retournant qu'un seul pic, il ne sera pas possible pour l'observateur de déterminer ce déséquilibre. Dans ce cas précis de microsatellites homozygotes, les sites sont dit *non-informatifs* et sont représentés par des valeurs manquantes dans la base de données à traiter. Précisons que la disparition éventuelle des deux allèles ne peut pas être observée. Le pourcentage moyen de données manquantes dans ce type de bases de données étant de l'ordre de 30%, il s'agit d'une caractéristique importante à prendre en compte. Des détails à ce sujet sont disponibles notamment dans les travaux de Skotheim *et al.* (2001), Tomlinson *et al.* (2002) et Panhard *et al.* (2003). Durant ce travail de thèse, nous n'avons pas spécifiquement étudié ce problème. Cependant, il nous a paru essentiel d'utiliser un procédé capable de gérer cet aspect là.

1.3 Analyse statistique des données

1.3.1 Objectifs et motivations biologiques

L'étude des expressions géniques, ou des anomalies alléliques dans le cadre de l'allélotypage, a pour but, entre autres, de déterminer et spécifier l'existence de liens entre ces données et certaines caractéristiques d'une pathologie (probabilité de rechute, classification, localisation, analyse de la survie...). Ainsi, un premier objectif consiste en le développement d'une modélisation fiable de tels liens afin de pouvoir décrire et/ou prédire la caractéristique étudiée.

Un second objectif consiste à proposer ou sélectionner des prédicteurs qui soient liés aux caractéristiques cliniques de la pathologie, telles que les récidives ou le décès par exemple. En effet, déterminer un ou plusieurs prédicteurs (probe sets ou microsatellites suivant la méthodologie) comme étant spécifiques d'une caractéristique d'une pathologie est d'une importance capitale pour dépister, détecter ou traiter efficacement la pathologie concernée ou encore pour en comprendre la physiopathologie. Sur le plan biologique, l'implication d'un ou plusieurs oncogènes, la perte d'un ou plusieurs suppresseurs de tumeurs, potentiellement corrélés à un taux de survie, ou à un risque de rechute, permet de proposer de nouveaux mécanismes de la progression tumorale. Ainsi, par exemple, pouvoir définir quel microsatellite ou quel ensemble de microsatellites est associé à la localisation préférentielle de la tumeur dans le côlon droit ou gauche ou à la récidive devrait permettre de proposer des séquences/associations d'altérations moléculaires associées à l'initiation et/ou à l'évolution de la pathologie cancéreuse.

1.3.2 Les régressions sur variables latentes

Afin de répondre aux objectifs biologiques qui nous sont donnés et que nous avons présentés dans la partie précédente (1.3.1), l'utilisation d'outils statistiques spécifiques est nécessaire. Les régressions sur variables latentes représentent un outil particulièrement bien adapté afin de traiter ce type de problème. En effet, afin de palier aux problèmes inhérents aux caractéristiques des bases de données que nous avons à traiter (1.2), des variables latentes, plus communément appelées *composantes*, sont créées en tant que combinaisons linéaires des prédicteurs d'origine. Ces composantes sont déterminées de façon à être non-corrélées et leur nombre est limité au rang de la matrice $\mathbf{X}$. Il est alors possible d'effectuer des techniques de régressions usuelles de $\mathbf{y}$ sur ces variable latentes, les considérant ainsi comme prédicteurs.

La première procédure utilisant cette méthodologie a été proposée par Hotelling (1957) et est nommée régression sur Composantes Principales (CP). Tout d'abord il s'agit d'effectuer une Analyse en Composantes Principales (ACP) sur la matrice $\mathbf{X}$. Rappelons que dans le cadre de l'ACP, l'objectif est de créer de nouvelles composantes $\mathbf{T}^{ACP} = \left(\mathbf{t}_1^{ACP}, \ldots, \mathbf{t}_K^{ACP}\right)$, $K \leqslant \mathrm{Rg}\left(\mathbf{X}\right)$, de variance maximale afin de représenter au mieux les données contenues dans la matrice $\mathbf{X}$. Pour ce faire, une diagonalisation de la matrice de variance-covariance $\mathbf{X}^T\mathbf{X}$ est effectuée, les vecteurs propres $(\mathbf{u}_1, \ldots, \mathbf{u}_K)$ associés aux valeurs propres $\phi_1 \geqslant \ldots \geqslant \phi_K$ les plus élevées sont alors utilisés afin de construire les composantes de la façon suivante :

$$\mathbf{t}_j^{ACP} = \mathbf{X}\mathbf{u}_j,\ j = 1, \ldots, K \tag{1.1}$$

Il s'agit ensuite d'effectuer une régression par MCO de $\mathbf{y}$ sur les composantes issues de cette ACP. Cette méthode permet ainsi un gain de stabilité lors des estimations des paramètres, les composantes étant orthogonales deux à deux et de variance maximale. Cependant, n'utiliser que les composantes de plus grande variance n'est pas nécessairement un choix optimal comme le remarque Jolliffe (1982). Quoi qu'il en soit, une limite de cette méthode est qu'elle ne tient pas compte de la réponse $\mathbf{y}$ lors de la création des variables latentes, maximisant uniquement la qualité de représentation de $\mathbf{X}$ dans le nouvel espace de dimension K ainsi créé.

Notre choix de méthodologie s'est porté sur une seconde technique de régression sur variables latentes, la régression des Moindres Carrées Partiels ou *Partial Least Squares* (PLS). Ce choix est motivé par plusieurs raisons.

Tout d'abord, au même titre que pour la régression sur CP, le problème de la grande dimension est résolu à travers la construction d'un nombre de composantes inférieur ou égal au rang de la matrice $\mathbf{X}$, composantes qui vont être utilisées comme régresseurs. Le problème de la colinéarité est ainsi lui aussi traité par la régression PLS (Wold *et al.*, 1984).

Ensuite, la régression PLS utilise des composantes maximisant non pas leur variance, comme c'est le cas pour la régression sur CP, mais leur covariance avec la réponse $\mathbf{y}$. Cette prise en compte de la réponse est essentielle. Ainsi, la régression PLS peut-être vu comme un compromis entre la régression par MCO et la régression sur CP. En effet, Höskuldsson (1992) a montré que la régression PLS représente un compromis idéal entre qualité d'ajustement, propriété de la régression par MCO, et stabilité des composantes obtenus, propriété liée à la régression sur CP.

Un argument supplémentaire en faveur de la régression PLS a été démontré par De Jong (1993a). Pour un nombre fixe de composantes, l'estimation $\hat{\mathbf{y}}^{PLS}$ produite par la régression PLS est liée à une erreur d'approximation plus faible que celle produite par la régression sur CP :

$$\forall k \in [\![1, \mathrm{Rg}(\mathbf{X})]\!], \ \left\|\mathbf{y} - \hat{\mathbf{y}}_k^{PLS}\right\|^2 \leqslant \left\|\mathbf{y} - \hat{\mathbf{y}}_k^{CP}\right\|^2 \tag{1.2}$$

avec $\hat{\mathbf{y}}_k^{PLS}$ et $\hat{\mathbf{y}}_k^{CP}$ représentant respectivement l'estimation de la réponse obtenue à l'aide du modèle de régression PLS et sur CP à k composantes.

Comme observé par Frank and Friedman (1993) et Krämer and Sugiyama (2011), les régressions PLS et sur CP possèdent des performances prédictives comparables. Cependant, dû à l'optimisation du procédé de construction des composantes PLS, celle-ci a besoin de moins de composantes que la régression sur CP pour atteindre des performances similaires. Il s'agit d'un élément important en faveur la régression PLS, notamment pour les problèmes où ces variables latentes peuvent être utilisées pour une interprétation concrète.

Enfin, la régression PLS est en mesure de gérer le problème des données manquantes, sans avoir recours à une imputation préalable. Nous reviendrons sur ce point dans le prochain chapitre.

Il est donc remarquable que cette technique de régression soit totalement adaptée au traitement des bases de données issues d'études génomiques, en tenant compte de l'intégralité de leurs caractéristiques problématiques que nous avons présenté dans la partie 1.2. La régression PLS, notamment à travers l'ensemble de ces arguments, est ainsi devenue un procédé

de modélisation de référence pour le traitement statistique de telles études (Boulesteix and Strimmer, 2007).

Nous allons donc consacrer le prochain chapitre de ce travail de thèse à la régression PLS afin d'en détailler les aspects qui nous paraissent essentiels.

Chapitre 2

La régression PLS

La régression PLS est un procédé de modélisation datant des années 1980, qui a pour origine le développement de l'algorithme NIPALS (Nonlinear estimation by Iterative PArtial Least Squares), introduit par Wold *et al.* (1966), pour l'analyse en composantes principales.

Son adaptation à l'approche PLS, approche introduite par Wold (1975), est à l'origine de l'écriture de la régression PLS sous sa forme algorithmique la plus connue. Il nous parait donc important de débuter ce chapitre en effectuant quelques rappels au sujet de cet algorithme NIPALS.

2.1 L'algorithme NIPALS

L'un des intérêts de l'algorithme NIPALS réside en sa capacité de traiter des jeux de données avec valeurs manquantes sans devoir procéder à une imputation préliminaire ou supprimer des lignes entières. Afin d'exposer de façon conventionnelle cet algorithme, nous allons suivre la présentation faite par Wold *et al.* (1987) et reprise par Tenenhaus (1998, ch.6). Posons $\mathbf{X} = \{x_{ij}\}$ un tableau individus×variables de rang K avec données manquantes. Notons $\mathbf{x}_1, \ldots, \mathbf{x}_p$ les colonnes de $\mathbf{X}$ que l'on suppose centrées.

Son écriture et son interprétation étant plus facile, commençons par écrire l'algorithme lorsqu'il n'y pas de données manquantes :

1. $\mathbf{X_0} = \mathbf{X}$
2. Pour $k = 1, \ldots, K$:
   - 2.1. Posons, par exemple, $\mathbf{t}_k$ comme étant le vecteur colonne de variance maximale dans $\mathbf{X}_{k-1}$.
   - 2.2. Répéter jusqu'à convergence de $\mathbf{p}_k$:
     - 2.2.1. $\mathbf{p}_k = \mathbf{X}_{k-1}^T \mathbf{t}_k / \mathbf{t}_k^T \mathbf{t}_k$
     - 2.2.2. Normer $\mathbf{p}_k$ à 1
     - 2.2.3. $\mathbf{t}_k = \mathbf{X}_{k-1} \mathbf{p}_k / \mathbf{p}_k^T \mathbf{p}_k$
   - 2.3. $\mathbf{X}_k = \mathbf{X}_{k-1} - \mathbf{t}_k \mathbf{p}_k^T$

Cette écriture algorithmique nous permet de relever un point fort essentiel, la rapidité de calcul. En effet, chaque coordonnée p_{kj} d'un vecteur $\mathbf{p}_k$ correspond au coefficient de régression de $\mathbf{t}_k$ sur $\mathbf{x}_{k-1,j}$. De la même façon, chaque coordonnée t_{kj} d'un vecteur $\mathbf{t}_k$ correspond au coefficient de régression de $\mathbf{p}_k$ sur la $j^{\text{ème}}$ ligne de $\mathbf{X}_{k-1}$. De plus, en remarquant que ces étapes correspondent à l'application du théorème de la puissance itérée pour le calcul du vecteur propre lié à la plus grande valeur propre, on en déduit que $\mathbf{p}_k$ sera le vecteur propre associé à la plus grande valeur propre de $\mathbf{X}_{k-1}^T\mathbf{X}_{k-1}$ et $\mathbf{t}_k$ celui lié à la même plus grande valeur propre de $\mathbf{X}_{k-1}\mathbf{X}_{k-1}^T$, ce qui peut asymptotiquement s'écrire de la façon suivante (Tenenhaus, 1998, p.63) :

$$\mathbf{X}_{k-1}^T\mathbf{X}_{k-1}\mathbf{p}_k = \psi_k\mathbf{p}_k \tag{2.1}$$
$$\mathbf{X}_{k-1}\mathbf{X}_{k-1}^T\mathbf{t}_k = \psi_k\mathbf{t}_k \tag{2.2}$$

On remarque donc que sans donnée manquante, cet algorithme produit bien les mêmes axes que ceux obtenu par l'ACP. Ainsi, en reprenant les notations de la partie 1.3.2 on obtient :

$$\mathbf{t}_k = \mathbf{t}_k^{ACP}$$
$$\mathbf{p}_k = \mathbf{u}_k$$

La différence entre l'ACP et l'algorithme NIPALS réside donc dans le fait qu'aucune diagonalisation matricielle n'a été réalisée directement. Ceci qui peut être intéressant en terme de temps de calcul sur des très grandes bases de données puisque, dans le cas de cet algorithme, seules des projections orthogonales sur un axe sont effectuées.

Comme énoncé au début de cette partie, un second intérêt fondamental de l'algorithme NIPALS réside dans sa capacité à traiter des jeux de données avec données manquantes. Pour ce faire, l'algorithme est adapté de la façon suivante :

1. $\mathbf{X_0} = \mathbf{X}$
2. Pour $k = 1, \ldots, K$:
   - 2.1. Posons, par exemple, $\mathbf{t}_k$ comme étant le vecteur colonne de variance maximale dans $\mathbf{X}_{k-1}$.
   - 2.2. Répéter jusqu'à convergence de $\mathbf{p}_k$:
     - 2.2.1. Pour $j = 1, \ldots, p$:

       $$p_{kj} = \frac{\sum_{i \in I} x_{k-1,ij} t_{ki}}{\sum_{i \in I} t_{ki}^2}$$

       avec $I = \{i : x_{ij} \text{ et } t_{ki} \text{ existent}\}$
     - 2.2.2. Normer $\mathbf{p}_k$ à 1

2.2.3. Pour $i = 1, \ldots, n$:

$$t_{ki} = \frac{\sum_{j \in J} x_{k-1,ij} p_{kj}}{\sum_{j \in J} p_{kj}^2}$$

avec $J = \{j : x_{ij} \text{ existe}\}$

2.3. $\mathbf{X}_k = \mathbf{X}_{k-1} - \mathbf{t}_k \mathbf{p}_k^T$

En calculant aux étapes 2.2.1 et 2.2.3, indépendamment pour chaque colonne et chaque ligne, les pentes des droites des moindres carrées sur les données disponibles, l'algorithme permet ainsi de traiter des jeux de données avec données manquantes sans imputation préalable. Cette méthode permet ainsi de retourner une estimation des valeurs manquantes de la façon suivante :

$$\hat{x}_{ij} = \sum_{k=1}^{K} t_{ki} p_{kj} \tag{2.3}$$

Précisons enfin que la convergence de l'algorithme est assurée lorsque la proportion de données manquantes n'est pas trop élevée, *i.e.* de l'ordre de 20 à 30% (Dray *et al.*, 2003), ce qui correspond au pourcentage moyen de manquants dans les bases de données issues d'études d'allélotypage 1.2. Cependant, à notre connaissance, aucune étude précise quant à la proportion limite de manquants pour assurer la convergence n'a été effectuée. Notons tout de même que cette proportion limite doit être dépendante de la structure et du type de ces données manquantes.

2.2 La régression PLS, généralités

2.2.1 Historique

L'algorithme PLS a été développé et proposé initialement par Wold *et al.* (1983) en application à une problématique multivariée dans le domaine de la chimie. Ces résultats, obtenus et publiés par les chimistes Svante Wold et Harald Martens, sont dans la continuité des travaux initiés par le père du premier cité, le statisticien Herman Wold. Tous deux publieront quelques années plus tard des mémoires personnelles sur l'historique du développement de la régression PLS à travers deux articles, Wold (2001) et Martens (2001).

A travers les travaux de ces deux chimistes qui, au sein de leur communauté, publieront et exposeront les intérêts de cette procédure (Lindberg *et al.*, 1983), la régression PLS va devenir une procédure habituelle en chimiométrie tel qu'en atteste une publication de Wold *et al.* (2001). Très rapidement, le succès de cette méthode va ainsi attirer grandement l'intérêt des statisticiens. On peut citer entre autres, pour les premiers, Höskuldsson (1988) et Frank and Friedman (1993). La recherche fondamentale portant sur les aspects mathématiques et statistiques de la régression PLS reste actuellement un secteur de recherche dynamique. L'étude des liens entre régression PLS et espaces de Krylov (Phatak and de Hoog, 2002) ou encore avec les polynômes orthogonaux (Blazere *et al.*, 2014) en sont deux exemples. D'autres

propriétés et résultats liés aux estimateurs ont également été obtenus notamment dans les travaux de Goutis *et al.* (1996), Phatak *et al.* (2002) ou encore De Jong (1995).

Au vue de l'intérêt de la communauté scientifique pour cette méthode, celle ci s'est rapidement propagée à d'autres secteurs d'activités tel que la bioinformatique, la médecine, la biologie ou encore la sociologie (Rosipal and Krämer, 2006). Afin d'illustrer cet intérêt grandissant au sein de la communauté scientifique pour la régression PLS, nous reprenons un graphique produit par Mehmood *et al.* (2012) représentant l'importante évolution du nombre de publications en lien avec la régression PLS au cours du temps (Fig. 2.1).

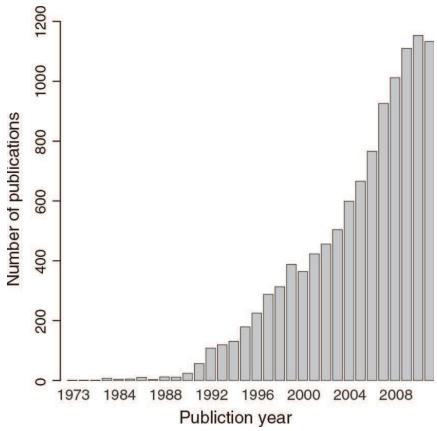


FIGURE 2.1 – Évolution du nombre de publications en lien avec la régression PLS.

Passons désormais à la présentation de cet algorithme de régression PLS.

2.2.2 L'algorithme PLS

Cadre de travail

Plusieurs variantes de l'algorithme PLS ont été développées dans la littérature, la plus connue étant celle basée sur l'algorithme NIPALS, algorithme que nous venons de présenter. Une autre variante, nommée Statistically Inspired Modification of PLS, donnant naissance à l'acronyme SIMPLS et développée par De Jong (1993b), s'est également imposée comme un algorithme de référence.

Cependant, tout au long de ce travail de thèse, nous nous sommes limités à l'étude de cas univariés. La régression PLS ainsi appliquée à $\mathbf{y} \in \mathbb{R}^n$ est mieux connue sous l'acronyme PLS1. Or il a été montré que dans ce cas précis, l'approche SIMPLS était équivalente à l'approche NIPALS (De Jong, 1993b). Nous allons donc uniquement nous concentrer sur cette dernière.

Nous nous plaçons dans l'espace des variables aléatoires réelles de carré intégrable au sens de Lebesgue noté L^2 :

$$L^2 = \left\{\text{v.a. } X \text{ tel que } \mathbb{E}\left(X^2\right) < \infty\right\} \tag{2.4}$$

Muni du produit scalaire $\langle X, Y\rangle = \mathbb{E}\left(XY\right)$, c'est un espace de Hilbert d'après le Théorème de Riesz-Fischer (Komornik, 2002, p.198). Ainsi, d'après le théorème de projection orthogonale, nous sommes assuré pour tout $x \in L^2$ et tout sous-espace vectoriel fermé de L^2 noté

K, de l'existence et l'unicité d'un point $y = P_K(x)$ à distance minimale de x. Ce point est caractérisé par la propriété suivante (Komornik, 2002, p.8) :

$$x - y \perp K \tag{2.5}$$

Ce résultat permet donc de garantir les propriétés d'orthogonalité liées à la méthode des MCO et sur lesquelles est basée la régression PLS. Se placer dans cet espace de Hilbert nous permet également de disposer d'autres résultats liés à ce type d'espace, notamment les inégalités de Cauchy-Schwartz (Komornik, 2002, p.5) ou de Minkowski (Komornik, 2002, p.198).

Finalement, précisons que tout au long de ce travail nous supposerons $\mathbf{y} \in \mathbb{R}^n$ et $\mathbf{X} \in \mathcal{M}_{n,p}(\mathbb{R})$ centrés-réduits.

Construction de l'algorithme

La régression PLS est caractérisée par la construction d'une structure d'espace latente, matérialisé par la formation de variables latentes $\mathbf{t}_k$, plus communément appelées *composantes* et mises sous forme d'une matrice $\mathbf{T}_K = (\mathbf{t}_1, \ldots, \mathbf{t}_K)$ avec $K \leqslant \mathrm{Rg}(\mathbf{X})$ (Wold *et al.*, 1983). Ces composantes sont définies comme combinaisons linéaires des variables d'origine, au même titre que les composantes obtenues en effectuant une ACP, et peuvent ainsi être exprimées de la façon suivante (Tenenhaus, 1998, p.107) :

$$\forall k \in [\![1; K]\!],\ \mathbf{t}_k = \mathbf{X}\mathbf{w}_k^* \tag{2.6}$$

Nous reviendrons un peu plus tard sur l'obtention des poids $\mathbf{w}_k^*,\ k \in [\![1; K]\!]$.

Dans le cadre d'une ACP, comme évoqué dans la partie 1.3.2, les composantes sont obtenues dans le but de maximiser leurs variances. Dans le cadre d'une régression PLS, les composantes associées sont quant à elles obtenues en tant que solution d'un problème d'optimisation sous contrainte, consistant à maximiser le carré de la covariance entre la dite composante et le vecteur réponse (Höskuldsson, 1988, p.218). Cela permet ainsi de tenir compte de la réponse associée. Formalisons quelque peu tout cela.

Notons $\mathbf{X}_k$ la matrice résiduelle de la régression linéaire multiple de $\mathbf{X}$ sur $\mathbf{T}_k$ et $\mathbf{y}_k$ le vecteur résidus de la régression linéaire simple de $\mathbf{y}$ sur $\mathbf{T}_k$. On pose $\mathbf{y}_0 = \mathbf{y}$ et $\mathbf{X}_0 = \mathbf{X}$. On définit alors successivement les composantes PLS de la façon suivante :

$$\forall k \in [\![1; K]\!],\ \mathbf{t}_k = \mathbf{X}_{k-1}\mathbf{w}_k \tag{2.7}$$

avec $\mathbf{w}_k$ solution du problème d'optimisation sous contrainte suivant (Boulesteix and Strimmer, 2007, p.3) :

$$\mathbf{w}_k = \underset{\mathbf{w} \in \mathbb{R}^p}{\mathrm{argmax}} \left\{ \mathrm{Cov}^2(\mathbf{y}_{k-1}, \mathbf{t}_k) \right\},\ \text{s.c. } \|\mathbf{w}_k\|_2^2 = 1 \tag{2.8}$$

$$= \underset{\mathbf{w} \in \mathbb{R}^p}{\mathrm{argmax}} \left\{ \mathbf{w}^T \mathbf{X}_{k-1}^T \mathbf{y}_{k-1} \mathbf{y}_{k-1}^T \mathbf{X}_{k-1} \mathbf{w} \right\},\ \text{s.c. } \|\mathbf{w}_k\|_2^2 = 1 \tag{2.9}$$

Utilisons la méthode des multiplicateurs de Lagrange :

Soit $\mathbf{L} = \mathbf{y}_{k-1}^T \mathbf{X}_{k-1} = (l_1, \ldots, l_p) \in \mathbf{R}^{1\times p}$. Posons :

$$J(\mathbf{w}) = -\mathbf{w}^T \mathbf{X}_{k-1}^T \mathbf{y}_{k-1} \mathbf{y}_{k-1}^T \mathbf{X}_{k-1} \mathbf{w} \tag{2.10}$$
$$= -\|\mathbf{w}\|_{\mathbf{L}}^2 \tag{2.11}$$

et

$$F(\mathbf{w}) = \mathbf{w}^T \mathbf{w} - 1 \tag{2.12}$$
$$= \|\mathbf{w}\|_2^2 - 1 \tag{2.13}$$

On obtient alors, pour $\mathbf{H} \in \mathbb{R}^p$:

$$\begin{cases} \partial J_{\mathbf{w}}(\mathbf{H}) = -2\langle \mathbf{w}, \mathbf{H} \rangle_{\mathbf{L}} = -2\mathbf{w}^T \mathbf{L}^T \mathbf{L} \mathbf{H} \\ \partial F_{\mathbf{w}}(\mathbf{H}) = 2\langle \mathbf{w}, \mathbf{H} \rangle = 2\mathbf{w}^T \mathbf{H} \end{cases} \tag{2.14}$$

Soit λ_1 le multiplicateur de Lagrange. On résout le système suivant :

$$\Rightarrow \begin{cases} -2l_1 \left(\sum_{j=1}^{p} w_j l_j \right) + 2\lambda_1 w_1 = 0 \\ \vdots \\ -2l_p \left(\sum_{j=1}^{p} w_j l_j \right) + 2\lambda_1 w_p = 0 \\ \mathbf{w}^T \mathbf{w} = 1 \end{cases}$$

$$\Rightarrow \begin{cases} l_1 \left(\sum_{j=1}^{p} w_j l_j \right) = \lambda_1 w_1 \\ \vdots \\ l_p \left(\sum_{j=1}^{p} w_j l_j \right) = \lambda_1 w_p \\ \sum_{j=1}^{p} w_j^2 = 1 \end{cases}$$

$$\Rightarrow \lambda_1^2 = \left(\sum_{j=1}^{p} l_j^2 \right) \times \left(\sum_{j=1}^{p} w_j l_j \right)^2$$

$$\Rightarrow \lambda_1 = \sqrt{\sum_{j=1}^{p} l_j^2} \times \sum_{j=1}^{p} w_j l_j = \|\mathbf{L}\|_2 \times \sum_{j=1}^{p} w_j l_j$$

Ce calcul entraîne donc le résultat suivant :

$$\mathbf{w}_k = \frac{\mathbf{X}_{k-1}^T \mathbf{y}_{k-1}}{\left\| \mathbf{X}_{k-1}^T \mathbf{y}_{k-1} \right\|} = \frac{1}{\sqrt{\sum_{j=1}^{p} \mathrm{Cov}^2(\mathbf{x}_{k-1,\,j}, \mathbf{y}_{k-1})}} \cdot \begin{pmatrix} \mathrm{Cov}(\mathbf{x}_{k-1,\,1}, \mathbf{y}_{k-1}) \\ \ldots \\ \mathrm{Cov}(\mathbf{x}_{k-1,\,p}, \mathbf{y}_{k-1}) \end{pmatrix} \tag{2.15}$$

L'algorithme PLS

Cette formulation des poids va ainsi permettre le développement de l'algorithme PLS. Celui-ci va intégrer, à travers quelques étapes sur lesquelles nous reviendrons, la prise en compte des données manquantes en s'inspirant de l'algorithme NIPALS. Cet algorithme PLS s'écrit donc de la façon suivante (Tenenhaus, 1998, p.99) :

1. Posons $\mathbf{X}_0 = \mathbf{X}$, $\mathbf{y}_0 = \mathbf{y}$ et $K \leqslant \mathrm{Rg}(\mathbf{X})$ le nombre de composantes à construire.
2. Pour $k = 1; \ldots, K$:
   - 2.1. $\mathbf{a}_k = \mathbf{X}_{k-1}^T \mathbf{y}_{k-1} / \mathbf{y}_{k-1}^T \mathbf{y}_{k-1}$
   - 2.2. $\mathbf{w}_k = \mathbf{a}_k / \|\mathbf{a}_k\|_2$
   - 2.3. $\mathbf{t}_k = \mathbf{X}_{k-1}\mathbf{w}_k / \mathbf{w}_k^T \mathbf{w}_k$
   - 2.4. $\mathbf{p}_k = \mathbf{X}_{k-1}^T \mathbf{t}_k / \mathbf{t}_k^T \mathbf{t}_k$
   - 2.5. $\mathbf{X}_k = \mathbf{X}_{k-1} - \mathbf{t}_k \mathbf{p}_k^T$
   - 2.6. $c_k = \mathbf{y}_{k-1}^T \mathbf{t}_k / \mathbf{t}_k^T \mathbf{t}_k$
   - 2.7. $\mathbf{y}_k = \mathbf{y}_{k-1} - c_k \mathbf{t}_k$

Commentons quelque peu cet algorithme.

Le premier aspect très intéressant de cette méthode réside dans le fait qu'aucune diagonalisation ou inversion de matrice n'est nécessaire. Il ne s'agit *in fine* que de calculs des pentes de droites de régressions linéaires simples par MCO. Ce procédé peut donc être interprété de façon totalement géométrique, comme une suite de projections orthogonales sur variables latentes. La présence de données manquantes n'empêche donc pas l'algorithme de fonctionner, les calculs se faisant alors uniquement sur les données disponibles, similairement à ce que nous avons déjà exposé dans le cadre de l'algorithme NIPALS dans la partie 2.1.

Par conséquent, sans données manquantes, les étapes 2.1 et 2.2 peuvent être résumées en une seule étape suivant le résultat 7.3 obtenu ci-dessus, à savoir :

$$\mathbf{w}_k = \mathbf{X}_{k-1}^T \mathbf{y}_{k-1} / \left\| \mathbf{X}_{k-1}^T \mathbf{y}_{k-1} \right\|_2$$

L'étape 2.3 peut alors également être simplifiée en l'écrivant de la façon suivante :

$$\mathbf{t}_k = \mathbf{X}_{k-1}\mathbf{w}_k$$

Finalement, l'équation de régression PLS s'écrit de la façon suivante :

$$\begin{aligned} \mathbf{y} &= \sum_{k=1}^{K} c_k \mathbf{t}_k + \epsilon && (2.16) \\ &= \sum_{k=1}^{K} c_k \left(\sum_{j=1}^{p} w_{jk}^* \mathbf{x}_j \right) + \epsilon && (2.17) \\ &= \sum_{j=1}^{p} \beta_j^{PLS} \mathbf{x}_j + \epsilon, && (2.18) \end{aligned}$$

avec $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ le vecteur erreur de dimension n.

2.3 Propriétés de la régression PLS

2.3.1 Propriétés structurelles

Nous venons de décrire l'algorithme de base de la régression PLS. Ainsi construite, la structure latente possède des propriétés intéressantes, dont certaines nous semblent essentielles. Nous allons donc nous permettre de les rappeler.

Proposition 2.3.1. *(Höskuldsson, 1988, p.214) (Tenenhaus, 1998, p.101)*
Les vecteurs et matrices issues de la régression PLS vérifient les propriétés suivantes :

a) $\mathbf{t}_k^T \mathbf{t}_l = 0,\ l > k$

b) $\mathbf{w}_k^T \mathbf{w}_l = 0,\ l > k$

c) $\mathbf{t}_k^T \mathbf{X}_l = 0,\ l \geqslant k$

d) $\mathbf{w}_k^T \mathbf{X}_l^T = 0,\ l \geqslant k$

e) $\mathbf{X}_k = \mathbf{X} \prod_{h=1}^{k} \left(Id_p - \mathbf{w}_h \mathbf{p}_h^T\right),\ \forall h \in [\![1, K]\!]$

Ces propriétés, et leurs démonstrations disponibles notamment dans Tenenhaus (1998, p.101), confirment les notions d'orthogonalité recherchées et liées à la structure latente. Ces résultats d'orthogonalité permettent ainsi d'effectuer les régressions linéaires par MCO de façon optimal dans ce nouvel espace latent créé.

La propriété e) va nous permettre de définir les vecteurs poids étoilés $\mathbf{w}_k^*,\ k \in [\![1; K]\!]$ introduits dans l'équation 2.6 à partir de la formulation initiale 2.7 des composantes :

$$\mathbf{t}_k = \mathbf{X}_{k-1}\mathbf{w}_k = \mathbf{X}\mathbf{w}_k^*,\ \forall k \in [\![1, K]\!] \tag{2.19}$$

$$\Rightarrow \begin{cases} \mathbf{w}_1^* = \mathbf{w}_1 \\ \mathbf{w}_k^* = \prod_{h=1}^{k-1} \left(Id_p - \mathbf{w}_h \mathbf{p}_h^T\right) \mathbf{w}_k,\ \forall k \in [\![2; K]\!] \end{cases} \tag{2.20}$$

Cependant, il est possible d'obtenir des formulations plus agréables pour ces vecteurs poids étoilés que celle que nous venons de construire.

Proposition 2.3.2. *(Tenenhaus, 1998, p.114)*

1. *Les vecteurs* $\mathbf{w}_k^*$ *vérifient l'équation de récurrence suivante :*

$$\mathbf{w}_k^* = \mathbf{w}_k - \mathbf{w}_{k-1}^* \mathbf{p}_k^T \mathbf{w}_k$$

avec $\mathbf{w}_1^* = \mathbf{w}_1$

2. *La matrice* $\mathbf{W}_k^* = (\mathbf{w}_1^*, \ldots, \mathbf{w}_k^*)$ *vérifie l'équation suivante :*

$$\mathbf{W}_k^* = \mathbf{W}_k \left(\mathbf{P}_k \mathbf{W}_k^T\right)^{-1}$$

avec $\mathbf{P}_k = (\mathbf{p}_1, \ldots, \mathbf{p}_k) \in \mathcal{M}_{p,k}(\mathbb{R})$.

Cette matrice $\mathbf{W}_k^*$ peut ainsi être considérée comme étant un inverse généralisé de $\mathbf{P}_k^T$.

Proposition 2.3.3. *(Tenenhaus, 1998, p.109)*

$$\mathbf{P}_k^T \mathbf{W}_k^* = Id_k \tag{2.21}$$

Enfin, une dernière propriété qui nous semble importante à rappeler est liée cette fois-ci aux paramètres de la régression PLS. De part son procédé d'optimisation et de réduction de la dimension de l'espace prédictif, la régression PLS procède à un « rétrécissement », ou *shrinkage* en anglais, de la norme du vecteur des coefficients de régression en fonction du nombre de composantes. En effet, De Jong (1995) a démontré que cette norme est une fonction strictement croissante du nombre de composantes.

Proposition 2.3.4. *(De Jong, 1995)*
Soit $r = \mathrm{Rg}(\mathbf{X})$*, alors :*

$$\left\|\beta_1^{PLS}\right\|_2 < \left\|\beta_2^{PLS}\right\|_2 < \ldots < \left\|\beta_r^{PLS}\right\|_2 \tag{2.22}$$

2.3.2 Degrés de liberté

Le développement des degrés de liberté liés à un modèle de régression PLS a été effectué par Krämer and Sugiyama (2011). L'établissement de ces degrés de liberté permet une comparaison précise des modèles obtenus en termes de complexité et de sensibilité. De plus, les critères d'information tels que l'AIC (Akaike, 1974) ou le BIC (Schwarz, 1978) ont également pu être adaptés, permettant ainsi de devenir des critères de choix de modèle, *i.e.* de sélection du nombre de composantes, plus pertinents. Précisons tout de même que ce développement ne concerne que la régression PLS usuelle et sans données manquantes.

Ayant utilisé le critère du BIC adapté (Krämer and Sugiyama, 2011) lors de nos recherches, critère dont nous rappellerons la forme dans le chapitre 5, ainsi que les degrés de liberté, notamment dans nos recherches présentées dans le chapitre 7, il nous semble important d'effectuer quelques rappels à ce sujet.

Établir les degrés de liberté d'un modèle linéaire revient à évaluer sa complexité, laquelle est intrinsèquement liée à la *matrice chapeau*, ou *hat matrix*, que l'on note $\mathbf{H}$. Cette matrice chapeau est définie comme la matrice permettant de passer de la réponse observée à la réponse estimée, *i.e.* :

$$\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y} \tag{2.23}$$

Lorsque $\mathbf{H}$ ne dépend pas de $\mathbf{y}$, les degrés de liberté sont définis comme suit.

Definition 2.3.1. *Hastie* et al. *(2009, p.232)*

$$DoF = \mathrm{trace}(\mathbf{H}) \tag{2.24}$$

Un exemple bien connu vérifiant cette propriété est le cas des modèles de régression par MCO. Dans ce cadre là, la matrice chapeau est la suivante :

$$\mathbf{H}^{MCO} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \tag{2.25}$$

Cette matrice, en plus de ne dépendre que de $\mathbf{X}$, est une matrice de projection. La trace de $\mathbf{H}^{MCO}$ est alors égale à son rang et l'on retrouve ainsi le résultat bien connu que les degrés de liberté associés à un tel modèle sont égaux à p (ou $p+1$ si l'on intègre l'estimation de l'ordonnée à l'origine). D'autres exemples de méthodes liées à une matrice chapeau ne dépendant pas de $\mathbf{y}$ sont la régression *Ridge* (Hoerl and Kennard, 1970) ou la régression sur CP.

Dans le cadre de la régression PLS, la matrice chapeau associée s'écrit :

$$\mathbf{H}_K = \mathbf{T}_K\left(\mathbf{T}_K^T\mathbf{T}_K\right)^{-1}\mathbf{T}_K^T \tag{2.26}$$

Celle-ci ne dépend donc pas uniquement de $\mathbf{X}$ mais également de $\mathbf{y}$, rendant ainsi la définition 2.3.1 obsolète. Efron (2004) propose une formulation plus générique des degrés de liberté sur laquelle vont s'appuyer Krämer and Sugiyama (2011).

Proposition 2.3.5. *(Efron, 2004, p.620)*

$$DoF = \mathbb{E}\left[\text{trace}\left(\frac{\partial\hat{\mathbf{y}}}{\partial\mathbf{y}}\right)\right] \tag{2.27}$$

Dans le cas où $\mathbf{H}$ ne dépend pas de $\mathbf{y}$ nous obtenons $\frac{\partial\hat{\mathbf{y}}}{\partial\mathbf{y}} = \frac{\partial\mathbf{H}\mathbf{y}}{\partial\mathbf{y}} = \mathbf{H}$ et retrouvons la définition initiale 2.3.1. Il est aussi intéressant de remarquer que cette définition traduit le fait que les degrés de liberté représentent en réalité la sensibilité de l'estimation obtenue par rapport aux données observées.

La problématique dans ce cadre de la régression PLS se résume donc à estimer la matrice $\frac{\partial\hat{\mathbf{y}}}{\partial\mathbf{y}}$. Pour ce faire, Krämer and Sugiyama (2011) vont s'appuyer sur les relations fortes existant entre la structure latente créée par la régression PLS et les espaces de Krylov. Nous allons donc rapidement introduire ces espaces avant de rappeler leur lien avec la structure latente créée par la régression PLS.

Plaçons nous dans le cadre d'un objectif de résolution du système linéaire suivant :

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{2.28}$$

avec $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$ et $\mathbf{b} \in \mathbb{R}^n$. Dans les cas non-triviaux, *i.e.* lorsque $\mathbf{A}$ est difficilement inversible amenant des problèmes de conditionnement, ou lorsque cette matrice est liée à des dimensions importantes, l'application de méthodes de résolution itératives devient essentielle, notamment afin de réduire les temps de calculs (Freund *et al.*, 1992). Certaines de ces méthodes vont ainsi utiliser des projections dans des sous-espaces particuliers, appelés les *espaces de Krylov*.

Definition 2.3.2. *(Boley, 1994, p.2) On appelle espace de Krylov d'ordre p, noté $\mathcal{K}_p(\mathbf{v},\mathbf{A})$, l'espace vectoriel généré par $\mathbf{v} \in \mathbb{R}^n$ et ses $p-1$ produits itérés par $\mathbf{A}$:*

$$\mathcal{K}_p(\mathbf{v},\mathbf{A}) = Vect\left\{\mathbf{v}, \mathbf{A}\mathbf{v}, \mathbf{A}^2\mathbf{v}, \ldots, \mathbf{A}^{p-1}\mathbf{v}\right\}. \tag{2.29}$$

Plus de détails sur les espaces de Krylov sont fournis, entre autres, par Freund *et al.* (1992), Boley (1994) ou Saad (2003, ch.6).

Le lien entre les espaces de Krylov et la régression PLS a été proposé initialement par Manne (1987) et Helland (1988). Ils ont ainsi montré que la suite des poids $\mathbf{w}_1, \ldots, \mathbf{w}_K$ pouvait être obtenue par orthogonalisation de la suite de Krylov $\mathcal{K}_K(s, C) = \left\{\mathbf{s}, \mathbf{Cs}, \ldots, \mathbf{C}^{K-1}\mathbf{s}\right\}$ avec $\mathbf{s} = \mathbf{X}^T\mathbf{y}$ et $\mathbf{C} = \mathbf{X}^T\mathbf{X}$. On peut ainsi obtenir une expression de $\hat{\beta}_k^{PLS}$ dépendant de l'espace engendré par cette suite de Krylov.

Proposition 2.3.6. *(Helland, 1990)*

$$\hat{\beta}_k^{PLS} = \underset{\beta \in \mathcal{K}_k(s,C)}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\beta\|^2 \tag{2.30}$$

Des résultats similaires ont ensuite été obtenu par De Jong (1993b) concernant les suites $\{\mathbf{w}_k^*\}$, $\{\mathbf{t}_k\}$ et $\{\mathbf{p}_k\}$.

Proposition 2.3.7. *(Tenenhaus, 1998, p.109)*
Soit $\mathbf{s} = \mathbf{X}^T\mathbf{y}$, $\mathbf{t} = \mathbf{Xs}$, $\mathbf{C} = \mathbf{X}^T\mathbf{X}$ *et* $\mathbf{D} = \mathbf{X}\mathbf{X}^T$. *Alors, il y a équivalence entre les espaces engendrés par les suites* $\{\mathbf{w}_k\}$, $\{\mathbf{w}_k^*\}$, $\{\mathbf{t}_k\}$ *et* $\{\mathbf{p}_k\}$ *et ceux engendrés par les suites de Krylov définis ci-après :*

1. $\{\mathbf{w}_1, \ldots, \mathbf{w}_K\} \sim \left\{\mathbf{s}, \mathbf{Cs}, \ldots, \mathbf{C}^{K-1}\mathbf{s}\right\}$
2. $\{\mathbf{w}_1^*, \ldots, \mathbf{w}_K^*\} \sim \left\{\mathbf{s}, \mathbf{Cs}, \ldots, \mathbf{C}^{K-1}\mathbf{s}\right\}$
3. $\{\mathbf{t}_1, \ldots, \mathbf{t}_K\} \sim \left\{\mathbf{t}, \mathbf{Dt}, \ldots, \mathbf{D}^{K-1}\mathbf{t}\right\}$
4. $\{\mathbf{p}_1, \ldots, \mathbf{p}_K\} \sim \left\{\mathbf{Cs}, \mathbf{C}^2\mathbf{s}, \ldots, \mathbf{C}^K\mathbf{s}\right\}$

Ces résultats ont permis l'élaboration d'un grand nombre d'avancées théoriques liées au domaine de la PLS. On peut citer entre autres l'équivalence entre la PLS1 et l'algorithme SIMPLS (De Jong, 1993b), la variance asymptotique des estimateurs d'une régression PLS (Phatak *et al.*, 2002), (Romera, 2010) ou les travaux menés par Blazere (2015) quant aux liens entre la régression PLS et les polynômes orthogonaux.

Krämer and Sugiyama (2011) se sont servis du résultat énoncé ci-dessus afin de développer ces degrés de liberté, à savoir que :

$$\{\mathbf{t}_1, \ldots, \mathbf{t}_K\} \sim \mathcal{K}_K(t, D) = \left\{\mathbf{t}, \mathbf{Dt}, \ldots, \mathbf{D}^{K-1}\mathbf{t}\right\} \tag{2.31}$$

Ils ont ainsi utilisé la représentation de $\hat{\mathbf{y}}_K$ qui en découle :

$$\hat{\mathbf{y}}_K = \overline{\mathbf{y}} + \mathcal{P}_{\mathcal{K}_K(t,D)} \tag{2.32}$$

De la même manière dont Phatak *et al.* (2002) ont procédé concernant les estimateurs de la régression PLS, Krämer and Sugiyama (2011) en ont déduit un résultat concernant la différentielle de $\hat{\mathbf{y}}_K$ suivant $\mathbf{y}$. Le résultat est le suivant.

Proposition 2.3.8. *(Krämer and Sugiyama, 2011)*

$$\frac{\partial \hat{\mathbf{y}}_K}{\partial \mathbf{y}} = \frac{1}{n} Id_n + \sum_{k=1}^{K} c_k \left(Id_n - \mathbf{T}\mathbf{T}^T \right) \mathbf{D}^k + \sum_{k=1}^{K} \mathbf{v}_k \left(\mathbf{y} - \hat{\mathbf{y}}_K \right) \mathbf{D}^k + \mathbf{T}\mathbf{T}^T \tag{2.33}$$

avec $\mathbf{B} = \left(\left\langle \mathbf{t}_{k_1}, \mathbf{D}^{k_2}\mathbf{y} \right\rangle \right) \in \mathcal{M}_K(\mathbb{R})$, $\mathbf{c} = \mathbf{B}^{-1}\mathbf{T}^T\mathbf{y} \in \mathbb{R}^K$ *et* $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_K) = \mathbf{T}\left(\mathbf{B}^{-1}\right)^T \in \mathcal{M}_{n,K}(\mathbb{R})$

La trace en découle directement, d'où la formulation suivante des degrés de liberté :

Proposition 2.3.9. *(Krämer and Sugiyama, 2011)*

$$\widehat{DoF}_K = 1 + \sum_{k=1}^{K} c_k \operatorname{trace}\left(\mathbf{D}^k\right) - \sum_{k_1,k_2=1}^{K} \mathbf{t}_{k_1}^T \mathbf{D}^{k_2} \mathbf{t}_{k_1} + \left(\mathbf{y} - \hat{\mathbf{y}}_K \right) \sum_{k=1}^{K} \mathbf{D}^k \mathbf{v}_k + K \tag{2.34}$$

2.4 Les extensions de la régression PLS

Afin d'améliorer ses performances, des adaptations de la régression PLS ont été développées. Nous allons dans ce travail de thèse en étudier deux plus particulièrement. La première extension consiste à adapter l'algorithme de régression PLS afin de prendre en compte la distribution liée à $\mathbf{y}$ en utilisant les procédés d'estimations effectués dans le cadre des modèles de régression linéaire généralisée. La deuxième modification, dénommée *Sparse* PLS, consiste à procéder à une sélection des régresseurs afin de décomplexifier le modèle.

2.4.1 Les extensions aux modèles linéaires généralisés

Dans la pratique, la réponse étudiée est souvent liée à une distribution précise, notamment discrète. Celle-ci peut-être binaire lorsqu'elle représente par exemple la présence ou non d'une anomalie allélique. Elle peut également être de nature multi-classe ordinale, représentant par exemple la qualité d'un certain produit, ou être modélisée par une loi de Poisson si elle représente un certain comptage.

Dans ces cas, et bien d'autres, il parait donc judicieux d'intégrer la notion de modèles linéaires généralisés (GLM) au procédé classique de la régression PLS.

Dans le cadre des modèles linéaires généralisés, l'estimation des paramètres est généralement obtenue par maximisation de la vraisemblance. En pratique, cette estimation est obtenue à l'aide d'un algorithme équivalent à l'algorithme de Newton-Raphson, connu sous le nom de *Iteratively Reweighted Least Squares*, correspondant ainsi à l'acronyme *IRLS*, qui a été développé par Green (1984). Chaque itération de cet algorithme se décompose en deux étapes :

$$\mathbf{z}^{(t)} = \mathbf{X}\beta^{(t)} + \left[\widetilde{\mathbf{W}}^{(t)} \right]^{-1} \left(\mathbf{y} - g^{-1}\left(\mathbf{X}\beta^{(t)} \right) \right) \tag{2.35}$$

$$\beta^{(t+1)} = \left(\mathbf{X}^T \widetilde{\mathbf{W}}^{(t)} \mathbf{X} \right)^{-1} \mathbf{X}^T \widetilde{\mathbf{W}}^{(t)} \mathbf{z}^{(t)} \tag{2.36}$$

avec g la fonction de lien associée à la distribution concernée et $\widetilde{\mathbf{W}}^{(t)} \in \mathcal{M}_n(\mathbb{R})$ une matrice diagonale tel que $\widetilde{W}_{ii}^{(t)} = (g^{-1})' \left(\beta^{(t)^T} \mathbf{x}_{(i)}^T\right)$, avec $\mathbf{x}_{(i)} \in \mathbb{R}^{1\times p}$ le vecteur représentant la $i^{\text{ème}}$ ligne de $\mathbf{X}$.

Ces itérations sont effectuées, soit jusqu'à convergence au sens de la déviance, soit jusqu'à atteindre le nombre d'itérations maximales choisi par l'utilisateur. Des explications détaillées quant à l'algorithme *IRLS* et son développement sont disponibles, entre autres, dans le livre de Gaussier and Yvon (2011).

Marx (1996) a été le premier à introduire la notion de modèles linéaires généralisés dans le domaine de la PLS. Pour ce faire, il a développé une méthode nommée *Iteratively Reweighted Partial Least Squares* (*IRPLS*), consistant à remplacer la régression linéaire pondérée dans l'équation 2.36 par une régression PLS pondérée à K composantes de $\mathbf{z}^{(t)}$ sur $\mathbf{X}$ avec une pondération matérialisée par la matrice $\widetilde{\mathbf{W}}^{(t)}$. Précisons qu'une régression PLS pondérée est l'analogue de la régression PLS en remplaçant l'ensemble des régressions par MCO par des régressions linéaires pondérées (Fort and Lambert-Lacroix, 2005). Une fois la matrice des composantes $\mathbf{T}^{IRPLS}$ ainsi construite, il s'agit d'effectuer l'algorithme *IRLS* appliqué à $\mathbf{y}$ et $\mathbf{T}^{IRPLS}$ pour obtenir les coefficients de la régression.

La recherche autour de l'extension de la régression PLS aux *GLM* fut très active dans les années qui ont suivi ce développement. En effet, des problèmes de convergence ont rapidement été observés, notamment dans le cadre logistique où la convergence de l'algorithme *IRLS* est dépendante de la structure des données (Albert and Anderson, 1984). Nous détaillons ces différents types de structure dans le chapitre 8. Ainsi, plusieurs travaux ont vu le jour afin de palier à ce problème. On peut considérer l'ensemble des procédés développés dans ce domaine comme appartenant à deux groupes distincts de méthodologie.

Le premier regroupe les procédés inspirés des travaux de Marx (1996) *i.e.* intégrant directement la notion de modèles généralisés dans la construction des composantes PLS. Ding and Gentleman (2005) ont ainsi proposé un procédé pouvant être considéré comme une adaptation de la méthode proposée par Marx (1996) et qui consiste à y intégrer la procédure de Firth (Firth, 1993) afin d'éviter les problèmes de non-convergence. Fort and Lambert-Lacroix (2005) ont proposé l'utilisation de l'estimateur de *Ridge* (Le Cessie and Van Houwelingen, 1992). Ainsi, au lieu d'utiliser la réponse d'origine $\mathbf{y}$, ils déterminent une pseudo-réponse $\tilde{\mathbf{y}}$ obtenu après convergence de l'algorithme *IRLS* où l'équation 2.36 est remplacé par une régression pondérée de *Ridge* :

$$\beta^{(t+1)} = \left(\mathbf{X}^T \widetilde{\mathbf{W}}^{(t)} \mathbf{X} + \lambda_r \Sigma^2\right)^{-1} \mathbf{X}^T \widetilde{\mathbf{W}}^{(t)} \mathbf{z}^{(t)} \tag{2.37}$$

avec Σ^2 une matrice diagonale tel que $\Sigma_{jj}^2 = (n-1)\,\widehat{\text{Var}}(\mathbf{x}_j)$ et λ_r le paramètre de contrainte lié à la régression de *Ridge*.

La procédure se termine en effectuant une régression PLS pondérée liée à $\tilde{\mathbf{y}}$ et une matrice poids obtenue également après convergence de l'algorithme *IRLS* présenté ci-dessus.

On peut également citer les travaux de Nguyen and Rocke (2004) et Bastien *et al.* (2005) qui entrent dans ce cadre là.

La deuxième catégorie regroupe des méthodes que l'on peut définir en deux étapes distinctes. Tout d'abord l'obtention des composantes PLS est effectuée à l'aide de la régression PLS usuelle, *i.e.* en ne tenant pas compte d'une distribution particulière de la réponse. Une fois les composantes obtenues, une régression logistique, une analyse discriminante quadratique ou encore une analyse discriminante linéaire sont des exemples de procédés appliqués aux composantes obtenues afin d'établir le modèle. Plus de détails sur les méthodes de classification sont disponibles dans le livre de Hastie *et al.* (2009, ch.4). Dans ce cadre, on peut citer les travaux de Nguyen and Rocke (2002a), Nguyen and Rocke (2002b) ou encore Boulesteix (2004).

Il nous a paru préférable et plus cohérent d'opter pour une méthodologie appartenant au premier groupe, *i.e.* dont la construction des composantes PLS tienne compte de la distribution originale de la réponse. Ainsi, dans le cadre de ce travail de thèse, nous avons utilisé la formulation proposée et décrite par Bastien *et al.* (2005) afin de développer et adapter nos nouveaux critères de sélection au cadre des modèles généralisés. En effet, son implémentation au sein du package *R plsRglm* (Bertrand *et al.*, 2014) permettant de traiter plusieurs types de distribution, cela nous a offert plus de souplesse dans nos recherches. Voici l'algorithme de construction tel qu'implémenté dans le package :

1. Construction de $\mathbf{t}_1$:
   - Pour $j = 1, \ldots, p$, déterminer a_{1j} le coefficient de la régression linéaire généralisée de $\mathbf{y}$ sur $\mathbf{x}_j$.
   - $\mathbf{w}_1 = \mathbf{a}_1 / \|\mathbf{a}_1\|$ avec $\mathbf{a}_1 = (a_{11}, \ldots, a_{1p})$
   - Construire $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1 / \mathbf{w}_1^T\mathbf{w}_1$
2. Pour $k = 2, \ldots, K$:
   - Déterminer la matrice résidus $\mathbf{X}_{k-1} = (\mathbf{x}_{k-1,1}, \ldots, \mathbf{x}_{k-1,p})$ de la régression par MCO de $\mathbf{X}$ sur $\mathbf{T}_{k-1}$
   - Pour $j = 1, \ldots, p$, déterminer a_{kj} le coefficient associé à $\mathbf{x}_{k-1,j}$ dans la régression linéaire généralisée de $\mathbf{y}$ sur $\mathbf{T}_{k-1}$ et $\mathbf{x}_{k-1,j}$.
   - $\mathbf{w}_k = \mathbf{a}_k / \|\mathbf{a}_k\|$ avec $\mathbf{a}_k = (a_{k1}, \ldots, a_{kp})$
   - Construire $\mathbf{t}_k = \mathbf{X}_{k-1}\mathbf{w}_k / \mathbf{w}_k^T\mathbf{w}_k$
3. Effectuer la régression linéaire généralisée de $\mathbf{y}$ sur $\mathbf{T}_K$.
4. Obtenir le modèle final en fonction des prédicteurs d'origine.

2.4.2 La régression *Sparse* PLS

L'objectif de procéder simultanément à une réduction de la dimension, ce qui est effectuée par la régression PLS, et à une sélection des prédicteurs relève autant d'un intérêt computationnel en permettant de réduire les temps de calculs, que d'un intérêt mathématique et statistique ou encore d'un intérêt biologique. En effet, à travers cette simplification

des modèles obtenus, les praticiens biologistes se voient aidés dans l'interprétation qui peut en être faite. C'est d'ailleurs dans cet optique là et pour répondre à ce besoin des praticiens biologistes que Lê Cao *et al.* (2008) seront les premiers à développer une adaptation de la régression PLS consistant à effectuer simultanément une réduction de la dimension et une sélection des prédicteurs, méthodologie qu'il dénommera *Sparse* PLS. Pour ce faire, ils se sont appuyés sur la formulation de la régression PLS effectuée à l'aide de la décomposition en valeurs singulières, formulation énoncée par Lorber *et al.* (1987). N'ayant pas utilisé cette formulation de la *Sparse* PLS dans le cadre de cette thèse, nous ne détaillerons pas plus cette méthodologie.

En réalité, précisons qu'un article interne au département de statistique de l'Université du Winsconsin, et mentionné par Lê Cao *et al.* (2008), a été publié en 2007 et propose lui aussi une méthodologie de *Sparse* PLS (Chun and Keles, 2007) qui aboutira à une publication officielle trois années plus tard (Chun and Keleş, 2010). Cet article contient ainsi un procédé consistant en l'intégration d'une contrainte de type $_1$ sur les vecteurs des poids $\mathbf{w}_h$. Formalisons quelque peu cela.

Chun and Keleş (2010) ont motivé leur travail à partir de la démonstration d'un théorème de non-consistance des estimations dans le cas $n \ll p$.

Théorème 2.4.1. *(Chun and Keleş, 2010, p.6) Supposons que* $\mathbf{y} = \mathbf{X}\beta + \epsilon$ *avec* $\epsilon \sim \mathcal{N}(0, \sigma^2)$, $\|\beta\|_2 < 0$ *et* σ *une constante. Alors :*

1. *si* $p/n \to 0$, *alors* $\left\|\hat{\beta}^{PLS} - \beta\right\|_2 \to 0$ *en probabilité.*
2. *si* $p/n \to c_0$ *avec* $c_0 > 0$, *alors* $\left\|\hat{\beta}^{PLS} - \beta\right\|_2 > 0$ *en probabilité.*

Précisons que des hypothèses supplémentaires sont nécessaires à la démonstration de ce théorème, hypothèses naturellement disponibles dans l'article de Chun and Keleş (2010).

Ainsi élaborer une méthodologie permettant de sélectionner un nombre restreint de prédicteurs pour l'élaboration du modèle relève d'un intérêt mathématique et statistique certain. Afin de parvenir à cela, les auteurs ont imposé une contrainte de type L_1 sur le vecteur poids $\mathbf{w}$ amenant ainsi à la formulation de la fonction objectif suivante :

$$\mathbf{w}_k = \underset{\mathbf{w}\in\mathbb{R}^p}{\operatorname{argmax}} \left\{\mathbf{w}^T \mathbf{X}_{k-1}^T \mathbf{y}_{k-1} \mathbf{y}_{k-1}^T \mathbf{X}_{k-1} \mathbf{w}\right\}, \text{ s.c. } \|\mathbf{w}\|_2^2 = 1, \|\mathbf{w}\|_1 < \lambda_{spls} \tag{2.38}$$

Jolliffe *et al.* (2003) ont développé une approche similaire dans le cadre de l'analyse en CP et ont pointé le fait que cette formulation n'était pas assez parcimonieuse et liée à un problème non-convexe, ce qui empêche l'application du théorème d'unicité d'extremum global.

Afin de palier à ces problèmes, ils se sont inspirés des travaux de Zou *et al.* (2006) en intégrant la pénalité L_1 non pas sur $\mathbf{w}$ mais sur un vecteur dit de substitution $\mathbf{v}$ tout en le gardant proche de $\mathbf{w}$, amenant ainsi la formulation suivante :

$$\min_{\mathbf{w},\mathbf{v}} \left\{-\kappa \mathbf{w}^T \mathbf{M} \mathbf{w} + (1-\kappa)(\mathbf{v}-\mathbf{w})^T \mathbf{M} (\mathbf{v}-\mathbf{w}) + \lambda_1 \|\mathbf{v}\|_1 + \lambda_2 \|\mathbf{v}\|_2^2\right\}, \text{ s.c. } \mathbf{w}^T\mathbf{w} = 1. \tag{2.39}$$

avec $\mathbf{M} = \mathbf{X}^T \mathbf{y} \mathbf{y}^T \mathbf{X}$.

Dans ce cas, l'utilisation de faibles valeurs de κ tend à solutionner le problème de non-convexité entraînant le problème de solution locale (Chun and Keleş, 2010, p.9). Après développement, ils obtiennent, dans le cas univarié, la solution suivante :

$$\hat{\mathbf{v}}_k = \left(\left|\mathbf{Z}^{(k)}\right| - \lambda_1/2\right)_+ \operatorname{sgn}\left(\mathbf{Z}^{(k)}\right) \tag{2.40}$$

avec $\mathbf{Z}^{(k)} = \mathbf{X}^T\mathbf{y}_{k-1}/\left\|\mathbf{X}^T\mathbf{y}_{k-1}\right\|$, $\mathbf{y}_{k-1}$ le vecteur des résidus de la régression linéaire de $\mathbf{y}$ sur $\mathbf{T}_{k-1}$, $\mathbf{y}_0 = \mathbf{y}$ et $(.)_+$ représentant la partie positive.

On peut exprimer cette formulation de la façon suivante :

$$\hat{\mathbf{v}}_k = \left(\left|\mathbf{Z}^{(k)}\right| - \nu \max_{1\leqslant j\leqslant p}\left|\mathbf{Z}_j^{(k)}\right|\right)_+ \operatorname{sgn}\left(\mathbf{Z}^{(k)}\right) \tag{2.41}$$

avec $0 \leqslant \nu \leqslant 1$ jouant le rôle du paramètre de parcimonie λ_1 présent dans (2.40).

De manière plus pratique, voici l'algorithme correspondant pour ν et K fixés :

- Soit $\widehat{\beta}^{SPLS} = 0$, $\mathcal{A} = \varnothing$, $k = 1$ et $\mathbf{y}_1 = \mathbf{y}$.
- Tant que $k \leqslant K$:
  1. Déterminer $\widehat{\mathbf{w}}_k$ tel qu'énoncé dans 2.41 avec $\mathbf{Z}^{(k)} = \mathbf{X}^T\mathbf{y}_1/\left\|\mathbf{X}^T\mathbf{y}_1\right\|$.
  2. $\mathcal{A} = \mathcal{A} \cup \{j : \widehat{w}_{j,k} \neq 0\}$.
  3. Effectuer la régression PLS à k composantes de $\mathbf{y}$ sur $\mathbf{X}_\mathcal{A}$, puis en extraire les nouvelles estimations des coefficients de régression $\widehat{\beta}^{SPLS}$.
  4. Poser $\mathbf{y}_1 = \mathbf{y} - \mathbf{X}_\mathcal{A}\widehat{\beta}^{SPLS}$ et $k = k + 1$.
- Retourner $\mathcal{A}$ et $\widehat{\beta}^{SPLS}$.

La détermination du couple d'hyperparamètres (ν, K) est ensuite effectuée par VC.

Enfin, précisons que des adaptations de la procédure ont été effectuées par Chung and Keles (2010) et Durif *et al.* (2015) afin de traiter les problèmes de classification. Nous y reviendrons dans les chapitres 6 et 7.

2.5 Détermination du nombre de composantes

2.5.1 Problématique de l'état de l'art

La détermination du nombre de composantes est un élément essentiel dans le processus de la régression PLS ainsi que pour ses extensions que nous venons de présenter dans la partie précédente 2.4. Ce nombre représente la dimension de l'espace dans lequel la réponse va être modélisée (projetée dans le cas de la régression PLS usuelle). Ainsi, retenir trop peu de composantes revient à une perte d'information pouvant être initialement modélisée. Au contraire, conclure en un trop grand nombre de composantes entraîne le phénomène bien connu de sur apprentissage (*over-fitting*) où la réponse sera bien modélisée mais le modèle sera lié à de faibles performances prédictives puisqu'incluant du bruit dans la matrice de prédicteurs (Wiklund *et al.*, 2007).

La détermination de cet hyperparamètre a été, et est toujours, un sujet de recherche important. Une quantité conséquente de travaux de recherche a ainsi été menée à ce sujet, menant à de nombreux critères ou procédés de sélection de ce nombre de composantes.

La régression PLS usuelle n'est en aucun cas liée à des hypothèses ou à des distributions précises si ce n'est l'indépendance des observations et l'appartenance des variables aléatoires parentes à l'espace L^2 (2.2.2). C'est d'ailleurs de là que lui vaut son qualificatif de *soft modeling* (Manne, 1987). Ceci empêche donc tout développement de tests de significativité fiable liés aux composantes et basés sur des lois théoriques (Wakeling and Morris, 1993). Ainsi, afin de déterminer le nombre de composantes à sélectionner, la recherche s'est essentiellement concentrée sur la comparaison des performances prédictives liées aux différents modèles, performances qui représentent souvent un objectif majeur dans l'établissement de modèles de régression PLS (Denham, 2000, p.352). La qualité de prédiction peut être évaluée à l'aide de la statistique nommée *Predictive Residual Error Sum of Squares* et introduite par Allen (1971) à des fins de sélection de modèles. Afin d'utiliser cette statistique, il faut disposer de deux jeux de données distincts. Le premier appelé jeu d'entraînement, ou *training set*, va être utilisé afin de procéder à l'estimation des paramètres du modèle. Une fois le modèle établi, il est appliqué au second jeu de données, appelé jeu test, afin d'obtenir les estimations de la réponse. Idéalement, il faut disposer d'un jeu test supplémentaire et indépendant du jeu de données d'entraînement (Forina *et al.*, 2004). Cependant, dû à des raisons logistiques ou financières, il est relativement rare en pratique de disposer de deux jeux de données indépendants si bien qu'une validation croisée (VC) est majoritairement effectuée (Wakeling and Morris, 1993) (Efron and Tibshirani, 1993, p.240). Cette statistique s'écrit alors de la façon suivante :

$$\text{PRESS} = \sum_{i=1}^{I} \left(y_i - \widehat{y_i}\right)^2 \tag{2.42}$$

avec I le nombre de données contenus dans le jeu test, y_i la $i^{\text{ème}}$ réponse du jeu test et $\widehat{y_i}$ la réponse estimée à l'aide du modèle préalablement établi sur le jeu de données d'entraînement.

La majorité des critères développés se basent ainsi sur cette statistique. Les premiers que l'on peut citer consistent simplement à déterminer le nombre de composantes à retenir comme étant celui lié, soit au modèle réalisant le minimum global de cette statistique sur l'ensemble des $r = \text{Rg}(X)$ modèles possibles, soit au premier modèle réalisant un minimum local ou encore au premier modèle lié à une valeur du PRESS inférieur à un certain seuil fixé par l'utilisateur (Forina *et al.*, 2004).

Wold (1978) a définit la statistique suivante :

$$R_k = \frac{\text{PRESS}_{k+1}}{\text{PRESS}_k} \tag{2.43}$$

avec PRESS_k la valeur de la statistique du PRESS lié au modèle à k composantes.

Il définit alors le nombre de composantes K à retenir de la façon suivante :

$$K = \min\{k \text{ tel que } R_k > 1\} \tag{2.44}$$

Osten (1988) montrera que le critère du minimum globale du PRESS n'est pas satisfaisant et préconise ainsi le critère du premier minimum local, à savoir le critère R de Wold (1978).

Une adaptation du seuil passant de 1 à 0,9 ou 0,95 renvoie ainsi au critère du R ajusté. Ce changement de seuil est motivé par Krzanowski (1987) et revient ainsi à durcir la condition d'inclusion d'une nouvelle composante.

Une autre statistique basée sur le PRESS est le critère dit du Q^2. Cette statistique est la suivante :

$$Q_k^2 = 1 - \frac{\text{PRESS}_k}{\text{RSS}_{k-1}} \tag{2.45}$$

avec $\text{RSS}_{k-1} = \sum_{i=1}^{n} (y_i - \widehat{y}_{k-1,i})$ la somme des carrés résiduelles où $\widehat{y}_{k-1,i},\ i \in [\![1, n]\!]$, représentent les estimations des éléments du vecteur réponse obtenues à l'aide du modèle à $k-1$ composantes.

Tenenhaus (1998, p.83) considère alors qu'une nouvelle composante $\mathbf{t}_k$ est significative si $Q_k^2 \geqslant 0,0975$.

En ce qui concerne la VC, plusieurs configurations sont possibles. La première, dénommée *leave-one-out CV*, consiste à ne sélectionner qu'un seul sujet comme jeu test, utilisant ainsi les $n-1$ données restantes afin d'établir le modèle. Ce procédé est alors répété n fois afin que chaque observation ait joué le rôle de jeu test. Il s'agit là certainement de l'option la plus utilisée dans la littérature (Gourvénec *et al.*, 2003) (Gómez-Carracedo *et al.*, 2007). Cependant, Gourvénec *et al.* (2003) remarquent que cette technique appliquée à l'évaluation du PRESS tend à retourner trop de composantes. De plus, Shao (1993) a démontré que cette méthode n'est asymptotiquement pas convergente dans le sens où la probabilité de sélectionner le modèle lié aux meilleurs performances prédictives ne converge pas vers 1 quand $n \to \infty$. Il montre ainsi que ce problème peut être théoriquement résolu en réservant n_v observations pour le jeu test, avec n_v satisfaisant $n_v/n \to 1$ quand $n \to \infty$. Deux autres procédés de VC pourraient ainsi représenter des améliorations de la *leave-one-out CV*. Le premier, dénommée *MCCV* pour *Monte Carlo CV* (Picard and Cook, 1984), consiste à réserver un nombre n_v bien plus important d'observations pour le jeu test (plus de la moitié des observations par exemple). Ces observations sont tirées de façon aléatoire et sans remise parmi l'ensemble des données disponibles. Ce processus est alors répété plusieurs fois (Gourvénec *et al.*, 2003). Le second, nommé *q-fold CV* (Breiman *et al.*, 1984), consiste à diviser le jeu de données initial en q groupes de taille identique (si possible) et de successivement en réserver un seul en tant que jeu test. Ce procédé est alors répété q fois afin que chacun des groupes ait été une fois le jeu test. Ainsi, si $q = n$ on retombe sur le procédé de la *leave-one-out CV*. Concernant le *MCCV*, Gourvénec *et al.* (2003) mentionne une question et problématique essentielle. Il prétend que nous sommes en droit de nous questionner sur la pertinence d'un modèle établi sur un jeu de données d'entraînement contenant une si petite partie des observations initiales. En d'autres termes, en réservant autant d'observations pour le jeu test, il est légitime de se demander si les observations restantes reflètent encore la structure initiale du jeu de données original. En ce qui concerne la *q-fold CV*, il est considéré comme étant un « bon »choix de créer entre 5 et 10 groupes, *i.e.* $q \in [\![5, 10]\!]$ (Kohavi, 1995) (Wiklund *et al.*, 2007). Malgré toutes ces tentatives d'utilisation et de modification de la VC, cette méthode reste controversée notamment dû à des problèmes de forte variabilité de l'estimation des erreurs de prédictions comme le reporte entre autres Efron and Tibshirani (1993, p.255) et Hastie *et al.* (2009, p.249). Ainsi, comme mentionné par Hastie *et al.* (2009, p.242), bien que diminuer le nombre q de groupes créés

pour la VC (ce qui revient à augmenter le nombre d'observations inclus dans le jeu test) entraîne une diminution de la variance, cela peut entraîner une augmentation du biais. D'autres arguments quant à la pertinence de l'utilisation des erreurs prédictives ainsi que celle de la VC sont exposés par Wiklund *et al.* (2007, p.429). Ces importants problèmes d'instabilité liés à la VC pour la détermination du nombre de composantes seront également repris par Boulesteix (2014).

D'autres méthodes n'ayant pas recours à la VC ont également été développées. Ainsi, l'utilisation de techniques bootstrap pour l'estimation des erreurs de prédictions a été proposée par Efron and Tibshirani (1993). En effet, ces techniques sont liées à une variabilité plus faible mais, malheureusement, dans certains cas, à un important biais comme le reportent Kohavi (1995) ou Efron and Tibshirani (1993, p.255).

Afin d'éviter la détermination arbitraire de certains seuils, des tests statistiques ont également été proposés et sont basés sur certaines hypothèses de distribution. Ainsi, Haaland and Thomas (1988) remarquant que sélectionner le nombre de composantes k^* réalisant le minimum du PRESS peut entraîner un phénomène de sur apprentissage, ils proposent une statistique de test concluant à des modèles plus parcimonieux. Pour ce faire, ils comparent les PRESS des modèles à $k < k^*$ composantes au PRESS minimal et sélectionnent le plus petit nombre de composantes dont le modèle lié retourne une valeur de PRESS qui n'est pas significativement plus grande que le PRESS minimal. Cette significativité est testée à l'aide d'une loi de Fisher, établi en supposant vérifiées les hypothèses que les erreurs de prédiction soient distribuées suivant une loi gaussienne d'espérance nulle et qu'elles soient mutuellement indépendantes, que ce soit au sein du même modèle ou entre les modèles concurrents. Cette statistique se résume donc de la façon suivante :

$$F_h = \frac{\text{PRESS}_k}{\text{PRESS}_{k^*}} \sim \mathcal{F}_{n,n} \tag{2.46}$$

avec PRESS_{k^*} la valeur du PRESS minimale obtenue sur le modèle à k^* composantes et PRESS_k celle liée au modèle à k composantes avec $k < k^*$. Ainsi la complexité retenue, *i.e.* le nombre de composantes, sera la plus petite valeur de k tel que F_h ne soit plus significatif.

Un second test a été développé par Osten (1988) et repose sur la statistique de test suivante :

$$F_o = \frac{\text{PRESS}_k - \text{PRESS}_{k+1}}{\text{PRESS}_{k+1}} (n - k - 1) \sim \mathcal{F}_{1,n-k-1} \tag{2.47}$$

Dans ce cas, le nombre de composantes retenu sera la plus petite valeur de k tel que F_o ne soit plus significatif.

Ces deux tests ne sont pas satisfaisants, comme le remarque Van der Voet (1994, p.316-317). En effet, la statistique F_h est basée sur des hypothèses qui ne sont pas vérifiées comme l'hypothèse d'indépendance des erreurs de prédictions des deux modèles concurrents. En ce qui concerne le statistique F_o, le fait que celle-ci puisse être négative si le modèle à $k+1$ composantes est lié à une valeur du PRESS supérieur à celle du modèle à k composantes représente une contradiction avec la distribution théorique supposée. Afin de se défaire d'hypothèses fortes sur des distributions et difficilement vérifiables dans le cadre de la PLS, des tests de permutations ont également été élaborés par Van der Voet (1994) et Wiklund *et al.*

(2007).

Un autre critère que nous pouvons citer ici sans trop de détails est le critère nommé H-error (Höskuldsson, 1996) mais que Denham (2000) ne trouvera pas lié à de bonnes performances lors de son étude. On peut également citer les critères de l'AIC (Akaike, 1974), du BIC (Schwarz, 1978) ou encore du C_p de Mallows (Mallows, 1973) qui se sont avérés ne pas être adaptés comme le remarque Höskuldsson (1996). Ceci étant en partie dû au fait que les degrés de liberté de la régression PLS, essentiels à ces critères, n'avaient pas encore été développés à ce moment là. N'ayant pas connaissance d'études comparatives concernant ces critères adaptés aux degrés de liberté développés par Krämer and Sugiyama (2011), nous avons décidé de porter une attention particulière au BIC adapté lors de nos recherches (Ch. 5).

Dans le cadre de l'extension de la régression PLS aux modèles linéaires généralisés, nous avons trouvé peu de critères qui soient spécifiques à ce cadre. Ainsi, nous pouvons citer le critère du minimum du nombre de mal-classés par VC (Rao, 2005, p.312), (Meyer *et al.*, 2010) pour une régression PLS logistique ou un critère, proposé par Bastien *et al.* (2005), basé sur des tests asymptotiques de Wald. Nous reviendrons sur ces critères dans le chapitre 5.

Effectuer un état de l'art exhaustif sur cette problématique de comparaison de modèles, dans notre cas de sélection du nombre optimal de composantes PLS, est en soit un défi à part entière tant le nombre de travaux portant sur ce sujet est important. Nous aurions ainsi pu ajouter les travaux de Marbach and Heise (1990), Höskuldsson (1992), Wakeling and Morris (1993), Holcomb *et al.* (1997), Messick *et al.* (1997), Green and Kalivas (2002) ou encore Lazraq *et al.* (2003). Malgré cela, nous ne pourrions pas prétendre avoir fait un état de l'art complet à ce sujet. Cependant, ce qui ressort de toute cette littérature est qu'aucun critère de sélection ne semble faire l'unanimité au sein de la communauté scientifique et ce malgré les tentatives de comparaison de critères effectuées entre autres par Höskuldsson (1996), Denham (2000), Li *et al.* (2002a), Gourvénec *et al.* (2003), Forina *et al.* (2004) ou encore Gómez-Carracedo *et al.* (2007).

L'élaboration d'un critère fiable de détermination du nombre de composantes reste donc un sujet de recherche ouvert et d'une grande importance tant la détermination de cet hyperparamètre est essentielle, non seulement pour la qualité de modélisation de la régression PLS, mais aussi pour la sélection de prédicteurs.

2.5.2 Objectifs de la thèse

La régression PLS est devenue l'un des procédés de référence pour le traitement des jeux de données obtenus à la suite d'études génomiques. Cependant, comme nous venons de le voir, aucun critère fiable dans la détermination du nombre de composantes ne fait actuellement référence. Or, il s'agit d'un point essentiel à l'établissement de modèles de régression fiables ainsi qu'à la sélection de prédicteurs.

Un premier objectif de cette thèse consiste donc en l'élaboration d'un critère d'arrêt dans

la construction des composantes PLS qui soit lié à certaines propriétés qui nous semblent essentielles. La première a été de ne pas se servir de la VC dû aux problèmes que celle-ci implique et que nous avons évoqué dans la partie précédente 2.5.1. De plus, nous désirions que ce critère ait un aspect universel dans le sens où la méthodologie soit applicable autant à la régression PLS usuelle qu'à ses extensions présentées dans la partie 2.4. Enfin, nous avons considéré comme étant essentiel que ce nouveau critère soit lié à une forte stabilité ainsi qu'à une forte robustesse au bruit aléatoire qui est une caractéristique omniprésente dans les bases de données que nous avons à traiter.

Pour ce faire, nous avons opté pour l'utilisation de techniques de bootstrap. En effet, comme indiqué entre autres par Efron and Tibshirani (1993, p.255) et Kohavi (1995), le bootstrap est une méthode globalement plus stable que la VC. Le bootstrap, comme indiqué dans la partie précédente, a déjà été appliqué afin d'estimer les erreurs de prédiction pour la détermination d'un nombre optimal de composantes. Cependant, cette utilisation du bootstap n'a pas débouché sur le développement d'un critère de référence. Nous avons donc cherché à en effectuer une autre utilisation, à savoir l'obtention d'intervalles de confiance pour les paramètres de la régression de $\mathbf{y}$ sur $\mathbf{T}_k$. En effet, la matrice $\mathbf{T}_k$ dépendant de la réponse, elle peut être considérée comme étant aléatoire. Ainsi, il est possible de considérer les observations $\left(y_i, \mathbf{x}_{(i)}\right)_{1 \leqslant i \leqslant n}$ comme étant des observations indépendantes identiquement distribuées (i.i.d.) suivant une distribution inconnue $\mathcal{F}$ $(p+1)$-dimensionnelle. La technique du bootstrap par paires est particulièrement adaptée à ce type de cas de figure (Efron and Tibshirani, 1993, p.113). Ce nouveau critère que l'on a développé, et que l'on explicitera dans le chapitre 5, permet donc d'éviter l'utilisation de la VC tout en cherchant à approcher la distribution des paramètres de régression, évitant ainsi de se baser sur la statistique du PRESS, afin de déterminer le nombre optimal de composantes. Ce travail de recherche a fait l'objet de l'écriture d'un article constituant le chapitre 5 de cette thèse.

Un second objectif a été d'utiliser ce nouveau critère afin de développer une méthode de sélection des prédicteurs ne nécessitant plus l'utilisation de la VC et qui soit, à nouveau, liée à la notion d'universalité recherchée. L'adaptation de notre nouveau critère de sélection du nombre de composantes à la *Sparse* PLS a également été effectuée. Ces recherches et les résultats obtenus ont fait l'objet de l'écriture d'un second article constituant le chapitre 7 de cette thèse.

A la vue du développement de méthodes basées sur le bootstrap, il nous a paru important de consacrer le prochain chapitre à des rappels sur cette technique avant d'exposer à proprement dit les résultats de recherche liés à cette thèse dans les chapitres suivants.

Chapitre 3

Le bootstrap

Ce chapitre doit être considéré comme une introduction à la méthodologie du bootstrap. Nous ne prétendons en aucun cas y être exhaustif, notamment quant aux propriétés théoriques du bootstrap. Il nous a seulement paru important d'en rappeler les bases afin de fournir aux lecteurs les informations nécessaires à la compréhension des méthodes développées dans les prochains chapitres.

3.1 Motivations

En statistique, un objectif courant est d'évaluer la qualité d'un estimateur $\widehat{\Theta}_n$ de Θ à partir d'un échantillon $X_1, \ldots, X_n$ indépendant identiquement distribué (i.i.d.) de variable aléatoire parente X à travers la détermination de son biais ou de sa variance. Un second objectif repose sur l'établissement d'une probabilité de couverture ou un risque de façon à obtenir un intervalle de confiance (IC) pour la statistique étudiée ou d'effectuer des tests. Pour ce faire, il faut connaître la loi théorique de $\widehat{\Theta}_n$ que l'on note $\mathcal{G}_n$. Or, $\widehat{\Theta}_n$ étant une fonction de l'échantillon de référence *i.e.* $\widehat{\Theta}_n = f(X_1, \ldots, X_n)$, il faut en premier lieu connaître la loi $\mathcal{F}$ portée par la variable aléatoire parente X.

Dans des cadres théoriques classiques et relativement triviaux, tel que dans le cadre gaussien, des résultats théoriques quant aux distributions portant sur des statistiques telles que la moyenne ou la variance de l'échantillon sont connus et permettent l'établissement d'IC. Cependant, en pratique, il est fréquent de ne pas connaître la loi $\mathcal{F}$, empêchant ainsi l'élaboration théorique de la loi $\mathcal{G}_n$. Même dans le cas où $\mathcal{F}$ est connu, entièrement ou à paramètre près, il n'est pas toujours possible d'obtenir $\mathcal{G}_n$ théoriquement au regard de la complexité de la statistique étudiée.

Afin de palier à ces limites, des méthodes computationnelles ont été élaborées ces dernières décennies, reposant sur la force de calcul grandissante du monde informatique. Ces méthodes consistent en la création d'échantillons supplémentaires, obtenus de façon algorithmique, permettant ainsi l'approximation de façon empirique des lois inconnues citées ci-dessus et ouvrant ainsi la possibilité d'élaboration d'IC, d'évaluation de la qualité des estimateurs ou encore la réalisation de tests.

Précisons que ce chapitre est principalement basé sur le livre de référence réalisé par Efron

and Tibshirani (1993) et invitons ainsi les lecteurs désireux d'en savoir plus sur ces méthodes à commencer par la lecture de cette œuvre ainsi que celle de Hall (1992).

3.2 Le *jackknife*

La méthode dit du *jackknife* est relativement ancienne puisque la première trace de cette méthodologie, qui ne portait pas encore ce nom, remonte à un article de 1949 écrit par Quenouille (1949) où il propose l'idée du *jackknife* pour la réduction de biais dans le cadre des séries temporelles. Il développera ensuite cette méthodologie ainsi que ses capacités de réduction du biais en 1956 (Quenouille, 1956) avant que le nom *jackknife* apparaisse vraiment pour la première fois dans Tukey (1958). Formalisons rapidement le procédé du *jackknife*, procédé que nous avons utilisé dans le chapitre 5 afin d'étudier et comparer les distribution du nombre de composantes PLS suivant différents critères d'arrêt.

L'idée originelle du *jackknife*, telle que décrite par Quenouille (1956), est basée sur la séparation de la base de données initiale en q groupes de taille h de telle sorte que $n = qh$. Nous allons cependant nous limiter au cas où $h = 1$, cas le plus utilisé dans la littérature (Miller, 1974).

Dans ce cas, la méthode consiste en la création de n échantillons *jackknife* formés de $n-1$ observations non-répétées et notés :

$$\mathbf{x}_{(i)} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)\,,\ i = 1, \ldots, n \tag{3.1}$$

Posons ainsi $\widehat{\theta}_{(i)} = f\left(\mathbf{x}_{(i)}\right)$ l'estimation de Θ obtenue à l'aide $\mathbf{x}_{(i)}$ et $\widehat{\theta}_n$ celle obtenue à l'aide de l'échantillon complet. Alors l'estimation *jackknife* du biais est définie par : (Efron and Tibshirani, 1993)

$$\hat{b}_{jack} = (n-1)\left(\widehat{\theta}_{(.)} - \widehat{\theta}_n\right) \tag{3.2}$$

avec $\widehat{\theta}_{(.)} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_{(i)}$.

L'estimation *jackknife* de l'écart-type est le suivant :

$$\widehat{se}_{jack} = \left(\frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{\theta}_{(i)} - \widehat{\theta}_{(.)}\right)^2\right)^{1/2} \tag{3.3}$$

Les explications sur l'obtention de ces formulations, et notamment sur la présence du facteur $n-1$ dans ces expressions, sont données par Efron and Tibshirani (1993, p.142).

Un des points intéressants du *jackknife* réside en l'insertion des estimations $\widehat{\theta}_{(i)}$ dans des *pseudo-valeurs* permettant une diminution du biais. Ces *pseudo-valeurs* sont définies de la façon suivante :

$$\widetilde{\theta}_{(i)} = n\widehat{\theta}_n - (n-1)\,\widehat{\theta}_{(i)} \tag{3.4}$$

On obtient ainsi l'estimation *jackknife* de Θ en déterminant la moyenne empirique de ces pseudo-valeurs :

$$\widetilde{\theta}_{(.)} = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\theta}_{(i)} \tag{3.5}$$

L'estimateur correspondant $\widetilde{\Theta}_{(.)}$ est alors, dans des cas précis, lié à un biais réduit. En effet, Quenouille (1956) considère que pour une grande partie des statistiques, le résultat suivant est vérifié :

$$\mathbb{E}\left(\widehat{\Theta}_n - \Theta\right) = \frac{a_1}{n} + \frac{a_2}{n^2} + \frac{a_3}{n^3} + \dots \tag{3.6}$$

Or en développant, on obtient ainsi que :

$$\mathbb{E}\left(\widetilde{\Theta}_{(.)} - \Theta\right) = \frac{a_2}{n^2} - \frac{a_2 + a_3}{n^3} + \dots \tag{3.7}$$

On a donc supprimé le terme d'ordre n^{-1}. Il est alors possible, par des moyens similaires, d'aller plus loin dans la réduction du biais en développant de nouvelles pseudo-valeurs (Quenouille, 1956). La formulation de l'estimation de l'écart-type décrite ci-dessus (3.3) peut également être décrite en fonction des $\widetilde{\theta}_{(.)}$ de la façon suivante :

$$\widehat{se}_{jack} = \left(\frac{1}{n\,(n-1)} \sum_{i=1}^{n} \left(\widetilde{\theta}_{(i)} - \widetilde{\theta}_{(.)} \right)^2 \right)^{1/2} \tag{3.8}$$

Ce résultat nous amène au second aspect important qui a été proposé par Tukey (1958), l'obtention d'IC pour Θ. Pour ce faire, il propose de traiter les estimateurs $\widetilde{\Theta}_{(i)}$ correspondant aux pseudo-valeurs $\widetilde{\theta}_{(i)}$ en les considérant approximativement comme des variables aléatoires indépendantes et identiquement distribuées, amenant ainsi le résultat suivant (Miller, 1974) :

$$\frac{\widetilde{\Theta}_{(.)} - \Theta}{\left(\frac{1}{n(n-1)} \sum\limits_{i=1}^{n} \left(\widetilde{\Theta}_{(i)} - \widetilde{\Theta}_{(.)} \right)^2 \right)^{1/2}} \sim \mathcal{T}_{n-1} \tag{3.9}$$

conduisant ainsi à l'obtention d'IC au seuil α pour Θ de la forme :

$$\left] \widetilde{\theta}_{(.)} - t_{n-1}^{(1-\alpha/2)} \widehat{se}_{jack} \;,\; \widetilde{\theta}_{(.)} + t_{n-1}^{(1-\alpha/2)} \widehat{se}_{jack} \right[\tag{3.10}$$

avec $t_{n-1}^{(1-\alpha/2)}$ le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-1$ degrés de liberté.

Cependant, il a été observé que ces IC n'étaient pas forcément satisfaisants suivant le contexte et les hypothèses de travail. Ainsi, bien que Miller (1974) dresse une liste de problèmes dans lesquels la statistique 3.9 obtenue à l'aide du *jackknife* vérifient bien l'hypothèse de Tukey (1958) quant à sa distribution, dont font partie les problèmes où θ est une fonction des paramètres d'une régression linéaire (Miller *et al.*, 1974), Efron and Tibshirani (1993) nous indique que si la statistique prise en compte n'est pas linéaire, il y a perte d'information ce qui peut rendre le *jackknife* inefficace.

3.3 Le bootstrap, une extension du *jackknife*

3.3.1 Historique et méthodologie

Le bootstrap est un procédé de ré-échantillonage introduit et proposé par Efron (1979). Ce procédé a pour but de généraliser la méthodologie du *jackknife* au sens suivant. Nous

avons vu que le *jackknife* permet de s'intéresser et d'analyser la distribution liée à la variable aléatoire

$$R\left(X, \mathcal{F}\right) = \widehat{\Theta}_n\left(X\right) - \Theta\left(\mathcal{F}\right) \tag{3.11}$$

notamment à travers son espérance *i.e.* le biais lié à $\widehat{\theta}_n$. Dans un second temps, une méthodologie décrite dans le paragraphe précédent (3.9) a également été développée afin de permettre l'obtention d'IC bien que celle-ci soit basée sur des hypothèses de loi qui n'ont pas de raisons théoriques d'être systématiquement vérifiées, si ce n'est dans le cas de grands échantillons (Efron, 1979).

Ainsi, Efron a développé cette technique du bootstrap qui cette fois-ci n'est plus liée à une variable aléatoire $R\left(X, \mathcal{F}\right)$ précise comme pour le *jackknife* et permet ainsi de traiter une quantité de problématique bien plus large, et ne nécessitant plus d'hypothèse sur les lois comme c'est le cas pour l'obtention d'IC avec le *jackknife*. Le *jackknife* réalise en réalité une approximation linéaire du bootstrap, ce point étant détaillé notamment par Efron and Tibshirani (1993, p.145) et Efron (1979)

Rappelons que l'objectif est d'approcher la loi $\mathcal{G}_n$ de $\widehat{\Theta}_n$. Pour ce faire, le principe du bootstrap consiste à approcher par simulation de Monte Carlo cette loi. Précisons tout de même qu'initialement Efron (1979) proposait deux autres méthodes, mais avec la forte augmentation de puissance de calcul liée à l'informatique, cette méthodologie est devenue la plus courante. Ainsi, globalement trois situations sont possibles :

1. $\mathcal{F}$ est entièrement connue.
2. $\mathcal{F}$ est partiellement connue dans le sens où $\mathcal{F}$ appartient à une famille de lois connues mais dépend d'un ou plusieurs paramètres inconnus à estimer.
3. $\mathcal{F}$ est totalement inconnue.

Dans le premier cas, il suffit donc d'approcher numériquement $\mathcal{G}_n$ de la façon suivante : Pour $b = 1, \ldots, B$:

- Simuler un échantillon de n observations issues de $\mathcal{F}$ et noté $\left(x_1^{(b)}, \ldots, x_n^{(b)}\right)$
- Déterminer $\widehat{\theta}_n^{(b)} = f\left(x_1^{(b)}, \ldots, x_n^{(b)}\right)$

Ainsi, les B variables aléatoires $\widehat{\Theta}_n^{(1)}, \ldots, \widehat{\Theta}_n^{(B)}$ sont i.i.d., distribuées comme $\widehat{\Theta}_n$. Ensuite, à partir de cette échantillon de taille B, on peut déterminer la fonction de répartition empirique de $\widehat{\Theta}_n$ de la façon suivante :

$$\mathcal{G}_{n,B}\left(t\right) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left(\widehat{\Theta}_n^{(b)} \leqslant t\right) \tag{3.12}$$

Puis en vertu du théorème de Glivenko-Cantelli, on sait qu'il y a convergence uniforme de la fonction de répartition empirique vers la fonction de répartition théorique. De même, il est possible d'approcher la moyenne de $\widehat{\theta}_n$, sa variance ou encore ses quantiles.

Cependant, en pratique, ce cas est rarement établi. On se trouve plus fréquemment dans une situation où $\mathcal{F}$ n'est que partiellement connue, nécessitant l'utilisation du bootstrap dit *paramétrique*, ou, le plus souvent, totalement inconnue et nécessitant alors l'utilisation du bootstrap *non paramétrique*. Nous allons détailler ce dernier.

L'idée du bootstrap non paramétrique consiste à utiliser une approximation connue de $\mathcal{F}$, à savoir sa fonction de répartition empirique $\mathcal{F}_n$, à la place de $\mathcal{F}$.

Definition 3.3.1. *On appelle échantillon bootstrap de* $X \sim \mathcal{F}$*, un échantillon* $\left(X_1^{(b)}, \ldots, X_n^{(b)}\right)$ *suivant la loi* $\mathcal{F}_n$*.*

On obtient alors une réalisation $\left(x_1^{(b)}, \ldots, x_n^{(b)}\right)$ de cet échantillon en effectuant n tirages avec remise parmi les n observations initiales $(x_1, \ldots, x_n)$, puisqu'ainsi chaque observation a une probabilité $1/n$ d'être tirée. On détermine ensuite l'estimation $\widehat{\theta}_n^{(b)}$ liée à l'échantillon bootstrap obtenu. Voici l'algorithme qui en découle :

Soit $(x_1, \ldots, x_n)$ une réalisation fixée de $(X_1, \ldots, X_n)$ avec $\mathcal{F}$ inconnue.

1. Pour $b = 1, \ldots, B$:
   - Tirer un échantillon $\left(x_1^{(b)}, \ldots, x_n^{(b)}\right)$ avec remise dans $(x_1, \ldots, x_n)$.
   - Calculer $\widehat{\theta}_n^{(b)} = f\left(x_1^{(b)}, \ldots, x_n^{(b)}\right)$.
2. Récupérer l'échantillon $\left(\widehat{\theta}_n^{(1)}, \ldots, \widehat{\theta}_n^{(B)}\right)$.
3. Répondre à la problématique posée liée à l'estimateur, que ce soit la détermination d'un biais, de sa variance, d'un IC ou autres à l'aide de cet échantillon.

L'obtention d'informations liées à $\mathcal{F}$ en utilisant sa fonction de répartition associée $\mathcal{F}_n$ est le principe dit du *plug-in*. Le bootstrap peut donc être considéré comme une application directe de ce principe (Efron and Tibshirani, 1993).

Précisons enfin que des résultats théoriques asymptotiques soutiennent la pratique du bootstrap (Bickel and Freedman, 1981) (Singh, 1981) (Hall, 1992) (Mammen, 2012).

3.3.2 Le bootstrap en régression linéaire

Nous venons d'introduire la méthodologie du bootstrap dans le paragraphe précédent. Pour ce faire, nous l'avons basée sur l'application d'un simple échantillon dans le but d'obtenir des informations sur une statistique précise. Cependant, la technique du bootstrap a été étendue à un domaine qui nous intéresse tout particulièrement dans le cadre de cette thèse, la régression linéaire.

Dans le cadre des modèles linéaires, deux méthodologies faisant intervenir le bootstrap sont largement décrites et utilisées afin d'obtenir des informations sur des statistiques dépendant des paramètres de régression à estimer. Chacune de ces deux méthodes est liée à un type de modèle différent, dépendant des conditions que l'on suppose vérifiées sur le modèle

en question.

La première méthode, décrite dès l'introduction du bootstrap dans l'article de Efron (1979), est liée à un modèle dit de *régression* (Freedman, 1981). Pour ce type de modèle, la matrice des régresseurs $\mathbf{X}$ est supposée fixe et les erreurs sont supposées indépendantes, provenant d'une même loi inconnue $\mathcal{F}$ d'espérance nulle, et homoscédastiques *i.e.* de variances identiques. Afin d'approcher cette distribution inconnue, on va donc appliquer le principe du bootstrap sur le vecteur des résidus obtenus à l'aide de la régression linéaire effectuée sur les données originelles. Formalisons cela.

Posons $\mathbf{y} = \mathbf{X}\beta + \epsilon$ le modèle de régression considéré avec $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathcal{M}_{n,p}(\mathbb{R})$ considéré fixe et $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ où $\epsilon_i \underset{i.i.d.}{\sim} \mathcal{F}$ d'espérance nulle. Le modèle probabiliste possède donc deux composantes inconnues, β et $\mathcal{F}$. L'estimation de β se fait, le plus communément, en résolvant les équations normales issues du procédé des moindres carrées, à savoir :

$$\widehat{\beta}^{MCO} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{3.13}$$

À partir de cette estimation, il est alors possible d'obtenir les résidus liés :

$$\widehat{\epsilon} = (\widehat{\epsilon_1}, \ldots, \widehat{\epsilon_n}) = \mathbf{y} - \mathbf{X}\widehat{\beta}^{MCO} \tag{3.14}$$

Il s'agit ensuite d'approcher $\mathcal{F}$ par la fonction de répartition empirique liée aux $\widehat{\epsilon_i}$. Pour ce faire, on tire aléatoirement et avec remise dans $\widehat{\epsilon}$ un échantillon $\epsilon^{(b)}$ de taille n et on détermine un échantillon bootstrap de la réponse de la façon suivante :

$$\mathbf{y}^{(b)} = \mathbf{X}\widehat{\beta}^{MCO} + \epsilon^{(b)} \tag{3.15}$$

Enfin, il s'agit d'obtenir l'estimation de β sur ce nouvel échantillon bootstrap de la façon suivante :

$$\widehat{\beta}^{(b)} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}^{(b)} \tag{3.16}$$

Ensuite, à partir de ces différentes estimations bootstrap de β, il est alors possible d'étudier des fonctions dépendantes de ce vecteur de paramètres ou d'obtenir des intervalles de confiance. Cependant, ce procédé est basé sur des hypothèses assez fortes qui, dans notre cadre d'étude, sont difficilement envisageables à chaque étape de construction d'une composante PLS. Ainsi, si l'homoscédasticité n'est pas vérifiée, l'estimateur bootstrap de la variance des paramètres est généralement biaisé (Wu, 1986). Il nous a donc paru plus raisonnable d'utiliser la seconde méthode d'adaptation du bootstrap aux techniques de régression, le *bootstrap par paires*.

Le bootstrap par paires a été introduit par (Freedman, 1981) et est adapté aux modèles dits de *corrélation*. Dans le cadre de ces modèles, on suppose alors $\mathbf{X}$ aléatoire et ϵ pouvant être lié à $\mathbf{X}$. Notons alors $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^T = (\mathbf{y}, \mathbf{X})$ de telle sorte que $\mathbf{z}_i = (y_i, x_{i1}, x_{i2}, \ldots, x_{ip})$.

La procédure consiste alors à tirer aléatoirement et avec remise dans $\mathbf{Z}$ de nouveaux échantillons bootstrap $\mathbf{Z}^{(b)}$ et d'en déterminer, pour chacun, l'estimation liée de la statistique $S(\mathbf{Z})$ étudiée (Fig. 3.1), dans notre cas $S(\mathbf{Z}) = \beta$ le vecteur des paramètres de régression linéaire liée à la base de données bootstrap obtenue. Ainsi, l'algorithme utilisé est le suivant :

- Pour $b = 1, \ldots, B$:
  1. Tirer aléatoirement et avec remise dans $\mathbf{Z}$ un échantillon $\mathbf{Z}^{(b)} = \left(\mathbf{z}_1^{(b)}, \ldots, \mathbf{z}_n^{(b)}\right)^T = \left(\mathbf{y}^{(b)}, \mathbf{X}^{(b)}\right)$.
  2. Déterminer $\widehat{\beta}^{(b)} = \left(\mathbf{X}^{(b)^T}\mathbf{X}^{(b)}\right)^{-1}\mathbf{X}^{(b)^T}\mathbf{y}^{(b)}$
- Répondre à la problématique liée à β à l'aide des réplications bootstrap $\left(\widehat{\beta}^{(1)}, \ldots, \widehat{\beta}^{(B)}\right)$

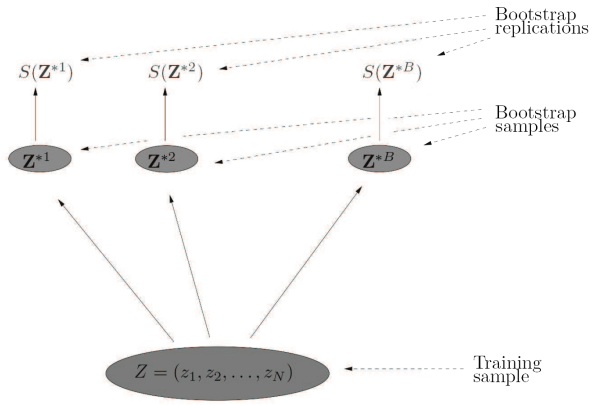


FIGURE 3.1 – Processus générique du bootstrap.

Ainsi, contrairement au bootstrap résiduel, cette procédure de bootstrap par paires repose uniquement sur la condition que $\mathbf{z}_i \underset{i.i.d.}{\sim} \mathcal{F}$ avec $\mathcal{F}$ une distribution sur $\mathbb{R}^{p+1}$. On peut ainsi qualifier cette procédure de *model-free* (Hastie *et al.*, 2009, ch.8) puisque n'utilisant que les lignes de la base de données originale et à aucun moment un modèle hypothétique n'est posé et utilisé. Cette grande souplesse dans l'utilisation de ce procédé nous a permis de pouvoir l'adapter à tout type de situation dans le cadre de la régression PLS.

De plus, des adaptations de ces procédures aux modèles linéaires généralisés ont été développées et initialement proposées par Moulton and Zeger (1991). Ainsi, les procédures développées et exposées dans les chapitres à venir ont pu être appliquées aussi bien au cadre de la régression PLS usuelle qu'à son extension aux modèles linéaires généralisés.

3.3.3 Les différents types d'intervalle de confiance

La finalité de l'utilisation du bootstrap dans le cadre de cette thèse réside en l'obtention d'IC des paramètres de régression liés à l'algorithme PLS. Nous allons donc rapidement rappeler les méthodes usuelles d'obtention de tels intervalles à l'aide des réplications bootstrap obtenues à travers le bootstrap par paires. Précisons tout de même qu'afin de faciliter les notations, nous allons nous placer dans le cadre d'un échantillon simple noté $\mathbf{x} = (x_1, \ldots, x_n)$. Cependant, toutes les procédures exposées dans la suite sont naturellement adaptables à l'obtention d'IC pour des coefficients de régression.

La méthode la plus classique d'obtention d'un IC pour un paramètre θ est celle reposant sur l'hypothèse de normalité de la loi associée à l'estimateur de ce paramètre (Efron, 1985). Autrement dit, l'hypothèse suivante est posée :

$$\widehat{\Theta}_n \sim \mathcal{N}\left(\Theta, \sigma^2\right) \tag{3.17}$$

Cette hypothèse implique ainsi que :

$$\mathcal{Z} = \frac{\widehat{\Theta}_n - \Theta}{\sigma} \sim \mathcal{N}\left(0, 1\right) \tag{3.18}$$

Il est alors possible d'estimer les bornes d'un IC au niveau $1 - 2\alpha,\ \alpha \in [0, 1/2]$ pour Θ de la façon suivante :

$$\Theta \in \left]\widehat{\theta}_n - z^{(1-\alpha)} \times \widehat{\sigma}, \widehat{\theta}_n + z^{(1-\alpha)} \times \widehat{\sigma}\right[\tag{3.19}$$

avec $z^{(1-\alpha)}$ le quantile d'ordre $(1 - \alpha)$ de la loi Normale centrée réduite.

Cette formulation est extrêmement utile en pratique mais correspond en réalité à un cadre asymptotique ou dit de *large-sample* (Efron and Tibshirani, 1993, p.154). Ainsi, dans le cadre de petits échantillons, il est préférable d'utiliser les quantiles d'une loi de Student en lieu et place des quantiles de la loi Normale. En effet, la loi de Student représente une meilleure approximation de la loi de cette statistique 3.18 puisque tenant compte du fait que l'écart-type σ est inconnu et donc à estimer. Ainsi, ces quantiles, bien qu'asymptotiquement proches de ceux d'une loi Normale centrée-réduite, permettent un élargissement des IC dans le cadre de petits échantillons afin de tenir compte du fait que σ soit inconnu (Efron and Tibshirani, 1993, p.158-159). Cependant, ce cadre reste trop restrictif et les hypothèses liées sont trop fortes pour que de telles méthodes soient applicables dans des cas plus généraux. Par exemple, ce procédé ne tient pas compte d'une éventuelle asymétrie de la distribution sous jacente. La technique du bootstrap a ainsi permis de considérablement améliorer l'obtention d'IC et ce à travers plusieurs développements.

Commençons par l'adaptation du bootstrap à la procédure décrite ci-dessus. Celle-ci a été proposée par Efron (1981) et consiste en une généralisation de 3.18 dans le sens où la distribution de $\mathcal{Z}$ n'est pas supposée connue à l'avance mais approchée par bootstrap. Pour ce faire, le procédé suivant est mis en place :

- Pour $b = 1, \ldots, B$:
  1. Tirer aléatoirement et avec remise dans $\mathbf{x}$ un échantillon $\mathbf{x}^{(b)} = \left(x_1^{(b)}, \ldots, x_n^{(b)}\right)$.
  2. Déterminer $\mathcal{Z}^{(b)} = \frac{\widehat{\theta}_n^{(b)} - \widehat{\theta}_n}{\widehat{\sigma}^{(b)}}$, avec $\widehat{\theta}_n = f(\mathbf{x})$, $\widehat{\theta}_n^{(b)} = f(\mathbf{x}^{(b)})$ et $\widehat{\sigma}^{(b)}$ l'écart-type de $\widehat{\theta}_n^{(b)}$ estimé à l'aide de $\mathbf{x}^{(b)}$.
- Posons $E = \left\{\mathcal{Z}^{(1)}, \ldots, \mathcal{Z}^{(B)}\right\}$. Déterminer le $\alpha^{\text{ème}}$ et le $(1-\alpha)^{\text{ème}}$ percentile de E, notés respectivement $\widehat{t}^{(\alpha)}$ et $\widehat{t}^{(1-\alpha)}$, définis de telle sorte que :

$$\#\left\{\mathcal{Z}^{(b)} \leqslant \widehat{t}^{(\alpha)}\right\}/B = \alpha \tag{3.20}$$

$$\#\left\{\mathcal{Z}^{(b)} \leqslant \widehat{t}^{(1-\alpha)}\right\}/B = 1 - \alpha \tag{3.21}$$

- Finalement, on retourne l'IC suivant :

$$\Theta \in \left]\widehat{\theta}_n - \widehat{t}^{(1-\alpha)} \times \widehat{\sigma}, \widehat{\theta}_n - \widehat{t}^{(\alpha)} \times \widehat{\sigma}\right[\tag{3.22}$$

Précisons que si aucun estimateur n'est connu pour $\widehat{\sigma}^{(b)}$, une deuxième procédure de bootstrap est à inclure consistant, pour chaque échantillon $\mathbf{x}^{(b)}$, à effectuer à nouveau des tirages aléatoires avec remise dans $\mathbf{x}^{(b)}$ afin d'approcher l'écart-type de chacun des $\widehat{\theta}_n^{(b)}$. Nous parlons alors d'un algorithme de double bootstrap (Hall, 1992). De plus, dans le cas où $B\alpha$ n'est pas un entier, des précisions sont données Efron and Tibshirani (1993, p.160) quant au choix des percentiles.

Ce procédé permet l'obtention d'IC non-symétrique autour de $\widehat{\theta}_n$, contrairement aux deux procédures usuelles citées précédemment. Cette capacité représente une importante amélioration en terme de taux de couverture (Efron and Tibshirani, 1993, p.161). Cependant, dans certains cas, travailler sur une fonction monotone du paramètre d'intérêt améliore la qualité des IC obtenus à l'aide de cette procédure. Par exemple, lorsque le paramètre d'intérêt est le coefficient de corrélation de Pearson ρ lié à une distribution normale bivariée, la transformation :

$$\phi(\rho) = \operatorname{argtanh}(\rho) = 1/2 \times \log\left(\frac{1+\rho}{1-\rho}\right) \tag{3.23}$$

améliore la qualité de l'IC obtenu puisque normalisant et stabilisant la variance de ce nouveau paramètre ϕ (Efron and Tibshirani, 1993, p.163), (Efron, 1981, p.148). Le problème qui en ressort est que la procédure développée et ses performances sont dépendantes de ces transformations, qui ne sont évidemment pas connues dans la majorité des cas.

Ainsi, une seconde procédure a été établie par Efron (1981), il s'agit des IC basés sur les percentiles de la distribution des réplications bootstrap. Une fois l'ensemble $\left\{\widehat{\theta}_n^{(b)}\right\}_{1\leqslant b\leqslant B}$ déterminé, l'IC retenu est le suivant :

$$\Theta \in \left]\widehat{\theta}_n^{(\alpha)}, \widehat{\theta}_n^{(1-\alpha)}\right[\tag{3.24}$$

avec $\widehat{\theta}_n^{(\alpha)}$ et $\widehat{\theta}_n^{(1-\alpha)}$ les quantiles de la distribution bootstrap du paramètre d'ordre respectivement α et $1-\alpha$.

Cette procédure relativement simple possède cependant la propriété intéressante suivante :

Proposition 3.3.1. *(Efron and Tibshirani, 1993, p.173)*

Supposons que la transformation $\widehat{\phi} = m\left(\widehat{\Theta}\right)$ *normalise parfaitement la distribution de* $\widehat{\Theta}$ *i.e. :*

$$\widehat{\phi} \sim \mathcal{N}\left(\phi, \sigma^2\right), \tag{3.25}$$

alors l'intervalle percentile basé sur $\widehat{\Theta}$ *est égal à :*

$$\left[m^{-1}\left(\widehat{\phi} - z^{(1-\alpha)}\sigma\right), m^{-1}\left(\widehat{\phi} - z^{(\alpha)}\sigma\right)\right] \tag{3.26}$$

En d'autres termes, cette procédure n'est pas dépendante d'une quelconque échelle et détermine automatiquement la transformation correcte, si elle existe. Cependant, cette procédure est tout de même encore soumise à une hypothèse forte à savoir que l'estimateur considéré doit être sans biais (Efron and Tibshirani, 1993, p.176).

Ainsi, afin de palier à cette limite, deux autres extensions ont été proposées. La première, nommée *bias corrected percentile method* (BC) a été proposée par Efron (1981) et notamment étudiée plus en détail par Efron (1985). La seconde est nommée *bias-corrected and accelerated percentile method* (BC_a) et a été introduite par Efron (1987). La première représentant un cas particulier de la seconde, nous allons nous contenter d'exposer brièvement la méthode BC_a qui est celle que nous avons utilisée tout au long de nos recherches afin d'obtenir des IC.

L'idée du bootstrap BC_a consiste à supposer l'existence d'une transformation $\widehat{\phi} = m\left(\widehat{\Theta}\right)$ strictement croissante et inversible ainsi que deux réels a et z_0 tel que :

$$\frac{\widehat{\phi} - \phi}{1 + a\phi} + z_0 \sim \mathcal{N}(0, 1) \tag{3.27}$$

Dans ce cadre là, a est appelé le paramètre d'*accélération* et z_0 la paramètre de *correction du biais* (Efron and Tibshirani, 1993). De cette formalisation, découlent des quantiles d'ordre différents de ceux utilisés pour la procédure du bootstrap percentile usuel où il s'agit simplement de choisir les quantiles $\widehat{\theta}_n^{(\alpha)}$ et $\widehat{\theta}_n^{(1-\alpha)}$ de la distribution bootstrap obtenue d'ordre respectivement α et $1 - \alpha$. Il s'agit donc de déterminer ces quantiles qui vont être choisis comme valeurs limites de l'IC au niveau $1 - 2\alpha$. Ainsi, dans le cadre du bootstrap BC_a pour l'obtention d'un IC au niveau $1 - 2\alpha$, on va obtenir l'IC suivant :

$$\Theta \in \left]\widehat{\theta}_n^{(\alpha_1)}, \widehat{\theta}_n^{(\alpha_2)}\right[\tag{3.28}$$

avec :

$$\alpha_1 = \Phi\left(\widehat{z}_0 + \frac{\widehat{z}_0 + z^{(\alpha)}}{1 - \widehat{a}\left(\widehat{z}_0 + z^{(\alpha)}\right)}\right) \tag{3.29}$$

$$\alpha_2 = \Phi\left(\widehat{z}_0 + \frac{\widehat{z}_0 + z^{(1-\alpha)}}{1 - \widehat{a}\left(\widehat{z}_0 + z^{(1-\alpha)}\right)}\right) \tag{3.30}$$

Φ et $z^{(1-\alpha)}$ représentent ici respectivement la fonction de répartition et le quantile d'ordre $(1 - \alpha)$ de la loi Normale centrée réduite. Enfin, en ce qui concerne les estimations des paramètres a et z_0 ainsi que les résultats théoriques liés au bootstrap BC_a, des précisions sont fournies notamment par Efron and Tibshirani (1993, ch.14-22) ainsi que par Efron (1987) et Shao and Tu (2012). Cette méthode possède ainsi les avantages du bootstrap percentile usuel tout en gérant la possibilité d'un biais de l'estimateur.

Précisons enfin que le bootstrap BC correspond au cas du bootstrap BC_a avec $a = 0$ (Efron, 1987).

Deuxième partie

Un nouveau critère d'arrêt basé sur le bootstrap pour la sélection du nombre de composantes

Chapitre 4

Résumé des travaux entrepris

4.1 Motivations

Le nombre de composantes PLS est un hyperparamètre fondamental de cette méthode. Comme abordé dans le chapitre 2, ce nombre traduit le dimensionnement de l'espace dans lequel la réponse va être projetée. Les estimations des paramètres de régression seront ainsi directement impactées par la détermination de cet hyperparamètre. De Jong (1995) établit ainsi un résultat dans ce sens (Prop. 2.22). Retenir un nombre de composantes inférieur au nombre optimal revient à une perte d'information et peut ainsi mener à l'établissement d'un modèle lié à une mauvaise représentation de certaines observations et/ou variables. Retenir un nombre de composantes supérieur au nombre optimal conduit à l'obtention d'un modèle sur-paramétré induisant généralement de mauvaises performances prédictives. (Wiklund *et al.*, 2007)

A titre d'exemple, nous avons représenté sur la Figure 4.1, l'incidence du nombre de composantes sur l'estimation des paramètres β^{PLS} du modèle.

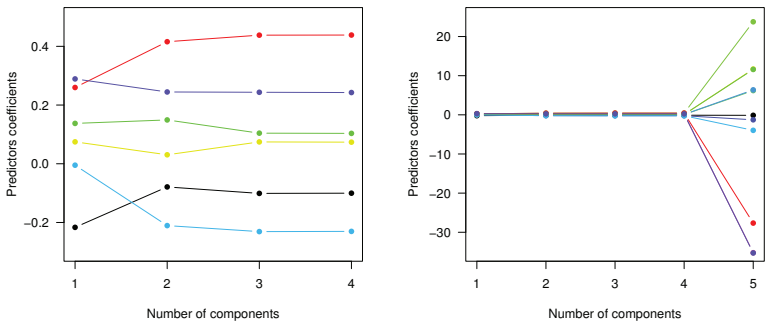


FIGURE 4.1 – Évolution des estimations des coefficients de régression β^{PLS} en fonction du nombre de composantes. Le graphique à gauche représentant un agrandissement de celui de droite sur les 4 premières composantes.

Ces graphiques ont été produits à l'aide d'une base de données obtenue par simulation

(Annexe A), d'après le même procédé que celui décrit dans le prochain chapitre, section 5.3.2. Le nombre de composantes optimal est égal à trois et la matrice des prédicteurs $\mathbf{X}$ a été construite à partir de quatre composantes. Cette représentation graphique, traduit, sur un exemple précis, l'importance capitale de l'utilisation d'un critère d'arrêt fiable dans la construction des composantes. Sans entrer dans les détails, ne s'agissant que d'un exemple parmi d'autres, la forte augmentation de la norme du vecteur $\widehat{\beta}^{PLS}$ lié au modèle à cinq composantes est intuitivement explicable. En effet, dû au fait que l'entière quantité d'information structurée dans $\mathbf{X}$ soit concentrée, sur cette simulation, dans un espace de dimension quatre, la cinquième composante se construit uniquement à l'aide du bruit aléatoire de faible variance introduit lors de cette simulation dans la matrice des prédicteurs $\mathbf{X}$. De part la méthodologie de la régression PLS consistant à maximiser la covariance entre $\mathbf{y}$ et ses composantes, il en ressort une cinquième composante à très faible variance mais possédant une corrélation avec $\mathbf{y}$ non négligeable, corrélation que l'on peut qualifier d'artificielle puisque n'étant constituée que de bruit aléatoire. Le coefficient de régression c_5 lié à cette cinquième composante sera donc disproportionné causant ainsi l'instabilité des estimations $\widehat{\beta}^{PLS}$ (Tab. 4.1). Rappelons qu'il ne s'agit pas ici de démontrer quoi que ce soit mais simplement de fournir un exemple, certes quelque peu extrême, permettant de représenter l'impact que peut avoir une mauvaise détermination de cet hyperparamètre qu'est le nombre de composantes pour une régression PLS.

TABLE 4.1 – Exemple : Résumé numérique des paramètres de composantes obtenus.

	$\mathbf{t}_1$	$\mathbf{t}_2$	$\mathbf{t}_3$	$\mathbf{t}_4$	$\mathbf{t}_5$
Variance	5.216	2.584	1.871	0.634	8.810×10^{-6}
$\text{Corr}(\mathbf{y}, \mathbf{t}_i)$	0.761	0.342	0.087	0.001	0.107
Coefficient c_i	0.333	0.213	0.064	0.002	36.211

La majorité des critères d'arrêt existants sont basés sur la validation croisée (VC). Bien qu'étant une méthode très répandue (Hastie *et al.*, 2009, ch.7), celle-ci est à utiliser avec précaution et peut-être liée à d'importants problèmes, notamment quant à la forte variabilité des résultats obtenus suivant la statistique utilisée et basée sur la VC. Boulesteix (2014) a notamment mentionné des problèmes liés à l'utilisation de cette méthode, problèmes que nous avons pu vérifier très rapidement sur certaines bases de données et que nous évoquons dans le chapitre suivant (5.6.1).

Comme évoqué dans la partie 2.5.1, ne pouvant considérer un critère existant comme étant globalement satisfaisant, l'objectif premier de cette thèse a donc été d'établir un nouveau critère d'arrêt dans la construction des composantes PLS. Ce nouveau critère se devait de répondre à certaines problématiques particulières que nous nous sommes fixées. Tout d'abord, il s'agissait de développer un critère ne nécessitant pas l'intervention de la VC afin d'éviter les problèmes inhérents à cette procédure. Ensuite, afin de faciliter sa future application en routine, nous désirions que ce nouveau critère soit en quelque sorte universel, dans le sens où il soit applicable, avec des propriétés similaires, à la PLS ainsi qu'à son extension aux modèles linéaires généralisés. Un troisième élément essentiel résidait dans l'établissement d'un critère de faible variabilité. En effet, être confronté à un critère qui, exécuté plusieurs fois sur la même base de données, retourne un nombre conséquent de déterminations différentes s'avère

très problématique. Nous avons ainsi pu observer ce phénomène de façon concrète suite à l'utilisation de critères basés sur la VC (Figures 5.16 et 5.18). Enfin, une des particularités de nos bases de données réside dans la présence de bruit, que ce soit dans la réponse étudiée ou dans les données disponibles au sein de la matrice des prédicteurs. Un exemple réside dans l'analyse même des puces. En effet, pour visualiser les endroits où la sonde s'est hybridée, il est possible d'obtenir, par excitation à l'aide d'un rayon laser, une image par fluorescence qui est captée à l'aide d'une caméra. Ainsi, beaucoup de sources d'erreurs sont possibles à ce niveau là telle une saturation du signal si l'activation est trop forte ou une perte si l'activation est trop faible. De plus, cette fluorescence faiblit en fonction du temps et de son exposition à la lumière notamment. Un exemple d'image, extrait du livre de Gibson and Muse (2004), obtenue sur une puce à ADN avec présence de bruit est représenté sur la Figure 4.2. Gibson and Muse (2004) nous précise ainsi que différents scanners testant les mêmes échantillons retournent des résultats différents. D'autres exemples de bruit, tel qu'un mauvais choix des échantillons témoins, sont mentionnés par Gibson and Muse (2004, ch.3). Le dernier objectif que nous nous soyons fixé réside donc dans le développement d'un critère d'arrêt qui soit robuste à différents niveaux de bruits, que ce soit dans $\mathbf{y}$ ou dans $\mathbf{X}$.

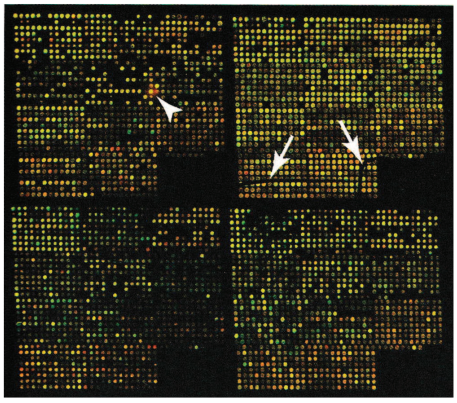


FIGURE 4.2 – Exemple de problèmes fréquents : La pointe de la tête de flèche à gauche cible un bruit de fond localement fort, les deux flèches à droite pointent vers des rayures (pas de signal).

4.2 Le bootstrap comme alternative à la validation croisée

Dans le cadre de la régression PLS, aucune hypothèse sur les distributions, à priori, n'est imposée. Comme exposé dans le chapitre 3, les techniques de bootstrap peuvent ainsi servir à palier ce manque en approchant les distributions inconnues de façon empirique. L'utilisation du bootstrap est bien connue dans le cadre des modèles de régression, comme expliqué à la section 3.3.2. A travers l'obtention d'IC bootstrap pour les paramètres de régression, cela permet, de façon équivalente, de tester la significativité d'une variable $\mathbf{x}_j$ en testant l'hypothèse nulle $\mathcal{H}_0 : \beta_j = 0$. Autrement dit, en établissant un IC au niveau $1 - \alpha$, si 0 n'appartient pas à cet IC, on est en mesure de rejeter $\mathcal{H}_0$ avec un risque d'erreur α. Plus

de précisions sur les liens entre l'établissement d'IC et les tests d'hypothèse sont fournis par Efron and Tibshirani (1993, p.156).

La régression PLS consistant en une succession de régressions linéaires, l'adaptation de méthode du bootstrap pour l'obtention d'un nouveau critère d'arrêt a été envisagée. La version du bootstrap par paires a été sélectionnée, principalement pour deux raisons. Tout d'abord, il est difficilement envisageable de supposer l'homoscédasticté des erreurs liées à chacun des modèles construit successivement par la régression PLS, critère essentiel à l'utilisation du bootstrap résiduel. De plus, les composantes $\mathbf{t}_k$ dépendant de $\mathbf{y}$, celles-ci définissent un sous-espace aléatoire et non pas fixe comme le note également Blazere (2015, p.130). Le bootstrap par paires est donc particulièrement adapté à ce type de situation, comme le précise Efron and Tibshirani (1993, p.113). Ainsi, l'idée a été d'utiliser cette technique afin de tester la significativité d'une nouvelle composante par rapport à $\mathbf{y}$ et par rapport à $\mathbf{X}$, ceci étant réalisé à travers l'établissement d'IC bootstrap lors des différentes phases de régressions linéaires concernant les composantes impliquées dans l'algorithme PLS. Ainsi, nous définissons une composante comme étant significative si elle l'est à la fois pour $\mathbf{y}$ et pour au moins un prédicteur $\mathbf{x}_j$. Le procédé développé va ainsi se dérouler jusqu'à obtention d'une composante non-significative. Le détail de cette procédure est explicité dans le prochain chapitre, section 5.2.3.

Une motivation supplémentaire d'utilisation du bootstrap réside dans le fait qu'il a été observé que le bootstrap est une méthode liée à une faible variabilité comparativement à la VC (Kohavi, 1995). Enfin, l'adaptation du bootstrap par paires aux modèles de régressions généralisés, proposé et détaillé par Moulton and Zeger (1991), nous a permis d'adapter notre nouveau critère au cadre de l'extension de la régression PLS aux modèles généralisés, atteignant ainsi l'objectif d'universalité recherché.

Plus de détails sur cette procédure sont donnés dans le chapitre 5.

4.3 Méthodologie

Afin de mettre à l'épreuve ce nouveau critère d'arrêt, nous l'avons comparé, sur différents points, à certains critères existants. Ces critères sont, dans le cadre d'une régression PLS ordinaire, ceux du Q^2 ainsi que le BIC adapté aux degrés de libertés développés par Krämer and Sugiyama (2011). Nous avons sélectionné le critère du Q^2 en tant que critère de référence. En effet, il s'agit d'un critère bien décrit, utilisé fréquemment et implanté dans le logiciel *R* (R Development Core Team, 2008), notamment dans le package *plsRglm* (Bertrand *et al.*, 2014), que nous avons utilisé dans le cadre de cette thèse ainsi que dans SIMCA-P, logiciel développé par l'équipe de S.Wold (Umetrics, 2005). Le critère du BIC adapté a été inclus dans l'étude car, à notre connaissance, aucune étude comparative liée à ce critère n'a été menée depuis son développement par Krämer and Sugiyama (2011). Il s'agissait là d'une opportunité afin de tester ses performances. Dans le cadre des extensions des régressions PLS aux modèles linéaires généralisés, nous avons sélectionnés le critère du nombre de mal-classés par VC pour une régression PLS logistique (Rao, 2005) (Meyer *et al.*, 2010) ainsi qu'un critère d'arrêt développé par Bastien *et al.* (2005). Les critères de l'AIC et du BIC usuel ont également été étudiés bien que ne possédant pas les degrés de liberté dans ce cadre des

modèles généralisés. Les détails sur ces critères sont disponibles dans le prochain chapitre, section 5.3.1. Précisons que dans le cadre des régressions généralisées, nous nous sommes limités au cas des lois binomiale et de Poisson.

N'ayant pas connaissance d'études précises menées quant à l'impact du bruit sur la précision des critères d'arrêt dans la construction de composantes PLS, nous avons donc réalisé d'intensives phases de simulation afin de suivre l'évolution de la qualité de ces différents critères en fonction de différents niveaux de bruits. Nous avons jugé de cette qualité à travers différentes caractéristiques. La première étant le biais observé lié à ces critères. En effet, le nombre de composantes à extraire étant connu, nous avons pu observer l'évolution de ce biais en fonction de l'évolution du bruit introduit dans nos bases de données simulées. Le second consiste en la comparaison de la variabilité des méthodes. Pour ce faire, à chaque niveau de bruit fixé, 100 jeux de données ont été simulés afin d'en extraire une estimation de la variance liée à chacun des critères. Enfin, nous avons évalué les performances prédictives des modèles issus de chacun des critères. Pour ce faire, pour chaque jeu de données simulé à l'aide desquels nous avons établis les différents modèles, nous avons simulé des observations supplémentaires ; ces observations formant ainsi un jeu de données test que nous avons utilisé afin de procéder à ces comparaisons. Les moyennes de performances prédictives ont ensuite été comparées à l'aide de test de Student. Les précisions quant à ces comparaisons sont disponibles dans le chapitre suivant.

Nous avons enfin testé notre méthode sur différentes bases de données réelles, nous permettant ainsi de vérifier certaines propriétés observées sur les bases de données simulées, notamment quant à la stabilité et la robustesse de ce nouveau critère.

4.4 Résultats

Nous nous sommes particulièrement intéressés à la robustesse des méthodes aux différents niveaux de bruits aléatoires qui peuvent être présents dans les bases de données réelles.

Dans le cadre de la régression PLS usuelle, il a été intéressant de remarquer, par exemple, que le Q^2 soit particulièrement sensible au niveau de bruit présent dans la réponse, l'amenant rapidement à sous-estimer le nombre de composantes à extraire. Ce problème de sous-estimation ayant déjà été remarqué par Tenenhaus (1998) suite à l'analyse de la base de données des « Processionnaires du Pin ». Le critère du BIC adapté est quant à lui un critère satisfaisant dans le cas où $n > p$ mais se voit lié à une variance déraisonnable dès lors que $n < p$. Notre nouveau critère basé sur le bootstrap retourne des résultats satisfaisants que ce soit en terme de stabilité ou de robustesse au bruit. De plus, nous avons pu conclure à de meilleures performances prédictives liées à ce nouveau critère et plus particulièrement dans le cadre de bases de données caractérisées par des niveaux de bruit non-négligeables.

Dans le cadre de la régression PLS logistique, notre critère peut être considéré comme un compromis entre les deux autres critères étudiés. En effet, la méthode du nombre de mal-classés par VC est lié au biais le plus faible alors que le critère développé par Bastien *et al.* (2005) est lié au biais le plus élevé. En ce qui concerne la variabilité, l'inverse a été observé. Notre critère s'est ainsi avéré être un compromis intéressant, d'autant plus qu'il s'avère lié à des performances prédictives au moins aussi bonnes que celles des autres critères étudiées.

Dans le cadre de la régression généralisée PLS liée à une distribution de Poisson, les

critères existant que nous avons étudiés retournent tous un nombre croissant de composantes en fonction du niveau de bruits aléatoire présent dans la réponse. Ce phénomène conduit ainsi à l'obtention de modèles sur-paramétrés, liés à des composantes tendant à modéliser ce bruit, et possédant de moins bonnes performances prédictives. Notre nouveau critère semble donc être le seul à recommander dans ce cadre là.

Chapitre 5

A New Universal Resample-stable Bootstrap-based Stopping Criterion for PLS Component Construction

Le chapitre suivant consiste en un article soumis au journal *Statistics and Computing* et qui est actuellement en cours de relecture. Il recouvre l'essentiel des travaux qui ont été menés sur cette problématique. Les premiers résultats de ces travaux ont également fait l'objet d'une communication sous forme de poster à la 8$^{\text{ème}}$ conférence internationale sur la PLS qui s'est déroulée à Paris en Mai 2014, 8$^{\text{th}}$ *International Conference on Partial Least Squares and Related Methods* (Annexe E). Un article contenant ces premiers résultats a ainsi été soumis et accepté pour publication dans les actes de la conférence (Annexe F). Les démonstrations des propositions énoncées dans ce chapitre sont disponibles en Annexe B.

Abstract

We develop a new robust stopping criterion for partial least squares regression (PLSR) component construction, characterized by a high level of stability. This new criterion is universal since it is suitable both for PLSR and extensions to generalized linear regression (PLSGLR). The criterion is based on a non-parametric bootstrap technique and must be computed algorithmically. It allows the testing of each successive component at a preset significance level α. In order to assess its performance and robustness with respect to various noise levels, we perform dataset simulations in which there is a preset and known number of components. These simulations are carried out for datasets characterized both by $n > p$, with n the number of subjects and p the number of covariates, as well as for $n < p$. We then use t-tests to compare the predictive performance of our approach with other common criteria. The stability property is in particular tested through re-sampling processes on a real allelotyping dataset. An important additional conclusion is that this new criterion given globally better predictive performances than existing ones in both the PLSR and PLSGLR (logistic and Poisson) frameworks.

Supplementary material is linked to this article.

Keywords: Bootstrap, PLSR, PLSGLR, Latent variable, Robustness

5.1 Introduction

Modeling relationships using traditional statistical methods like ordinary least squares regression (OLSR), between a univariate response and highly correlated covariates, is rarely recommended, and for datasets including more covariates than subjects, is not even possible. However, with recent technological and computing advances, providing consistent analysis of such datasets has become a major challenge, particularly in domains such as medicine, biology and chemometrics. For such reasons, statistical methods have been developed, including partial least squares regression (PLSR), introduced by Wold *et al.* (1983) and described in detail by Höskuldsson (1988), amongst others. PLSR has become a standard tool in chemometrics (Wold *et al.*, 2001) and for dealing with genomic datasets (Boulesteix and Strimmer, 2007). Indeed, due to its appealing properties, PLSR is able to efficiently deal with high-dimensional settings, and notably, resolves the collinearity problem (Wold *et al.*, 1984).

In this paper, we focus on the PLS univariate response framework, better known as PLS1. Let n be the number of observations and p the number of covariates. Then, $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ represents the response vector, with $(.)^T$ denoting the transpose, and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p) \in \mathcal{M}_{n,p}(\mathbb{R})$ the covariate matrix, with $\mathcal{M}_{n,p}(\mathbb{R})$ the set of matrices with n rows and p columns. Note that without loss of generality, $\mathbf{X}$ and $\mathbf{y}$ are supposed centered, and scaled to unit variance. PLSR consists of building $K \leqslant \text{rk}(\mathbf{X})$ orthogonal latent variables $\mathbf{T}_K = (\mathbf{t}_1, \ldots, \mathbf{t}_K)$, also called components or scores vectors, in such a way that $\mathbf{T}_K$ optimally describes the common information space between $\mathbf{X}$ and $\mathbf{y}$. In order to do so, these components are built up as linear combinations of the original covariates, i.e.,

$$\mathbf{t}_k = \mathbf{X}_{k-1}\mathbf{w}_k,\ 1 \leqslant k \leqslant K, \tag{5.1}$$

where $\mathbf{X}_0 = \mathbf{X}$, and $\mathbf{X}_{k-1},\ k \geqslant 2$, represents the residual covariate matrix obtained through the OLSR of $\mathbf{X}$ on $\mathbf{T}_{k-1}$. $\mathbf{w}_k = (w_{1k}, \ldots, w_{pk})^T$ is obtained as the solution of the following maximization problem (Boulesteix and Strimmer, 2007):

$$\begin{aligned}\mathbf{w}_k =& \underset{\mathbf{w}\in\mathbb{R}^p}{\text{argmax}} \left\{\text{Cov}^2\left(\mathbf{y}_{k-1}, \mathbf{t}_k\right)\right\} &(5.2)\\ =& \underset{\mathbf{w}\in\mathbb{R}^p}{\text{argmax}} \left\{\mathbf{w}^T\mathbf{X}_{k-1}^T\mathbf{y}_{k-1}\mathbf{y}_{k-1}^T\mathbf{X}_{k-1}\mathbf{w}\right\}, &(5.3)\end{aligned}$$

with the constraint $\|\mathbf{w}_k\|_2^2 = 1$, and where $\mathbf{y}_0 = \mathbf{y}$, and $\mathbf{y}_{k-1}$ represents the residual vector obtained from the OLSR of $\mathbf{y}$ on $\mathbf{T}_{k-1}$.

These components can also be directly linked to the original covariate matrix:

$$\mathbf{t}_k = \mathbf{X}\mathbf{w}_k^* = \sum_{j=1}^{p} w_{jk}^*\mathbf{x}_j,\ 1 \leqslant k \leqslant K, \tag{5.4}$$

where $\mathbf{w}_k^* = \left(w_{1k}^*, \ldots, w_{pk}^*\right)^T$ is the vector of the original covariates' weights, dependent on $\mathbf{y}$ (Wold *et al.*, 2001). As demonstrated by Tenenhaus (1998, p.114), by noting $\mathbf{W}_k^* =$

$(\mathbf{w}_1^*, \ldots, \mathbf{w}_k^*) \in \mathcal{M}_{p,k}(\mathbb{R})$, this matrix satisfies the following equation:

$$\mathbf{W}_k^* = \mathbf{W}_k \left(\mathbf{P}_k \mathbf{W}_k^T\right)^{-1}, \tag{5.5}$$

where $\mathbf{P}_k = (\mathbf{p}_1, \ldots, \mathbf{p}_k) \in \mathcal{M}_{p,k}(\mathbb{R})$ is the matrix containing the k vectors of regression coefficients from the OLSR of $\mathbf{X}$ on $\mathbf{T}_k$, also known as $\mathbf{X}$-loadings.

Let K be the selected number of components. The final regression model is thus:

$$\begin{aligned}
\mathbf{y} =& \sum_{k=1}^{K} c_k \mathbf{t}_k + \epsilon & (5.6)\\
=& \sum_{k=1}^{K} c_k \left(\sum_{j=1}^{p} w_{jk}^* \mathbf{x}_j \right) + \epsilon & (5.7)\\
=& \sum_{j=1}^{p} \beta_j^{PLS} \mathbf{x}_j + \epsilon, & (5.8)
\end{aligned}$$

with $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ the $n \times 1$ error vector and $(c_1, \ldots, c_K)$ the regression coefficients from the OLSR of $\mathbf{y}$ on $\mathbf{T}_K$, also known as $\mathbf{y}$-loadings.

In order to take into account specific distributions linked to the response, an extension to the generalized linear regression method, noted PLSGLR, was introduced by Marx (1996). This led to further research and developments related to the field (Nguyen and Rocke, 2002b; Boulesteix, 2004; Ding and Gentleman, 2005). Note that in this case, $\mathbf{y}$ is naturally not centered or scaled to unit variance. In this paper, the process developed by Bastien *et al.* (2005) and implemented in the R package *plsRglm* (Bertrand *et al.*, 2014) is used. In this context, the regression model is the following:

$$g(\theta) = \sum_{k=1}^{K} c_k \left(\sum_{j=1}^{p} w_{jk}^* \mathbf{x}_j \right), \tag{5.9}$$

with θ the conditional expected value of $\mathbf{y}$ for a continuous distribution, or the probability vector of a discrete distribution with a finite support. The link function g depends on the distribution of $\mathbf{y}$.

As mentioned above, both PLSR and its extension to generalized models rely on determining a tuning parameter: the number of components. The obtention of an optimal number of components K_{opt} is crucial to get reliable estimations of the original covariates' regression coefficients. Concluding that $K < K_{opt}$ leads to a loss of information, meaning that connections between some covariates and $\mathbf{y}$ are not correctly modeled. Concluding that $K > K_{opt}$, i.e., over-fitting, can lead to models with poor predictive ability (Wiklund *et al.*, 2007).

Despite the fact that PLSR has become a versatile and standard tool in many domains like chemometrics, bioinformatics, medicine and social science (Rosipal and Krämer, 2006), choosing well the number of components is still an open and important problem (Wiklund *et al.*, 2007). Indeed, the relative lack of theoretical hypotheses, leading PLSR to be called a *soft-modeling* process (Manne, 1987), precludes the development of typical statistical tests based on theoretical distributions for testing parameters (Wakeling and Morris, 1993). Therefore, a substantial number of papers deal with this question by introducing new statistics or

comparing several statistics' abilities. Most developed criteria are based on the predictive residual error sum of squares (PRESS), introduced by Allen (1971) for model selection. To be calculated, this statistic ideally needs an independent test set. However, notably due to logistical constraints, this additional set is rarely available (Efron and Tibshirani, 1993, p. 240). Therefore, cross-validation (CV) techniques are usually used to obtain an estimation of PRESS-based statistics. Issues concerning CV methods for establishing prediction ability are reported in the literature, notably linked to the high variability of obtained results (Efron and Tibshirani, 1993, p. 255; Hastie *et al.*, 2009, p. 249; Wiklund *et al.*, 2007, p. 429; Boulesteix, 2014). Such issues are observed in this paper. An alternative to CV methods is the well-known bootstrap technique introduced by Efron (1979). Using this process for estimating prediction errors has already been proposed, notably by Efron and Tibshirani (1993), and also adapted to selecting the optimal number of components in PLS and principal component regression (PCR) (Wehrens and Linden, 1997; Denham, 2000; Amato and Vinzi, 2003; Mevik and Cederkvist, 2004). However, it has also been established that though the use of the bootstrap for predictive error estimation efficiently reduces the variability issue, it can also lead to large bias (Efron and Tibshirani, 1993, p. 255; Kohavi, 1995). Much further literature is also available, introducing new criteria or comparing criteria: Höskuldsson (1996), Van der Voet (1994), Li *et al.* (2002a), Green and Kalivas (2002), Gourvénec *et al.* (2003), Gómez-Carracedo *et al.* (2007) are some examples. Performing a global state-of-the-art review on this subject would be difficult due to the vast number of previous works. However, it is clear that there is not yet one precise criteria that can be considered reliable *in general*. In the PLSGLR framework, it is also notable that very few criteria adapted to this situation have been proposed, and none of them can currently be considered as a good general benchmark.

The aim of the article is twofold. First, we wish to establish a new criterion that can be considered universal, i.e., both reliable and easily adaptable to both the PLSR and PLSGLR frameworks. To the best of our knowledge, no previous criteria features this property. Second, this new criterion has to avoid CV methodology and related issues such as instability. Therefore, we develop a new bootstrap-based criterion to select the number of PLS components. The originality of the approach is due to the fact that it tests directly both the $\mathbf{X}$- and $\mathbf{y}$-loadings. To do this, the establishment of bootstrap-based confidence intervals (CI) is achieved. By focusing on the unknown distribution of the regression coefficients rather than predictive error-based statistics as previously proposed, we open up the possibility of directly testing the significance of successive components, which is pertinent for both the PLSR and PLSGLR frameworks. This method avoids the use of CV techniques and related issues.

In this article, we first explain the context and give theoretical details, before introducing the new algorithmic bootstrap-based criterion as pseudo-code in Section 5.2. In Section 5.3, we present existing criteria that have been chosen for comparison purposes, and then describe the simulation set-up we use to make comparisons. In Section 5.4, we analyze results obtained in the PLSR framework, followed by PLSGLR results for logistic regression (PLS-LR) and Poisson regression (PLS-PR) in Section 5.5. In Section 5.6, we focus on some real datasets and compare our new criterion to relevant existing ones. Using a real allelotyping dataset, we also compare the robustness of our new bootstrap-based criterion through resampling, approximating the distribution of the extracted number of components. Lastly, in Section 5.7, we discuss the observed advantages and disadvantages of each criterion.

5.2 A new boostrap based stopping criterion

5.2.1 Context

As mentioned in Section 5.1 and to the best of our knowledge, no criterion for tuning the number of components can currently be considered the benchmark. In addition, being derived from CV and thus linked to issues discussed in Section 5.1, known criteria are often based on arbitrary or empirical threshold values (Krzanowski, 1987), or theoretical asymptotic distributions (Haaland and Thomas, 1988; Osten, 1988), which are not appropriate for general reliable establishment of PLS models. For such reasons, we have developed a new criterion which is not based on CV processes and does not depend on arbitrary threshold values. Futhermore, our aim is not to directly focus on predictive ability-based statistics (already well-developed in the literature), but rather on scores vectors themselves, by searching for a way to test their significance, like is done for OLSR with Student-type tests. However, as PLSR methodology is a soft-modeling process (5.1), no such global distribution can be used.

The bootstrap is a well-known method for approximating unknown distributions. Bootstrap techniques adapted to the regression framework have already been proposed by Efron (1979) and Freedman (1981). As a bootstrap-based criterion could be a useful way to avoid CV, it was proposed for PLS component selection, notably by Denham (2000) and Amato and Vinzi (2003). However, to the best of our knowledge, a bootstrap-based process has never been used in order to test the various loadings involved, and represents an option for choosing an optimal number of PLS components, which covers all our goals.

5.2.2 Bootstrapping pairs in PLSR

Let $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^T = (\mathbf{y}, \mathbf{X}) \in \mathcal{M}_{n,p+1}(\mathbb{R})$, i.e., $\mathbf{z_i} = (y_i, x_{i1}, \ldots, x_{ip})$, $1 \leqslant i \leqslant n$. The so-called bootstrapping pairs method was introduced by Freedman (1981) and consists of building R new datasets by re-sampling with replacement in $\mathbf{Z}$ in order to mimic the generation of the original data. This leads to an empirical approximation of the distribution linked to a statistic $\mathcal{S}(\mathbf{Z})$. This technique only relies on the assumption that the originals pairs $(y_i, \mathbf{x}_{i\bullet})$, where $\mathbf{x}_{i\bullet}$ represents the i^{th} row of $\mathbf{X}$, are randomly sampled from some unknown $(p+1)$-dimensional distribution. It was developed to treat so-called correlation models, in which covariates are considered as random, and ϵ may be related to them. In this way, it is appropriate to "estimate the regression plane for a certain population on the basis of a simple random sample" (Freedman, 1981, p. 1219).

Constructing a new component $\mathbf{t}_k$ as described in Section 5.1 implies that $\mathbf{w}_k = \frac{\mathbf{X}_{k-1}^T \mathbf{y}_{k-1}}{\left\|\mathbf{X}_{k-1}^T \mathbf{y}_{k-1}\right\|}$. This property leads to the following result.

Proposition 5.2.1. *Let $\mathbf{y}_0 = \mathbf{y}$ and $\mathbf{X}_0 = \mathbf{X}$. Let $\mathbf{y}_{k-1}$ and $\mathbf{X}_{k-1}$, $k \geqslant 2$, be respectively the residual response vector and covariate matrix obtained through both the OLSR of $\mathbf{y}$ and $\mathbf{X}$ on $\mathbf{T}_{k-1}$. Suppose that,*
$\forall k \in [\![1, K]\!], \exists i \in [\![1, p]\!], \mathbf{x}_{i,(k-1)}^T \mathbf{y}_{(k-1)} \neq 0$.
Then, the PLS component building process implies that:
$\forall k \in [\![1, K]\!]$, $c_k > 0$ *and, conditionally on $\mathbf{X}$, c_k follows a positive distribution.*

As a consequence of this result, bootstrapping pairs $(y_i, \mathbf{x}_{i\bullet})$ by applying PLSR to each bootstrap sample in order to test Y-loadings, is not straightforward.

Furthermore, this method does not appear to be relevant since it approximates the uncertainty of the subspace spanned by the scores vectors, though this is not the initial aim. Our goal is to test particular PLS components based on the original dataset, since these latent variables are built and used for modeling specifically these original data. In other words, a method able to test the significance of these particular random latent variables, defined as specific linear combinations obtained through the PLSR processed on the original dataset, is to be looked for. As a concrete example, let $\mathbf{t}_1^{ori} = \sum_{j=1}^{p} w_{j1}^{ori}\mathbf{x}_j$ be the first PLS component based on the original data. By bootstrapping pairs $(y_i, \mathbf{x}_{i\bullet})$ and applying the PLS process to a bootstrap sample $(\mathbf{y}, \mathbf{X})^b$, the obtained weights $\mathbf{w}_1^b$ are naturally different from $\mathbf{w}_1^{ori}$, so the uncertainty of the specific random variable $\sum_{j=1}^{p} w_{j1}^{ori}\mathbf{x}_j$ is not tested by this ill-adapted process, but rather uncertainty about the construction of this first component.

To succeed in testing these specific components, a bootstrapping pairs $(y_i, \mathbf{x}_{i\bullet})$ process has to be performed, while keeping fixed the weights $\mathbf{W}_k^{ori}$ obtained on the original data, for the construction of the components linked to each bootstrap sample. Thus, the specific uncertainty of the particular linear combination of the original variables is approximated. Performing this process is equivalent to bootstrapping pairs $(y_i, \mathbf{T}_{k,i\bullet})$, where $\mathbf{T}_{k,i\bullet}$ represents the i^{th} row of $\mathbf{T}_k$, i.e., sampling from an empirical distribution conditional on the scores vectors $\mathbf{T}_k$.

As the PLS components are built both for modeling the response and summarizing the original relevant information in $\mathbf{X}$, we propose to test each new component $\mathbf{t_k}$ by approximating the conditional distribution of the $\mathbf{X}$- and $\mathbf{y}$-loadings given $\mathbf{T}_k$. This is done by bootstrapping pairs $(y_i, \mathbf{T}_{k,i\bullet})$ and $(\mathbf{x}_{ij}, \mathbf{T}_{k,i\bullet})$, $\forall j \in [\![1, p]\!]$. We also propose to define the significance of a new component in terms of its significance for both $\mathbf{y}$ and $\mathbf{X}$, so that the extracted number of components K is defined as the last one which is significant for both.

5.2.3 Adapted bootstrapping pairs as a new stopping criterion

Based on our definition of the significance of a new component, a double bootstrapping pairs algorithm was constructed. The first step consists of bootstrapping pairs $(\mathbf{x}_{ij}, \mathbf{T}_{k,i\bullet})$, $\forall j \in [\![1, p]\!]$. We propose that a component is considered significant for $\mathbf{X}$ if and only if it is significant for at least one of the original covariates. Components are successively tested until we reach the first non-significant one. This step leads to a maximal number of components $k_{\max}$ that can be extracted. The second step consists of bootstrapping pairs $(y_i, \mathbf{T}_{k,i\bullet})$ to test the significance against $\mathbf{y}$ of each successive component $\mathbf{t}_k$, with $k \leqslant k_{\max}$. To avoid confusion between the number of covariates and $\mathbf{X}$-loadings, we set m as the total number of original covariates.

The algorithm of this double bootstrapping pairs implementation is thus as follows:

I Bootstrapping $(\mathbf{X}_{i\bullet}, \mathbf{T}_{k,i\bullet})$:

Let $k = 0$.

Repeat

1 $k = k + 1$.

2 Compute the k^{th} component, defining $\mathbf{T}_k = (\mathbf{t}_1, \ \ldots, \mathbf{t}_k)$.

3 Bootstrap pairs $(\mathbf{X}_{i\bullet}, \mathbf{T}_{k,i\bullet})$, returning R bootstrap samples:

$$(\mathbf{X}, \mathbf{T}_k)^{b_1}, \ldots, (\mathbf{X}, \mathbf{T}_k)^{b_R}.$$

4 For each $(\mathbf{X}, \mathbf{T}_k)^{b_r}$, do m OLS regressions:

$$\mathbf{x}_l^{b_r} = \sum_{j=1}^{k} \left(\hat{p_{lj}}^{b_r} . \mathbf{t}_j^{b_r}\right) + \hat{\delta}_{lk}^{b_r}.$$

5 $\forall p_{lk}$, construct a $(100 \times (1-\alpha))\%$ bilateral BC_a CI, noted:

$$\mathrm{CI}_l = \left[\mathrm{CI}_{l,1}^k, \mathrm{CI}_{l,2}^k\right].$$

Until $\forall l \in \{1, \ldots, m\}$, $0 \in \mathrm{CI}_l$.

Return $k_{\max} = k - 1$ and $\mathbf{T}_{k_{\max}}$.

II Bootstrapping $(y_i, \mathbf{T}_{k,i\bullet})$:

Note that for the PLSGLR case, the relevant generalized regression is performed at step 9.

Let $k = 0$.

Repeat

6 $k = k + 1$.

7 Compute $\mathbf{T}_k$ by extracting the k first columns from $\mathbf{T}_{k_{\max}}$.

8 Bootstrap pairs $(y_i, \mathbf{T}_{k,i\bullet})$, returning R bootstrap samples:

$$(\mathbf{y}, \mathbf{T}_k)^{b_1}, \ldots, (\mathbf{y}, \mathbf{T}_k)^{b_R}.$$

9 For each pair $(\mathbf{y}, \mathbf{T}_k)^{b_r}$, do the OLS regression:

$$\mathbf{y}^{b_r} = \sum_{j=1}^{k} \left(\hat{c}_j^{b_r} . \mathbf{t}_j^{b_r}\right) + \hat{\epsilon}_k^{b_r}.$$

10 Since $c_k > 0$, construct a $(100 \times (1-\alpha))\%$ unilateral BC_a CI:

$$\mathrm{CI} = \left[\mathrm{CI}_1^k, +\infty\right[\text{ for } c_k.$$

While $\mathrm{CI}_1^k > 0$ and $k \leqslant k_{\max}$.

Return the final extracted number of components $K = k - 1$.

Results linked to this bootstrap-based criterion are referred to as **BootYT** in the following.

5.3 Simulation

5.3.1 Existing criteria used for comparison

To perform our benchmarking study, several existing criteria were used.

In the PLSR framework, the Q^2 criterion was selected since it represents a standard criterion, implemented notably in both the R package *plsRglm* (Bertrand *et al.*, 2014) and the SIMCA-P software (Umetrics, 2005). This criterion is based on q-fold CV methods (Breiman *et al.*, 1984) and was computed for both $q = n$, leading to the universal standard CV method called *leave-one-out* CV (Gómez-Carracedo *et al.*, 2007), and $q = 5$ (5-CV) following recommendations of Kohavi (1995) and Hastie *et al.* (2009), so as to reduce variability in the CV method. The BIC criteria, corrected with the estimated degrees of freedom (DoF) by Krämer and Sugiyama (2011), was also included since to the best of our knowledge, no published study has analyzed its performance.

In the PLSGLR framework, there are a limited number of relevant criteria available; we thus present only two here: the number of misclassified values (Meyer *et al.*, 2010), and a criterion introduced by Bastien *et al.* (2005). Both are available in the R package *plsRglm*. The usual AIC and BIC criteria were also included.

- In PLSR:

  1. **Q^2**. For each new component $\mathbf{t}_k$, the following statistic is evaluated:

     $$Q_k^2 = 1 - \frac{\text{PRESS}_k}{\text{RSS}_{k-1}},$$

     where RSS_{k-1} represents the Residual Sum of Squares when the number of components is $k-1$, and PRESS_k the PRESS when the number of components is equal to k. Tenenhaus (1998) considers that a new component $\mathbf{t}_k$ improves significantly the prediction of $\mathbf{y}$ if:

     $$\sqrt{\text{PRESS}_k} \leqslant 0.95\sqrt{\text{RSS}_{k-1}} \iff Q_k^2 \geqslant 0.0975.$$

     Results linked to this criterion using both leave-one-out and $q = 5$ CV are referred to in the text and plots by **Q2lv1o** and **Q2K5** respectively.

  2. **BIC**. The R package *plsdof*, based on the work of Krämer and Sugiyama (2011), was used to compute this criterion. It works as follows:

     $$\text{BIC} = \text{RSS}/n + log(n)(\gamma/n)\widehat{\sigma_\epsilon}^2,$$

     where γ represents the DoF of model (7.5) and $\widehat{\sigma_\epsilon}^2$ is defined by Krämer and Sugiyama (2011).
     The selected model is the one which represents the first local minimum of this adapted BIC criterion; related results are referred to by **BICdof**. Results linked to models obtaining the global minimum are also returned under the acronym **BICglob**.

- In PLSGLR:

1. **AIC**. The AIC criterion (Akaike, 1974) can be computed whatever the distribution involved. However, no corrected DoF have yet been suggested for the PLSGLR framework.
2. **BIC**. As in the case of AIC, the BIC (Schwarz, 1978) is calculable without correcting the DoF.
3. **CV − MClassed**. This criterion can only be used for PLS-LR. Via 5-CV, it determines for each model the number of misclassified predicted values. The selected model is the one corresponding to this statistic's minimal value.
4. **p_val**. Bastien *et al.* (2005) define a new component $\mathbf{t}_k$ as non-significant if it contains no significant covariate. Asymptotic Wald tests are used to conclude as to the significance of the various covariates.

5.3.2 Simulation plan

To compare these criteria, simulations were performed by adapting the *simul_ data_ UniYX* function, available in the R package *plsRglm* (Bertrand *et al.*, 2014). First, four orthonormal vectors of size p are built. Let $\mathbf{T} \in \mathcal{M}_{4\times p}(\mathbb{R})$ be the matrix containing them. Then, rows of $\mathbf{X}$ are successively obtained using $\mathbf{X}_{i\bullet} = \mathbf{R}_i\mathbf{T} + \epsilon_i$, where $\mathbf{R}_i = (r_{1i}, \ldots, r_{4i}) \in \mathbb{R}^4$ is a vector of random draws from $\mathcal{N}(0, \sigma_j)$, $j = 1, \ldots, 4$ respectively, with $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (10, 8, 6, \sigma_4)$, and ϵ_i is drawn from $\mathcal{N}(0, 10^{-2})$. Only the first three orthonormal vectors are linked to the simulated response, so that σ_4 varies during the simulations in order to understand the impact of increasing variability of this uninformative fourth component on the various criteria. Different processes were used to simulate response vectors, depending on the desired distribution. As a constant in all simulation schemes, the first three orthonormal vectors are involved, so that whatever the framework, simulations are performed to obtain a relevant subspace of dimension 3. Also, a noise parameter in $\mathbf{y}$ helps us to determine the robustness of the criteria being examined with respect to increasing values of σ_5, which characterizes its standard deviation. Fixed sets of values for σ_4 and σ_5 are given, depending on the framework, and described in the corresponding sections. For more details about these simulation processes, see Supplementary Materials (Annex A).

Simulations were performed for two different cases, for both the PLSR and PLSGLR frameworks. The first was the $n > p$ situation with $n = 200$ and $p \in \Omega_{200} = [\![7, 50]\!]$. The second was the $n < p$ situation where $n = 20$ and $p \in \Omega_{20} = [\![25, 50]\!]$. For each fixed pair (σ_4, σ_5), which represents the standard deviation of the uninformative fourth component in $\mathbf{X}$ and the additional random noise standard deviation in $\mathbf{y}$, respectively, we simulated 100 datasets. Each dataset is based on p_l covariates, $1 \leqslant l \leqslant 100$. The p_l numbers are obtained by sampling with replacement in Ω_n. Testing these criteria on 100 different datasets allows us to calculate a mean value of the number of components for each fixed couple (σ_4, σ_5) as well as an estimated variance that represents the inherent stability of the various criteria. Lastly, the number of bootstrap replicates was fixed at $R = 500$ and CI were constructed by setting $\alpha = 0.05$. More in-depth details of the data simulation framework is available in Supplementary Materials (Annex A).

The aim of the simulations is twofold. First, a comparison of the chosen criteria, both through their results for the number of components and their robustness against different

random noise variances. Second, predictive abilities are compared using predictive normalized mean squared errors (PNMSE), calculated on 80 additional simulated samples per dataset in the $n < p$ framework, used as test sets. Normalization is performed by dividing the predictive mean squared errors (PMSE) related to the obtained model by the PMSE linked to the trivial one (constant model equal to the mean of the training data). Furthermore, as mentioned in Krämer and Sugiyama (2011, p. 702), "the large test sample size ensures a reliable estimation of the test error." Then, for each pair of values (σ_4, σ_5), asymptotic t-tests with Welch-Satterthwaite DoF approximation (Welch, 1947) are performed to compare the PNMSE averages over the 100 simulated datasets related to each criterion. All tests have been run at level $\alpha = 0.05$.

5.4 PLSR results

As mentioned in Section 5.3.2, the simulated subspace is spanned by three orthonormal vectors (components). By modeling using uninformative elements in $\mathbf{X}$, a model based on four components is thus overfitted. Any supplementary component will be built from random noise present in $\mathbf{X}$.

5.4.1 Initial selection

To select the best method, both between the Q2lv1o and the Q2K5 criteria, and between the BICdof and BICglob ones, results related to datasets with $n > p$, for the following sets of values for noise standard deviations (NSD), are considered:

$$(A) : \begin{cases} \sigma_4 \in \{0.01, 0.21, \ldots, 5.81\} \\ \sigma_5 \in \{0.01, 0.51, \ldots, 20.01\}. \end{cases}$$

The averages of the selected numbers of components over the 100 simulated datasets per couple are calculated. These averages, denoted by *nb_comp* and related to the BIC and Q^2 criteria, are presented graphically in Fig.5.1 and 5.2 respectively as functions of σ_4 and σ_5, respectively denoted *sigma4* and *sigma5*.

Based on results shown in Fig. 5.1, BICglob has stability issues. We observe that this is mainly due to the adapted DoF not necessarily increasing as the number of components rises. Therefore, since adding a component can surprisingly lead to smaller DoF, this criterion is related to both over-determination and stability issues. The BICdof process, by searching for the first local minimum of the adapted BIC criterion, allows us to focus on the comparison between models related to k components, $1 \leqslant k \leqslant K + 1$. Based on our observations, these successive models are mainly linked to increasing DoF, avoiding issues related to the BICglob process. Therefore, the BICglob criterion should be avoided, and we retain BICdof for further comparisons.

Concerning the Q^2 criterion, results displayed in Fig. 5.2 point to a negligible effect of different values of q on these average numbers of components. Since reducing the value of q implies a variance decrease in related results (Hastie *et al.*, 2009, p.243), the Q2K5 criterion is retained here for further comparisons.

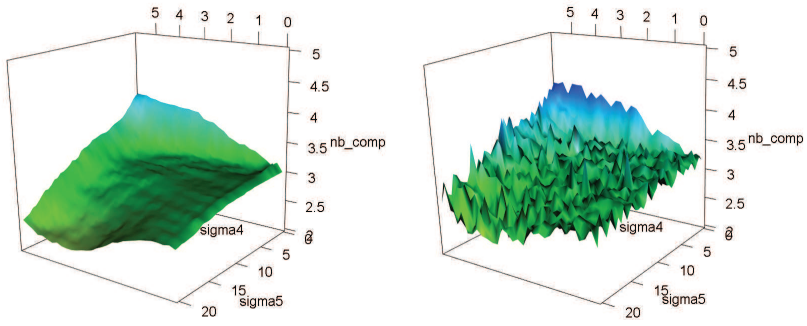


Figure 5.1: PLSR, $n > p$, sets of NSD values from (A), Evolution of average of selected numbers of components (nb_comp) over 100 datasets per pair (σ_4, σ_5) for BIC based criteria; left: BICdof; right: BICglob.

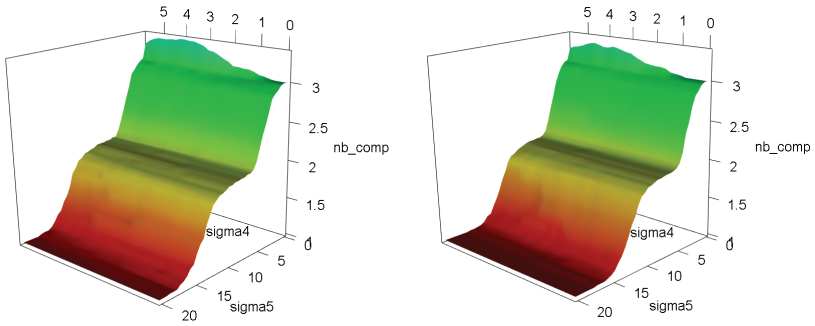


Figure 5.2: PLSR, $n > p$, sets of NSD values from (A), Evolution of averages of selected numbers of components (nb_comp) over 100 datasets per pair (σ_4, σ_5) for Q^2 based criteria; left: Q2lv1o; right: Q2K5.

In light of these initial observations, only three methods are retained: Q2K5, BICdof and our new bootstrap-based criterion.

5.4.2 PLSR: the $n > p$ case

To compare the three retained methods when $n > p$, the following enlarged sets of values for NSD are considered:

$$(B) : \begin{cases} \sigma_4 \in \{0.01, 0.21, \ldots, 5.81\} \cup \{6.01, 7.01, \ldots, 30.01\} \\ \sigma_5 \in \{0.01, 0.51, \ldots, 20.01\} \end{cases}$$

The means of number of components over the 100 simulated datasets per pair (σ_4, σ_5) are displayed for the three criteria in Fig. 5.3. Variances of these numbers of components over the 100 simulated datasets per pair were also estimated, and are shown using boxplots in Fig. 5.4. Note that these variances approximate the inter-dataset variability for fixed values of σ_4 and σ_5, not the intra-dataset one.

In these results, we see that the Q2K5 criterion is the least robust against increasing noise variability in $\mathbf{y}$, characterized by increasing values of σ_5 (sigma5). This lack of robustness

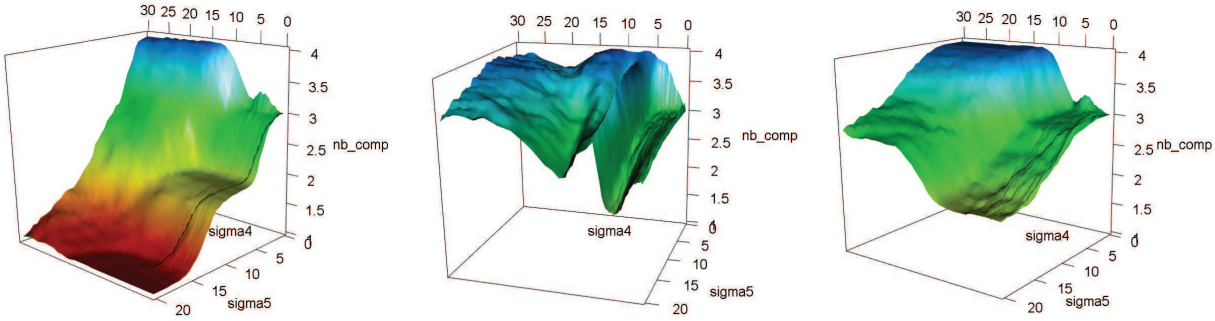


Figure 5.3: PLSR, $n > p$, sets of NSD values from (B), evolution of averages of selected numbers of components (nb_comp) over 100 datasets per pair (σ_4, σ_5); from left to right: Q2K5, BICdof and BootYT criteria.

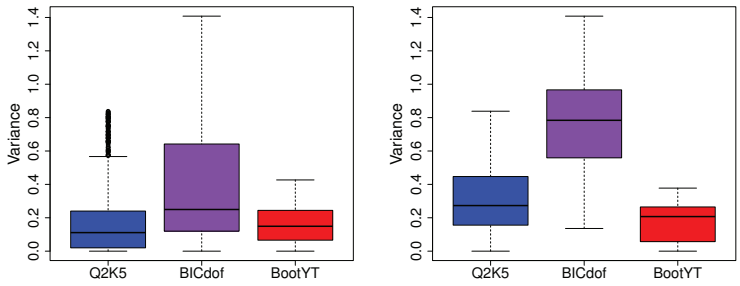


Figure 5.4: PLSR, $n > p$, sets of NSD values from (B); left: boxplots of estimated variance in the number of components over the 100 datasets per pair (σ_4, σ_5) for all involved values of σ_4 and σ_5; right: boxplots of estimated variance in the number of components over the 100 datasets per pair (σ_4, σ_5) for all involved values of σ_5 and $\sigma_4 \geqslant 15.01$.

leads it to globally underestimate the number of components. BICdof has a low computational requirements and is also the most robust against increasing values of σ_5. 86.37% of all its selected numbers of components are equal to three or four. However, as seen in Fig. 5.4, the BICdof features the highest global variability in number of components selected over the 100 datasets involved per pair (σ_4, σ_5). This is even more acute for datasets characterized by a fourth component standard deviation that is higher than that involved in the relevant subspace, i.e.,

$$\sigma_4 > \sqrt{\sum_{i=1}^{3} \sigma_i^2} = \sqrt{200} \simeq 14.14. \tag{5.10}$$

In this particular case, our new bootstrap-based criterion retains stability, while the median of the BICdof results, for instance, more than triples (0.25 to 0.79) compared to that of the whole data. Moreover, BootYT is the most robust against increasing variability of the uninformative fourth component in $\mathbf{X}$.

As a preliminary conclusion based on these initial results, advising the use of a certain choice among the BICdof or BootYT criteria is not relevant in the $n > p$ case. Due to its lack of robustness against noise variability in $\mathbf{y}$, the Q2K5 criterion should be avoided.

5.4.3 PLSR: the $n < p$ case

As suggested by Krämer and Sugiyama (2011), a small training sample size allows us to consider high-dimensional settings.

Mean and variance analyses

In this $n < p$ framework, the following sets of values for NSD are considered for criteria comparison:

$$(C) : \begin{cases} \sigma_4 \in \{0.01, 1.01, \ldots, 6.01\} \\ \sigma_5 \in \{0.01, 0.51, \ldots, 20.01\} . \end{cases}$$

Averages of numbers of components over the 100 datasets per pair (σ_4, σ_5) are displayed in Fig. 5.5. Graphical representations of variances are also shown in Fig. 5.6.

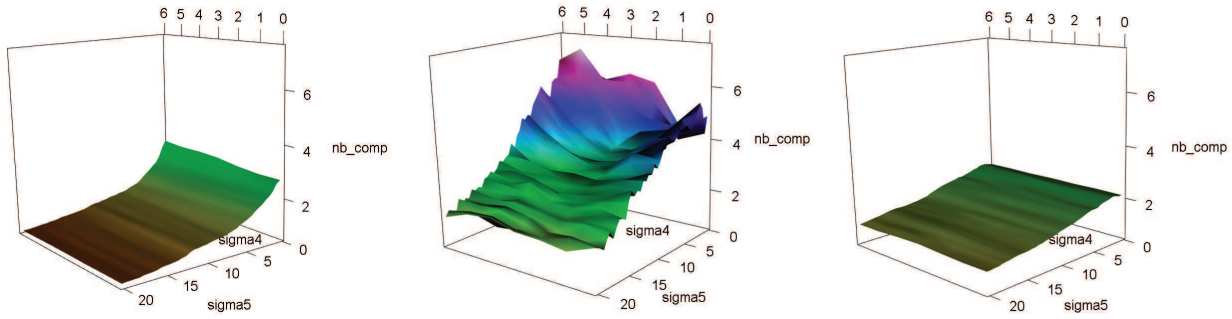


Figure 5.5: PLSR, $n < p$, sets of NSD values from (C), evolution of averages of selected number of components (nb_comp) over 100 datasets per pair (σ_4, σ_5); from left to right: Q2K5, BICdof and BootYT criteria.

In Fig. 5.5, we see that the BICdof appears to suffer from overfitting issues. Moreover, based on the results in Fig. 5.6, it returns results linked to out-of-range values of the variance, compared with the other two criteria. These two issues are mainly due to the extraction of 1678 (5.8%) results equal to 19 components, whereas 26184 (91.2%) trials give four or less components. By more carefully analyzing this phenomenon, it appears that the rate of 19 components is a globally decreasing function of σ_5. If these extreme results are considered non-representative of the criterion, the apparent lack of robustness, as well as the apparent over-fitting issues, may not be so important. However, these extreme results suggest inherent issues leading to a lack of reliability of this BIC criterion, and cannot not be ignored.

Our new boostrap-based criterion underestimates the number of components but is robust to increasing noise levels in $\mathbf{y}$, thus returning averages of number of components between 1.2 and 2.2. Moreover, the related low variance seen gives good evidence of stability. The Q2K5

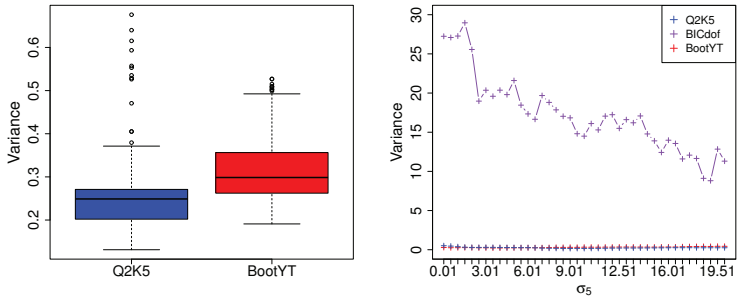


Figure 5.6: PLSR, $n < p$, sets of NSD values from (C); left: boxplots of estimated variance of the number of components over the 100 datasets per pair (σ_4, σ_5) for both the Q2K5 and BootYT criteria; right: evolution of estimated variances as a function of σ_5 for the three criteria studied.

criterion has comparable stability but is less robust to increasing noise levels in $\mathbf{y}$ than our criterion, meaning that in general, it is linked to significant under-fitting issues.

PNMSE analysis

The results of t-tests for PNMSE mean comparisons are shown in Fig. 5.7.

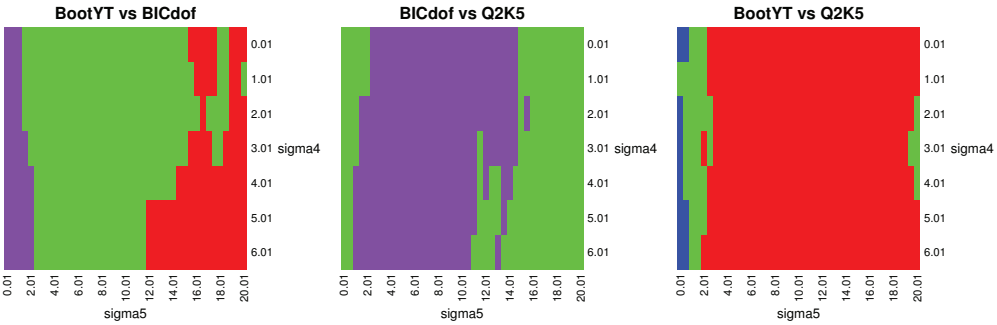


Figure 5.7: PLSR, $n < p$, sets of NSD values from (C); graphical representation of t-test results for PNMSE averages comparison; color code: BootYT better (red), BICdof better (purple), Q2K5 better (blue), no significant difference (green).

Results related to the smallest values of σ_5 require special consideration. Due to the consequent lack of noise in $\mathbf{y}$, models related to an over-determined number of components are not linked to the usual poor predictive ability issue since these supplementary scores vectors only try to model negligible noise. This implies that PNMSE are globally subject to the same rule as the MSE, i.e., the higher the number of components, the lower the PNMSE. As a direct consequence, the BICdof, which globally leads to over-fitted models (Fig. 5.5), returns by far the lowest PNMSE. This fact lead to only focus on the extracted number of components when $\sigma_5 \simeq 0$, so the Q2K5 criterion is to be advised in this particular case. However, such noiseless properties are rarely satisfied in real datasets. In all other cases,

the BootYT criterion returns models which are at least comparable if not better predictive performance than the other two.

5.4.4 PLSR: Conclusion

As a global conclusion, and in light of the results shown, the BootYT criterion can be seen as an interesting compromise between the other two in the $n > p$ framework, retaining their advantages but without their drawbacks. Indeed, this criterion offers both better robustness than Q2K5 against noise variability in $\mathbf{y}$, and better robustness than BICdof against variability of the uninformative fourth component in $\mathbf{X}$. It also features appealing stability compared to that of BICdof, especially for high σ_4 values. Concerning the $n < p$ case, extreme numbers of components selected by the BICdof criterion means that it is not pertinent to compare it to the other methods. While it returns 19338 (67.380%) results between two and four components, the over-determination issue cannot be ignored, while our criterion returns all its results below five. The BootYT criterion is also more robust against noise variability in $\mathbf{y}$ than Q2k5. Lastly, concerning predictive abilities, our new criterion has comparable if not better performance than the other two, with the exception of the case of negligible noise variability in $\mathbf{y}$, for which Q2K5 is advised. Recommendations are summarized in Table 5.1.

Table 5.1: Recommended criteria for PLSR.

	$n > p$		$n < p$	
	Low σ_4 values	High σ_4 values	Low σ_4 values	High σ_4 values
Negligible σ_5 values	BootYT / BICdof	BootYT	Q2K5	Q2K5
Non-negligible σ_5 values	BootYT / BICdof	BootYT	BootYT	BootYT

5.5 PLSGLR results

In this section, results related to the comparison between the bootstrap-based criterion and four other criteria (AIC, BIC, CV-MClassed and p_val, see Section 5.3.1) are presented. Note that, in this framework, due to specific distributions linked to $\mathbf{y}$ and the link-functions g used, an increase in σ_5 does not lead to a linear increase of noise variance in $\mathbf{y}$ as it does for PLSR simulated datasets. However, the bijectivity of these link functions ensures that a spanned subspace of dimension three is extracted.

5.5.1 PLS-LR results

PLS-LR: the $n > p$ case

The following sets of values for NSD are considered for criteria comparison:

$$(D) : \begin{cases} \sigma_4 \in \{0.01, 0.51, \ldots, 9.51\} \\ \sigma_5 \in \{0.01, 0.51, \ldots, 15.51\} \end{cases}.$$

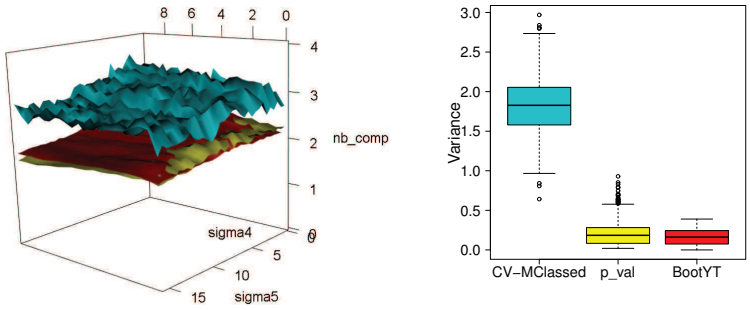


Figure 5.8: PLS-LR, $n > p$, sets of NSD values from (D); left: evolution of average of selected numbers of components (nb_comp) over 100 datasets per pair (σ_4, σ_5); right: boxplots of estimated variance of the number of components over the 100 datasets per pair (σ_4, σ_5).

Both the means and variances of numbers of components over the 100 datasets per pair (σ_4, σ_5) are displayed in Fig. 5.8.

Based on these, CV-MClassed performs well in estimating the optimal number of components, on average. However, the downside is the higher variances related to its results than those of the others. Therefore, this criterion should be used with caution. The BootYT and p_val criteria return similar results in the $n > p$ case. Both of them slightly underestimate the optimal number of components, but show stability in their results.

The uncorrected DoF lead the AIC and BIC criteria to globally overestimate the number of components (Supplementary Material, Annex C). Thus, these criteria should be avoided until the development of a DoF correction in this PLSGLR framework, and will not be considered in the $n < p$ case.

PLS-LR: the $n < p$ case

Here, the following sets of values for NSD are considered for criteria comparison:

$$(E) : \begin{cases} \sigma_4 \in \{0.01, 0.51, \ldots, 9.51\} \\ \sigma_5 \in \{0.01, 0.51, \ldots, 9.51\} \end{cases}.$$

Both the averages and variances of numbers of components over the 100 datasets per pair (σ_4, σ_5) are displayed in Fig. 5.9.

The CV-MClassed criterion retains both the same property of well estimating, on average, the number of components, and still has the variability issue. Concerning the two other criteria, we observe a greater underestimation issue linked to the p_val criterion than for BootYT. Furthermore, they both feature low variability.

PLS-LR: PNMSE and misclassified values analysis

Since the binary response obtained by the model is equal to 1 if the estimated response is over 0.5, 0 if not, returning higher PNMSE does not necessarily lead to a higher number

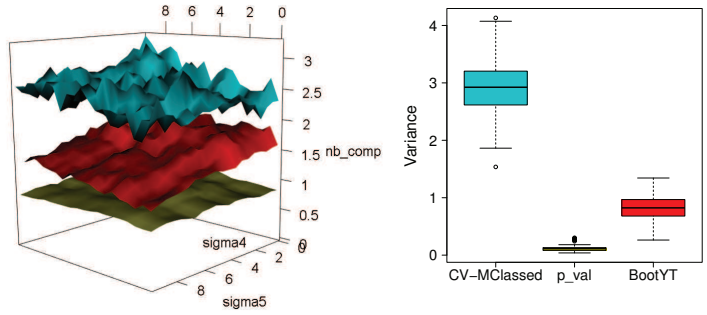


Figure 5.9: PLS-LR, $n < p$, sets of NSD values from (E); left: evolution of average of selected numbers of components (nb_comp) over 100 datasets per pair (σ_4, σ_5); right: boxplots of estimated variance of the number of components over the 100 datasets per pair (σ_4, σ_5).

of misclassified values. Thus, we also computed the number of misclassified predicted values (*M_classed*) for each of the three criteria. The results of t-tests are shown in Fig.5.10.

The bootstrap-based criterion is never less efficient than the others. If there is globally no significant difference between bootstrapping pairs and the p_val criterion related to the PNMSE, BootYT performs better than it in terms of number of misclassified predictions. Next, there are only a few cases where bootstrapping pairs are significantly better than CV-MClassed for the number of misclassified predictions. But, in terms of the PNMSE, the BootYT criterion is better than the latter as it returns significantly smaller PNMSE values, especially for high values of σ_5.

PLS-LR: Conclusion

From these simulations, it reasonable to assume that the bootstrap-based criterion is globally more efficient than the others. In the $n > p$ case, it has similar stability to p_val. However, it globally underestimates the optimal number of components, while CV-MClassed does not, but with high variability. As for the $n < p$ case, BootYT has better predictive performance than the two other criteria in terms of both PNMSE and predictive misclassified values. It also has low variability, important for any future implementation. Lastly, AIC and BIC are clearly not useful, since corrected DoF have not yet been established (Supplementary Material, Annex C). These conclusions are summarized in Table 5.2.

Table 5.2: Recommended criteria for PLS-LR.

Aim	Optimal number of components	Stability	Predictive abilities
	CV-MClassed	BootYT / p_val	BootYT

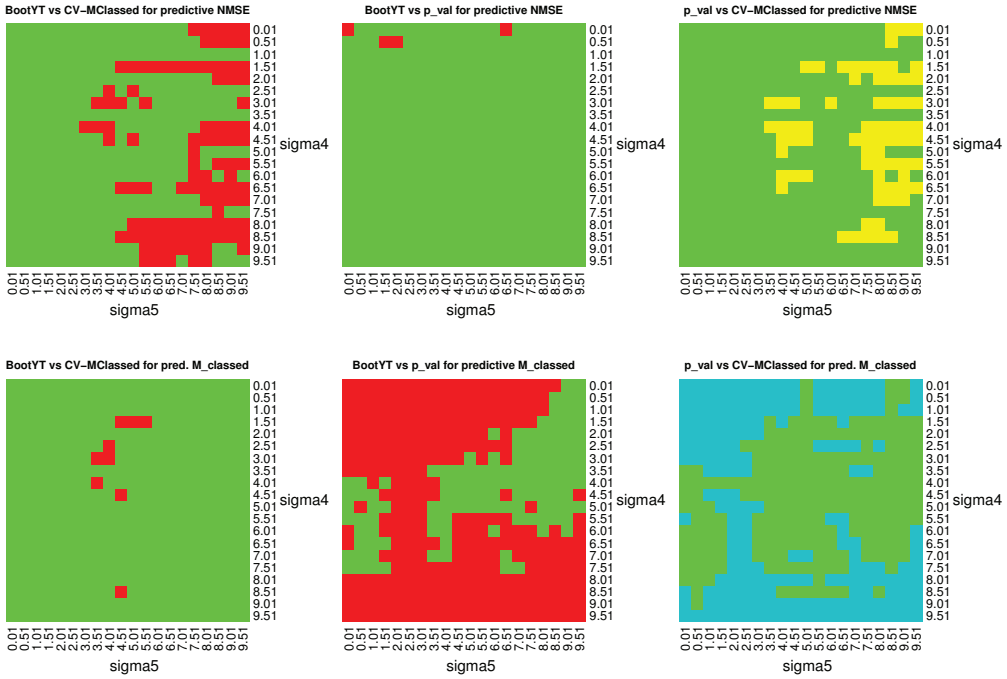


Figure 5.10: PLS-LR, $n < p$, sets of NSD values from (E); graphical representation of t-test results for both PNMSE and misclassified values averages comparison; color code: BootYT better (red), CV-MClassed better (turquoise), p_val better (yellow), no significant difference (green).

5.5.2 PLS-PR results

PLS-PR: Row mean analysis

In the PLS-PR case, the following sets of values for NSD are considered for respectively the $n > p$ (F) and $n < p$ (G) cases:

$$(F): \begin{cases} \sigma_4 \in \{0.01, 0.51, \ldots, 9.51\} \\ \sigma_5 \in \{0.01, 0.21, \ldots, 2.21\} \cup \{2.51, 3.01, \ldots, 7.01\} \end{cases}$$

$$(G): \begin{cases} \sigma_4 \in \{0.01, 0.51, \ldots, 9.51\} \\ \sigma_5 \in \{0.01, 0.21, \ldots, 2.21\} \cup \{2.51, 3.01, \ldots, 5.01\}. \end{cases}$$

Averages of number of components over the 100 datasets per pair (σ_4, σ_5), related to the four criteria considered (AIC, BIC, p_val and BootYT), are shown in Fig. 5.11 for both the $n > p$ and $n < p$ frameworks.

Apart from the bootstrap-based criterion, all criteria return an increasing number of components as σ_5 increases. These results lead us to conclude that our new bootstrap-based stopping criterion is the only one which is relevant for Poisson distributions, in that it selects, on average, a decreasing number of components as σ_5 increases. Based on these plots, no additional analyses of the numbers of components was done. Only the two criteria that give results, on average, closest to the expected result, are retained for further comparisons related to MSE, namely the p_val and BootYT criteria.

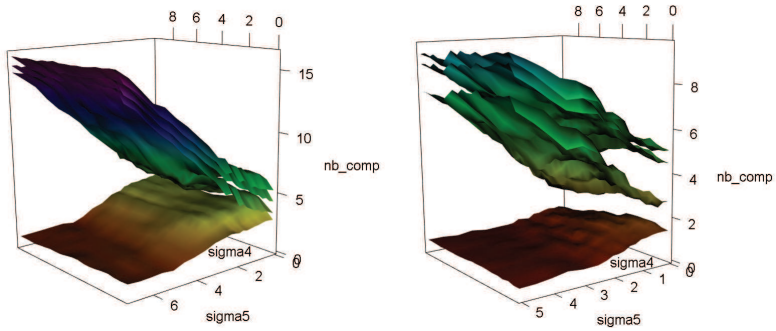


Figure 5.11: PLS-PR, evolution of average of selected numbers of components (nb_comp) over 100 datasets per pair (σ_4, σ_5); left: $n > p$, sets of NSD values from (F); right: $n < p$, sets of NSD values from (G); from top to bottom: AIC, BIC, p_val and BootYT results.

PLS-PR: MSE analysis

First, training log(MSE) were computed using the $n < p$ framework, and their means over all datasets related to each value of σ_5 are shown in Fig. 5.12.

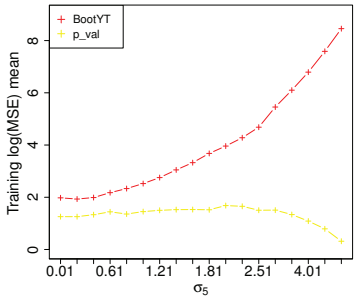


Figure 5.12: PLS-PR, $n < p$, sets of NSD values from (G); evolution of training log(MSE) means.

The global decrease in log(MSE) linked to p_val confirms that, as expected by the increasing number of extracted components observed in Section 5.5.2, this criterion models the random noise in $\mathbf{y}$. In contrast, the bootstrap-based criterion shows a systematic increase in log(MSE), which empirically suggests that it better succeeds in separating the real information from the noise.

Variances of PNMSE results over datasets related to each pair (σ_4, σ_5) were computed. Means of these variances, related to fixed values of σ_5, are displayed in Fig. 5.13.

While results obtained by the bootstrap-based criterion are linked to acceptable variances when $\sigma_5 \leqslant 1.61$, the out-of-range variances linked to the p_val results, due to the models' over-complexity observed in Section 5.5.2, lead to non-significant differences in mean while using t-tests on these datasets.

To obtain consistent t-test outcomes, models obtained in the $n > p$ framework were used. One hundred additional samples were simulated for each dataset to build test sets. Both means and means of variances of PNMSE over datasets for fixed values of σ_5 were computed, and are shown in Fig. 5.14.

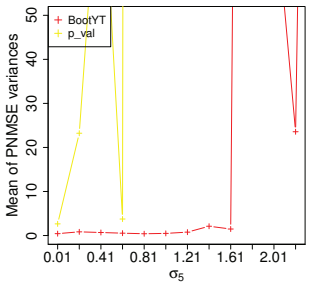


Figure 5.13: PLS-PR, $n < p$, sets of NSD values from (G); evolution of means of PNMSE variances for each σ_5.

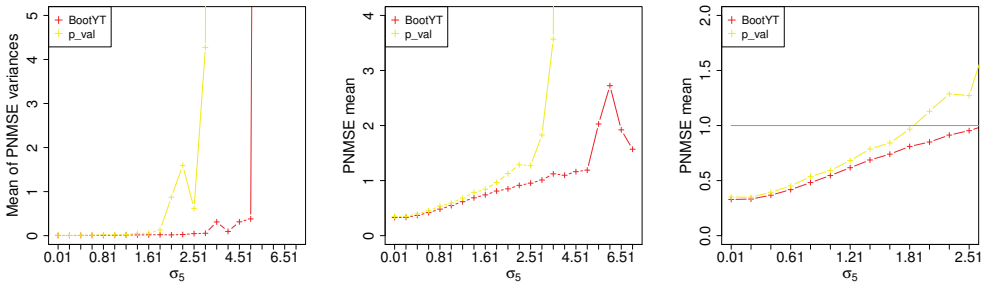


Figure 5.14: PLS-PR, $n > p$, sets of NSD values from (F); left: evolution of means of PNMSE variances for each σ_5; center: evolution of PNMSE means for each σ_5; right: evolution of PNMSE means for $\sigma_5 \leqslant 2.51$.

Based on these plots, it is clear that models built with the bootstrap-based criterion are on average better than the trivial ones when $\sigma_5 \leqslant 2.51$, while the p_val criterion fails to build better models than the trivial ones when the NSD in $\mathbf{y}$ is higher than 1.81. Both criteria return low variances in PNMSE for $\sigma_5 \leqslant 3.01$, so t-tests return consistent outcomes in this range of values. Results of these t-tests are displayed in Fig. 5.15.

Based on these t-tests results, our new criterion is to be recommended when setting up a predictive model. Note that non-significant differences for $\sigma_5 \geqslant 3.51$ are due to the high increase in variances linked to the p_val results (see Fig. 5.14).

PLS-PR: Conclusion

In the case of response vector $\mathbf{y}$ linked to a Poisson distribution, the bootstrap-based criterion stands out as the only one which should be used. Indeed, the others can be interpreted as increasing functions of σ_5, so they model the random noise in $\mathbf{y}$, leading to over-fitting issues. As a direct consequence, they return models with poor predictive abilities compared to the new criterion.

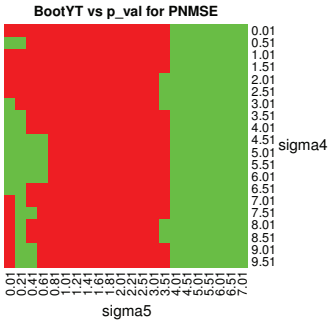


Figure 5.15: PLS-PR, $n > p$, sets of NSD values from (F); plot of t-test results for PNMSE means comparison; color code: BootYT better (red), no significant difference (green).

5.6 Applications on real datasets

5.6.1 Illustration of CV issues: first applications on real datasets

As mentioned by Boulesteix (2014), important issues concerning the stability of the q-fold CV procedure for the choice of tuning parameters, here the number of components, have been observed. These issues are directly induced by the value of q and by the random character of this resampling-based procedure while splitting the original dataset into two distinct sets, a training one and a test one. To illustrate consequences on the tuning parameter, we treated two real datasets.

The first dataset was collected from patients with colon adenocarcinoma. It has 104 observations on 33 binary qualitative explanatory variables, and one binary response variable representing the cancer stage according to the Astler-Coller (AB vs CD) classification (Astler and Coller, 1954). This binary response leads us to perform PLS-logistic regressions. This dataset, named *aze_compl*, is available in the *R* package *plsRglm* (Bertrand *et al.*, 2014).

We ran 100 times the selection process for the number of components using the CV-MClassed criterion, with $q \in \{3, 5, 10, 15, 30\}$. Then, we performed the same process using our new criterion. Results are shown in Fig. 5.16.

Results obtained through q-fold CV, with $q \neq n$, are displayed in Fig.5.16, and typical examples for these types of issue. Depending on the choice of q and the way the different folds are split, the extracted number of components can be dramatically different. In addition, obtaining a complete distribution of the number of components is essentially impossible, due to the high number of different possibilities for splitting the original datasets into q groups.

Proposition 5.6.1. *Let $n = pq + r$, $0 \leqslant r \leqslant q - 1$ be the Euclidean division of n by q. Then, the number of distinct partitions of the original dataset into r $(p+1)$-elements subsets and $(q-r)$ p-elements subsets for a CV does not depend on the order of their placement, and is equal to:*

$$f(n,q) = \frac{n!}{r!\,(q-r)!} \times \left(\frac{1}{(p+1)!}\right)^{r} \times \left(\frac{1}{p!}\right)^{q-r}. \tag{5.11}$$

Leave-one-out CV, which is the only complete CV ($f(n,n) = 1$, i.e., there is only one way to choose n folds out of n observations), selects one component. However, it suffers from

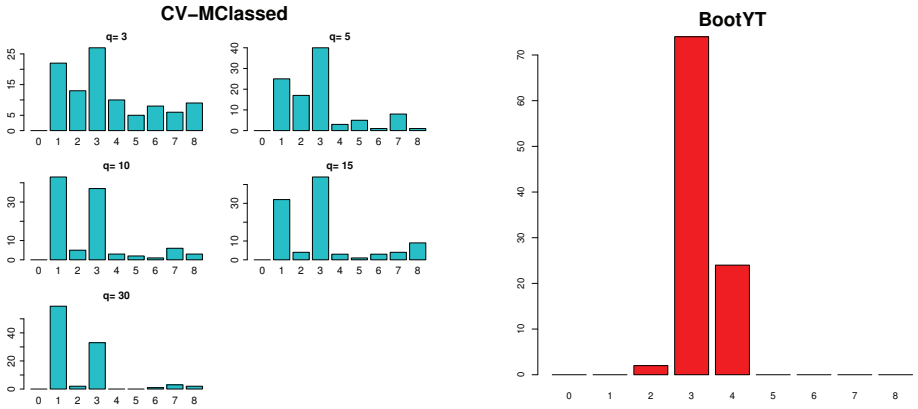


Figure 5.16: Extracted number of components using q-fold CV-MClassed (left) and BootYT (right) criteria.

variance issues concerning the bias-variance tradeoff on the estimation of the prediction error (Hastie *et al.*, 2009; Kohavi, 1995). Our new criterion is more stable on this dataset and leads the user to choose the number of components, in this case three, via a more accurate process.

The second example is a benchmark dataset, called "Processionnaire du Pin", which is treated in depth by Tenenhaus (1998). It has 33 observations each with 10 explanatory variables, and is also available in the *R* package *plsRglm* under the name *pine*. More details on this dataset are available in Tenenhaus (1998).

The same process was applied to this second example, with the usual PLS regressions. Thus, we can compare the Q^2 criterion obtained through q-fold CV, $q \in \{2, 3, 5, 10, 15\}$, and our new criterion. The Q^2 criterion obtained through leave-one-out CV chooses one significant component. All results are shown in Fig. 5.17.

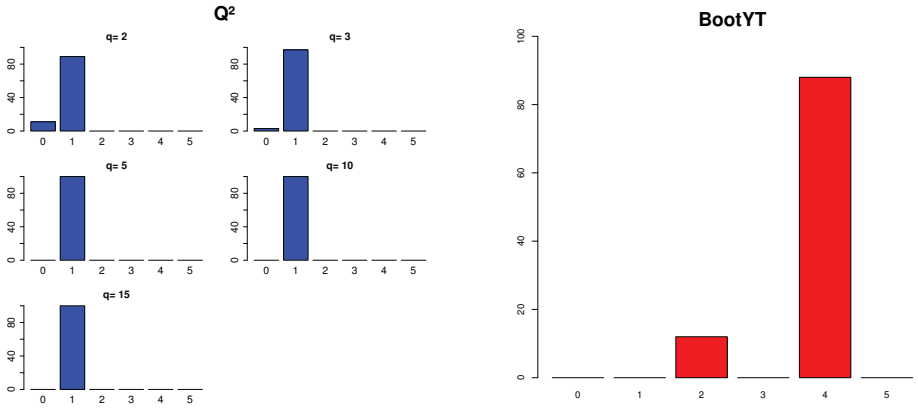


Figure 5.17: Extracted numbers of components using q-folds Q^2 (left) and BootYT (right).

Here, q-fold CV does not suffer from stability issues as seen before, since the Q^2 criterion is much more stable than the minimization of the number of misclassified values. However,

extracting one component is not recommended. Tenenhaus (1998), after a complete analysis of this dataset, showed that four components is the best decision. This under-estimating issue linked to the Q^2 criterion confirms the simulation results obtained in Section 5.4.2. Thus, while the Q^2 criterion under-estimates this optimal number of components, our new bootstrap-based criterion selects four components more than 80% of the time.

5.6.2 Application on an allelotyping dataset

In this section, we focus on an allelotyping study. Our method is applied to a dataset that concerns 267 subjects with colon cancer. Measures were made on 33 microsatellites, in search of an allelic imbalance that indicates an abnormal number of allele copies of a nearby gene of interest. The aim of the study was to find the microsatellite subsets that would best discriminate left and right colon tumors. Thus, the univariate response corresponds to the original location of a colon tumor, leading to a binary response $\mathbf{y}$, taking the value 0 (resp. 1) if it was on the right colon (resp. left). More details are available in Weber *et al.* (2007).

This dataset contains missing values, so a preprocessing step was performed in order to complete it, using the *R* package *mice*. As $\mathbf{y}$ is a 0-1 response, we used the three following stopping criteria in component construction: our new bootstrap-based criterion, CV-MClassed, and p_val. We performed 100 times the selection of the number of components using both the $q = 5$ CV-MClassed criterion and our new one, leading to the distribution of the extracted number of components shown in Fig. 5.18. Then, we computed the mean of the 100 values of extracted numbers of components related to the $q = 5$ CV-MClassed criterion, obtaining 7.99, which is higher than that obtained for BootYT. These results match the simulation conclusions (Section 5.5.1).

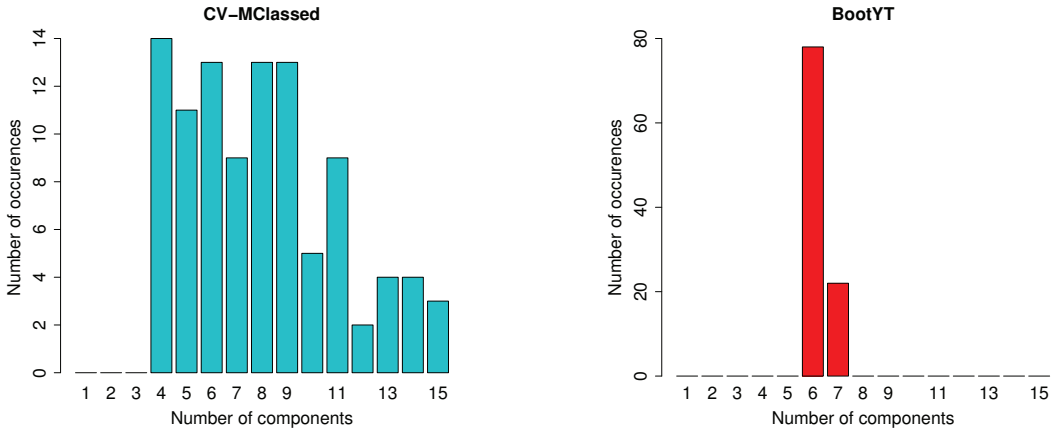


Figure 5.18: Extracted number of components using $q = 5$ CV-MClassed (left) and BootYT (right).

Based on the distributions in Fig. 5.18, the major default of the CV-MClassed criterion is clear, namely the dependence of the extracted number of components on the way the group has been randomly formed. Thus, performing a single CV to find the number of components

using this criterion, must be avoided. As expected, the BootYT criterion returns stable results and selects, in almost 80% of cases, 6 components.

We also tested the robustness of these three criteria through a bootstrap re-sampling process with 100 bootstrap iterations, as well as a jackknife method. These two resampling methods lead to distributions of the extracted number of components linked to each of the three criteria. Results are shown in Fig. 5.19.

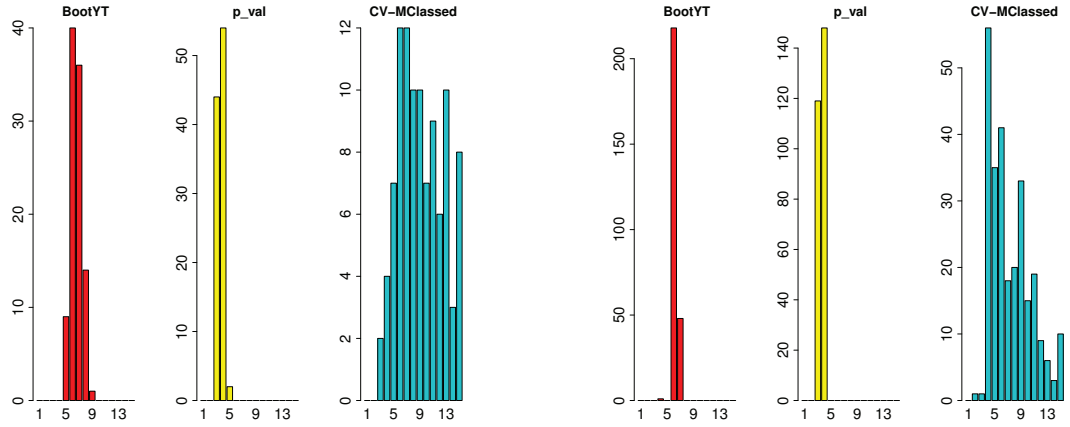


Figure 5.19: Distribution of extracted number of components through bootstrap (left) and jackknife (right) re-sampling.

These results confirm the high resampling robustness of our new criterion compared to CV-MClassed. The p_val criterion has comparable robustness but, based on our simulation results, higher bias. Based on these results and our simulations, we can reasonably conclude that for this dataset, the optimal number of components is 6.

5.7 Discussion

A new bootstrap-based stopping criterion for PLS component construction was developed, characterized by a high level of stability and robustness with respect to noise level compared to other well-known criteria.

Its implementation requires a function integrating the bootstrap process while building the PLS components. Though not yet available online, an equivalent form can be coded using existing R functions, notably from the *plsRglm* and *boot* packages. It requires recomputing, for each increment of k, the entire set of PLS components. While this style of implementation performs more operations than needed, it does not take much more time, since the main part of the computational cost comes from the bootstrap process.

Indeed, our bootstrap-based criterion has a shortcoming with respect to computational time, which is greater than the other methods since it requires, in the PLSR framework, $[(k_{\max}+1)\times p_l+\ (K+1)]\times R$ least squares regressions per dataset. An initial improvement has already been made by developing parallel processing for this. We note also that the development of a corrected DoF in the PLSGLR framework would also allow the development

of an adapted corrected BIC formulation. This could provide an interesting alternative to the bootstrap-based criterion since it might save a great amount of computational time.

Nevertheless, our new bootstrap-based criterion represents a reliable, consistent and robust stopping criterion for selecting the optimal number of PLS components. It avoids the use of CV techniques and, to the best of our knowledge, is the first to directly focus on the different loadings involved. Thus, it can be performed both in the PLSR and PLSGLR frameworks, and allows users to test the significance of a new component with a preset risk level α.

In the $n > p$ PLSR framework, our simulations confirm that both BICdof and BootYT are appropriate and well-designed criteria. Our new bootstrap-based criterion is also an appropriate alternative in the $n < p$ case, since the BICdof criterion suffers from high variance and overestimation issues, especially for models with low random noise levels in $\mathbf{y}$. Furthermore, both the BICdof and Q2K5 criteria are more sensitive than the bootstrap-based criterion to increasing noise levels in $\mathbf{y}$ in this case.

As for the PLSGLR framework, our simulation results, based on two specific distributions (binomial and Poisson), lead us to recommend this new bootstrap-based criterion. Indeed, in the PLS-LR case, we show that, depending on the statistic used (testing NMSE or number of misclassified predictions) to test predictive ability, the bootstrap-based criterion is never significantly worse than either the CV-MClassed or p_val criteria. Concerning results obtained for a response vector following a Poisson distribution, the bootstrap-based criterion is the only one which returns consistent results, by retaining a decreasing number of components while the random noise level in $\mathbf{y}$ increases. Adding this to the MSE analysis and the obtained t-test results, it is reasonable to advise using the new criterion in this framework.

Troisième partie

Détermination de prédicteurs significatifs en PLS

Chapitre 6

Vers des améliorations de méthodes de sélection

6.1 Objectifs et motivations

La détermination de prédicteurs significatifs est un objectif important pour de multiples raisons, en partie exposées dans la partie 2.4.2. Dans le domaine de la chimiométrie, Höskuldsson (2001) évoque ainsi également l'importance de la sélection des prédicteurs quant aux performances de la régression PLS. Ainsi, outre la régression *Sparse* PLS introduite dans la partie 2.4.2, de multiples procédés ont été développés dans le cadre de la régression PLS. Un aperçu de ces méthodes est fourni par Gauchi and Chagnon (2001) qui effectue dans cet article une comparaison d'une vingtaine de procédés de sélection et une intéressante revue de la littérature à ce sujet a été effectuée par Mehmood *et al.* (2012). Cette revue de la littérature aura pour conclusion que si de nombreuses méthodologies ont été développées, aucune ne peut être considérée comme étant meilleure que les autres.

L'ensemble des procédés peut être globalement séparé en deux ou trois catégories définies respectivement par Lazraq *et al.* (2003) et Mehmood *et al.* (2012). La première catégorie, commune aux deux articles et dénommée *dimension-wise selection* par Lazraq *et al.* (2003) ou *embedded methods* par Mehmood *et al.* (2012), regroupe les méthodes procédant à la sélection de variable à chaque création d'une nouvelle composante (dimension) tel que la *Sparse* PLS. La seconde catégorie, dénommée *model-wise selection* par Lazraq *et al.* (2003) regroupe les procédés consistant en l'établissement d'un modèle PLS complet en premier lieu avant d'effectuer une sélection des variables. Mehmood *et al.* (2012) sépare cette catégorie en deux sous-groupes. Le premier regroupe les procédés nommés *filter methods* consistant en l'élaboration d'un unique modèle de régression PLS à partir duquel la sélection des prédicteurs sera effectuée. Le second regroupe les procédés dénommés *wrapper methods* consistant, dans les grandes lignes, en l'utilisation des méthodes du premier sous-groupe de façon itérative (Mehmood *et al.*, 2012). Nous retrouvons dans la catégorie *model-wise selection* des techniques développées à l'aide du bootstrap qui, à la vue de ce travail de thèse, ont tout particulièrement retenu notre attention.

L'introduction de techniques bootstrap pour la sélection de prédicteurs est initialement

proposée par Lazraq *et al.* (2003). Ce procédé consiste, à l'aide du bootstrap par paires effectué sur le couple $(\mathbf{y}, \mathbf{X})$, en l'établissement d'IC pour chaque variable, les IC contenant 0 conduisant ainsi à l'exclusion de la variable liée. Notons que dans ce cadre là, l'établissement d'IC est effectué en supposant la normalité des distributions liées aux coefficients β_K^{PLS} obtenus. Le procédé de bootstrap, dans ce cadre là, est donc uniquement utilisé afin de permettre l'établissement d'une estimation de la moyenne et de la variance liées à chaque coefficient de régression. Il conclura ainsi que ce procédé, sans être uniformément meilleur que celui conseillé par Gauchi and Chagnon (2001) après sa comparaison, retourne des résultats satisfaisant et peut-être très utile dans certaines situations. Un second procédé, proposé par Bastien *et al.* (2005), consiste en l'utilisation du bootstrap par paires effectué sur le couple $(\mathbf{y}, \mathbf{T_K})$ afin, à nouveau, d'extraire des IC pour chaque prédicteur.

Nous nous sommes ainsi plus particulièrement intéressés à deux procédés, chacun représentant, à notre avis, les méthodes de référence de chacune des deux catégories explicitées ci-dessus. Le premier procédé est la *Sparse* PLS développée par Chun and Keleş (2010) et appartenant à la première catégorie. Le second est le procédé utilisant le bootstrap développé par Lazraq *et al.* (2003). Ces deux procédés, comme l'ensemble des méthodes liées à la régression PLS, sont naturellement extrêmement dépendant du choix de nombre de composantes.

La *Sparse* PLS possède cette particularité, à la vue de son algorithme (2.4.2), que pour une valeur de contrainte fixée, le nombre de prédicteurs sélectionné est une fonction croissante du nombre de composantes. Ainsi, sélectionner le nombre de composantes par VC rend ce procédé de sélection de variables, de façon pratique, beaucoup moins précis. Ceci est dû à la forte variabilité des résultats obtenus par VC comme nous avons pu l'expliquer et l'observer dans les chapitres précédents. Des illustrations de ce problème à l'aide de graphiques sont proposées dans le prochain chapitre (Figures 7.1 et 7.7).

Le procédé basé sur le bootstrap est lui aussi extrêmement dépendant du nombre de composantes déterminé initialement sur la base de données étudiée. En effet, une fois sa détermination arrêtée, celui-ci est gardé fixe pour l'établissement de tous les modèles liés aux différents échantillons bootstrap. La détermination d'un nombre non optimal de composantes peut donc modifier considérablement les variables sélectionnées à l'issue du processus. De plus, il semble raisonnable de se questionner sur l'hypothèse qui est implicitement faite par Lazraq *et al.* (2003), à savoir que le nombre de composantes reste le même pour chaque échantillon bootstrap obtenu.

Ces faits, en pratique, rendent l'application de ces méthodes de sélection bien périlleuse. Le second objectif de cette thèse a donc été de développer des méthodes de sélection plus fiables en intégrant notamment notre nouveau critère de sélection du nombre de composantes par bootstrap que nous avons montré être plus stable que les méthodes basées sur la VC (Chapitre 5).

6.2 Modifications des procédés envisagées

Nous avons tout d'abord proposé une adaptation à la *Sparse* PLS de notre nouveau critère d'arrêt dans la construction de composantes PLS. Ainsi, pour chaque valeur du paramètre de parcimonie, une détermination du nombre de composantes optimale est effectuée à l'aide de

notre nouveau critère. Pour ce faire, étant donnée qu'à chaque itération de l'algorithme de la régression *Sparse* PLS le nombre de variables entrant dans la construction des composantes peut être modifié, il en devient nécessaire de retester l'ensemble des composantes à chaque étape jusqu'à l'obtention d'une itération où une des composantes est détectée comme étant non-significative. Ainsi, au lieu de déterminer le meilleur couple d'hyperparamètre par VC, il suffit de déterminer le meilleur paramètre de parcimonie par VC parmi ceux proposés par l'utilisateur. En effet, désormais, à chaque valeur correspond un unique modèle caractérisé par un nombre de composantes déterminé à l'aide de notre critère basé sur le bootstrap. Le nombre de modèles à comparer se voit donc considérablement réduit.

Un second procédé de sélection, basé sur la méthode introduite par Lazraq *et al.* (2003), a été développé. Ce développement fait suite à des résultats théoriques développés et présentés en Annexe G amenant des indications sur le nombre de composantes lié à un échantillon bootstrap. Ces résultats fournissent en tout état de cause une preuve du fait que le nombre de composantes lié à un échantillon bootstrap n'ait pas de raison d'être le même que celui lié au jeu de données originel. Or, la méthode proposé par Lazraq *et al.* (2003) en fait indirectement l'hypothèse.

Nous avons ainsi développé une méthode que nous avons qualifiée de dynamique. En effet, celle-ci consiste en l'utilisation du bootstrap par paires sur le couple $(\mathbf{y}, \mathbf{X})$ couplée à la détermination, pour chaque échantillon bootstrap ainsi obtenu, du nombre de composantes optimal à l'aide de notre nouveau critère d'arrêt. Il est important de noter que cette méthode a été rendue envisageable par le développement de notre nouveau critère d'arrêt dans la construction de composantes PLS et par les propriétés qui lui sont allouées, notamment concernant sa stabilité. A des fins de comparaison, la même procédure a été développée mais en déterminant le nombre de composantes lié à chacun des échantillons bootstrap à l'aide du critère du BIC adapté.

Enfin précisons que, grâce à l'universalité du critère développé et présenté dans le chapitre précédent, il nous a été possible d'adapter ces deux procédés au cadre de l'extension de la régression PLS aux modèles linéaires généralisés. Les détails sur ces deux procédés sont fournis dans les sections 7.2.1 et 7.2.2 du chapitre suivant.

6.3 Méthodologie et résultats

Une fois les codes développés, nous nous sommes intéressés à comparer les capacités et les caractéristiques des différents procédés présentés ci-avant. Précisons que nous avons également inclus la régression Lasso comme méthode de référence, méthode introduite par Tibshirani (1996) et développée sous sa forme algorithmique nommée *LARS*-Lasso par Efron *et al.* (2004).

Dans le cadre de la régression PLS usuelle, la comparaison entre ces différentes méthodes couplées à différents critères d'arrêt dans la construction de composantes a été effectuée à l'aide de données issues de simulations. Précisons tout de même que seules les réponses ont été simulées. En effet, afin de garder toute la complexité liée à une base de données réelle, nous avons utilisé comme matrice $\mathbf{X}$ une base de donnée issue d'une étude sur puces à ADN.

Pour plus de détails, nous renvoyons le lecteur à la partie 7.3.2 du chapitre suivant. Les réponses ont ensuite simulées suivant différents niveaux de bruits afin de suivre l'évolution de la pertinence et des caractéristiques des différentes méthodes prises en compte.

Dans le cadre de la régression PLS généralisée, nous avons comparé les méthodes dans le cas logistique. Ce choix provient essentiellement de deux facteurs. Le premier étant que des adaptations de la *Sparse* PLS au cadre logistique ont été effectuées par Chung and Keles (2010) et Durif *et al.* (2015), nous permettant ainsi de comparer nos développements à ces méthodes. En effet, bien que le procédé développé par Chung and Keles (2010) soit également adapté au problème multi-classe, il n'existe pas, à notre connaissance, d'adaptation de la *Sparse* PLS à d'autres types de distributions, nous empêchant ainsi la comparaison de notre nouveau procédé entièrement basé sur le bootstrap, qui lui est directement adaptable à tout type de distribution entrant dans le cadre des modèles linéaires généralisés, à d'autres procédés de sélection existants. Le deuxième facteur étant que l'étude issue d'une analyse sur puce à ADN que nous avons utilisé dans cette étude, a été établie à l'aide de cellules tumorales extraites du colon. Nous possédions ainsi l'information sur la localisation d'origine de la tumeur, à savoir sur la partie distale ou proximale. Il nous a ainsi été possible de coder cette information de façon binaire et d'étudier si l'expression génique d'un ou plusieurs probe sets pouvait être significative de la localisation d'origine de la tumeur. Des résultats biologiques connus nous permettant enfin de soutenir ou d'infirmer nos conclusions.

Nous avons particulièrement prêté attention à trois caractéristiques des méthodes qui nous ont semblé essentielles pour ce type de procédure de sélection. La première étant naturellement leur précision quant aux variables jugées significatives. Le deuxième point réside en la stabilité des résultats issus des différents procédés lorsque ceux-ci sont exécutés plusieurs fois sur la même base de données. Enfin, leurs performances prédictives ont également été comparées et ce en fonction de différents niveaux de bruit.

Dans le cadre de la régression PLS, nous avons ainsi pu vérifier que nos développements, sur l'ensemble des études menées, améliorent les procédés de base dont ils sont issus. Il nous a également été possible de procéder à une recommandation, au sens large, des méthodes à utiliser suivant le niveau de bruit présent dans la réponse et l'objectif recherché. En ce qui concerne les résultats liés au cadre logistique, notre nouveau procédé dynamique basé intégralement sur le bootstrap est lié à la meilleure stabilité et a été en capacité d'isoler un seul probe set comme étant significatif tout en gardant des performances prédictives similaires aux modèles obtenus à l'aide des autres procédés. De plus, ce probe set est biologiquement connu comme étant lié à une notion de spécificité de la localisation des tumeurs du colon, ce qui a renforcé nos conclusions issues de ce nouveau procédé de sélection.

Chapitre 7

New developments of Sparse PLS regressions

L'ensemble de ces recherches et des résultats qui en découlent ont fait l'objet d'un résumé soumis et accepté pour une communication sous la forme d'un poster à la 8$^{\text{th}}$ *International Conference of the ERCIM WG on Computational and Methodological Statistics* (CMStatistics 2015) qui s'est déroulé du 12 au 14 Décembre 2015 à Londres, UK. Ce chapitre est actuellement en cours de mise en forme pour soumission au journal *Computational Statistics and Data Analysis*. Des résultats théoriques quant au nombre de composantes à extraire sur un échantillon bootstrap ainsi que les démonstrations sont disponibles en Annexe G.

Abstract

Methods based on the so-called Partial Least Squares (PLS) regression, which recently gained much attention in the analysis of high-dimensional genomic datasets, were developed since the early 2000s to perform variable selection. Most of these techniques rely on some tuning parameters that are commonly determined by cross-validation (CV) based methods, which raise important stability issues. Therefore, we developed a new dynamic bootstrap-based method for significant predictors selection, suitable for both PLS regression and its incorporation into generalized linear models (GPLS). It relies on the establishment of bootstrap confidence intervals that both allow to test the significances of predictors at a preset first species risk α and avoid the use of CV. We also developed an adapted version of the Sparse PLS (SPLS) and SGPLS regression, using a recently introduced non-parametric bootstrap-based technique for the determination of the numbers of components. We compared their variable selection reliability, as well as their stability concerning the tuning parameters determination and their predictive ability, using datasets simulations for PLS framework and real microarray gene expression datasets for PLS-Logistic classification. We observe that our new dynamic bootstrap-based method features an interesting property in that it succeed better than all the other approaches in separating the random noise in $\mathbf{y}$ from the relevant information, leading thus to a better accuracy and predictive abilities, especially for non-negligible noise levels.

Supplementary material is linked to this article.

Keywords: Variable selection, PLS, GPLS, Bootstrap, Stability.

7.1 Introduction

Partial Least Squares (PLS) regression, which was introduced by (Wold *et al.*, 1983), is a well known dimension reduction method, notably in chemometrics and spectrometric modeling (Wold *et al.*, 2001). In this paper, we focus on the PLS univariate response framework, better known as PLS1. Let n be the number of observations and p the number of covariates. Then, $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ represents the response vector, with $(.)^T$ denoting the transpose. The original underlying algorithm, which was developed to deal with continuous responses, consists in building latent variables $\mathbf{t}_k,\ 1 \leqslant k \leqslant K$, also called components, as linear combinations of the original predictors $\mathbf{X} = (\mathbf{x}_1, \ldots \mathbf{x}_p) \in \mathcal{M}_{n,p}(\mathbb{R})$, where $\in \mathcal{M}_{n,p}(\mathbb{R})$ represents the set of matrices of n rows and p columns, so that,

$$\mathbf{t}_k = \mathbf{X}_{k-1}\mathbf{w}_k,\ 1 \leqslant k \leqslant K, \tag{7.1}$$

where $\mathbf{X}_0 = \mathbf{X}$, and $\mathbf{X}_{k-1},\ k \geqslant 2$, represents the residual covariate matrix obtained through the OLSR of $\mathbf{X}_{k-2}$ on $\mathbf{t}_{k-1}$. $\mathbf{w}_k = (w_{1k}, \ldots, w_{pk})^T \in \mathbb{R}^p$ is obtained as the solution of the following maximization problem (Boulesteix and Strimmer, 2007):

$$\begin{aligned} \mathbf{w}_k =& \underset{\mathbf{w}\in\mathbb{R}^p}{\operatorname{argmax}} \left\{\operatorname{Cov}^2(\mathbf{y}_{k-1}, \mathbf{t}_k)\right\} && (7.2)\\ =& \underset{\mathbf{w}\in\mathbb{R}^p}{\operatorname{argmax}} \left\{\mathbf{w}^T\mathbf{X}_{k-1}^T\mathbf{y}_{k-1}\mathbf{y}_{k-1}^T\mathbf{X}_{k-1}\mathbf{w}\right\}, && (7.3) \end{aligned}$$

with the constraint $\|\mathbf{w}_k\|_2^2 = 1$, and where $\mathbf{y}_0 = \mathbf{y}$, and $\mathbf{y}_{k-1},\ k \geqslant 2$, represents the residual vector obtained from the OLSR of $\mathbf{y}_{k-2}$ on $\mathbf{t}_{k-1}$.

The final regression model is thus:

$$\begin{aligned} \mathbf{y} =& \sum_{k=1}^{K} c_k\mathbf{t}_k + \epsilon && (7.4)\\ =& \sum_{j=1}^{p} \beta_j^{\mathrm{PLS}}\mathbf{x}_j + \epsilon, && (7.5) \end{aligned}$$

with $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ the $n \times 1$ error vector and $(c_1, \ldots, c_K)$ the regression coefficients related to the OLSR of $\mathbf{y}_{k-1}$ on $\mathbf{t}_k$, $\forall k \in [\![1, K]\!]$, also known as $\mathbf{y}$-loadings.

More details are available notably in Höskuldsson (1988) or Tenenhaus (1998).

This particular regression technique based on a reduction of the original dimension avoids any matrix inversion or diagonalization by only using deflation. It permits to deal with high-dimensional datasets efficiently and notably solves the collinearity problem (Wold *et al.*, 1984).

With the high technological advances of these last decades, the PLS regression has been gaining more attention and has been successfully applied in many others domains, notably

in the genomic area. The development of both microarray and allelotyping techniques result in high-dimensional datasets from which information has to be efficiently extracted. To this end, PLS regression has become a benchmark as an efficient statistical method for prediction, regression and dimension reduction (Boulesteix and Strimmer, 2007). Practically speaking, the observed response related to such studies does commonly not follow a continuous distribution. Regular aims of gene expressions datasets are to describe classification problems, such as cancer stages, diseases relapse events or tumor classification. Therefore, PLS regression had to be adapted to take into account these specific discrete distributions of responses. This problem has been an intensive subject of research during these last years, leading to globally two types of adapted PLS regression for classification. The first one, studied and developed notably by Nguyen and Rocke (2002b), Nguyen and Rocke (2002a) and Boulesteix (2004), is a two-stages methodology. The first step consists in building the standard PLS components by treating the response as a continuous one and, in a second step, classification methods are performed, such as logistic discrimination (LD) or quadratic discriminant analysis (QDA). A second type of adapted PLS regression consists in building the PLS components using either an adapted or the original Iteratively Reweighted Least Squares (IRLS) algorithm and finally performing a generalized linear regression on these components. Such a method was first introduced by Marx (1996). Different adaptations or improvements, using notably Ridge regressions (Le Cessie and Van Houwelingen, 1992) or Firth's procedure (Firth, 1993) to avoid non-convergence or infinite parameter estimations, had then been developed notably by Nguyen and Rocke (2004), Ding and Gentleman (2005), Fort and Lambert-Lacroix (2005) or Bastien *et al.* (2005). In this work, we focus on the second type of adapted PLS regression, referred to as GPLS.

As previously mentioned, a particularity of these datasets lie in their high-dimensional setting *i.e.* $n \ll p$. Chun and Keleş (2010) demonstrate that the asymptotic consistency of the PLS estimators does not hold in the very large p and small n case, so that filtering or selecting predictors become necessary to obtain consistent estimations of parameters. However, all the methods described above proceed to classification fitting using the entire set of predictors. These datasets containing frequently thousands of predictors, like for microarray datasets, a variable filtering pre-processing should be applied. A commonly used pre-processing method for classification relies on the BSS / WSS-statistic which is calculated as:

$$\mathrm{BSS}_j \,/\, \mathrm{WSS}_j = \frac{\sum_{q=1}^{Q} \sum_{i:y_i \in G_q} (\hat{\mu}_{jq} - \hat{\mu}_j)^2}{\sum_{q=1}^{Q} \sum_{i:y_i \in G_q} (x_{ij} - \hat{\mu}_{jq})^2}, \tag{7.6}$$

with $\hat{\mu}_j$ the sample mean of $\mathbf{x}_j$ and $\hat{\mu}_{jq}$ the sample mean of $\mathbf{x}_j$ within class G_q for $q \in \{1, \ldots, Q\}$. Then, predictors related to highest values are retained, but no specific rule exists to choose the number of predictors to retain. A Bayesian based technique, available in the R-package *limma*, became a common way to avoid this concern by computing a Bayesian based p-value for each predictor and, therefore, allowing users to choose a number of relevant predictors based on a threshold p-value (Smyth, 2004).

This method could not be considered as a parsimonious one but rather as a pre-processing stage for the exclusion of uninformative covariates. Reliably selecting relevant predictors in PLS regression has several interests. Practically speaking, it would allow the users to identify the original covariates which are significantly linked to the response, like is done for OLS regression with Student-type tests. Statistically speaking, it would avoid the establishment of over-complex models and insure the consistency of PLS estimators. Several methods for variable selection have already been developed (Mehmood *et al.*, 2012). Lazraq *et al.* (2003) group these techniques into two main categories. The first one, named the model-wise selection category, consists of first establishing the PLS model before performing a variable selection. The second one, which is called the dimension-wise selection category, consists of selecting variables on one PLS component at a time.

A dimension-wise method, introduced by Chun and Keleş (2010) and called Sparse PLS (SPLS), has become a benchmark for selecting relevant predictors by using PLS methodology. This technique is adapted for continuous response and consists of a simultaneous dimension reduction and variable selection, computing sparse linear combinations of covariates as PLS components. This is achieved by introducing an L_1 constraint on the weight vectors $\mathbf{w}_k$, leading to the following formulation of the objective function:

$$\mathbf{w}_k = \underset{\mathbf{w}\in\mathbb{R}^p}{\operatorname{argmax}} \left\{\mathbf{w}^T\mathbf{X}_{k-1}^T\mathbf{y}_{k-1}\mathbf{y}_{k-1}^T\mathbf{X}_{k-1}\mathbf{w}\right\}, \text{ s.c. } \|\mathbf{w}\|_2^2 = 1, \|\mathbf{w}\|_1 < \lambda \tag{7.7}$$

where λ determines the amount of sparsity. More details are available in Chun and Keleş (2010). Two tuning parameters are thus involved: $\eta \in [0, 1]$ as a rescaled parameter equivalent to λ, and the number of PLS components K, which are determined through CV-based mean squared error (CV MSE). We refer to this technique as SPLS CV in the following.

Chung and Keles (2010) developed an extension by integrating this SPLS technique into the generalized linear models (GLM) methodology, leading this technique to solve classification problems. They also integrate the Firth's procedure in order to deal with non-convergence issues. Both tuning parameters are selected using CV MSE. We refer to this method as SGPLS CV in the following.

A well known model-wise selection method was introduced by Lazraq *et al.* (2003). It consists of bootstrapping pairs $(\mathbf{y}_i, \mathbf{x}_{i\bullet})$, $1 \leqslant i \leqslant n$, where $\mathbf{x}_{i\bullet}$ represents the i^{th} row of $\mathbf{X}$, before applying a PLS regression with a preset number of components K on each bootstrap sample. By performing this method, approximations of distributions related to predictors' coefficients are achieved. It leads to bootstrap-based confidence intervals (CI) for each predictor and so opens up the possibility of directly testing their significance with a fixed first species risk α. Advantage of this method is twofold. First, this method, by focusing on PLS regression coefficients, is relevant for both PLS and GPLS frameworks. Second, it only depends on one single tuning parameter K which has to be previously determined.

An important related concern has to be mentioned. While performing this technique, approximations of distributions are achieved conditionally to a fixed dimension of the extracted subspace. In other words, it approximates the uncertainty of these coefficients conditionally to the fact that estimations are performed in a K-dimensional subspace for each bootstrap sample. The determination of an optimal number of components is crucial for achieving

reliable estimations of the regression coefficients (Wiklund *et al.*, 2007). Thus, since this determination is specific to the involved dataset, it should be performed for each bootstrap samples in order to obtain reliable bootstrap-based CI. We established some theoretical results which confirm this point (supplementary material, Annex G).

Determining tuning parameters by using q-fold cross-validation (q-CV) based criteria may induce important issues concerning the stability of extracted results (Hastie *et al.* (2009, p.249); Boulesteix (2014); (Magnanensi *et al.*, 2015)). Thus, using such criteria for successive determinations of the number of components should be avoided. As mentioned, amongst others, by Efron and Tibshirani (1993, p.255) and Kohavi (1995), bootstrap-based criteria are known to be more stable than CV-based ones. In this way, Magnanensi *et al.* (2015) developed a robust bootstrap-based criterion for the determination of the number of PLS components which is characterized by a high level of stability and suitable for both PLS and GPLS regression frameworks. Thus, this criterion opens up the possibility of reliable successive determinations for each bootstrap sample.

In this article, we introduce a new dynamic bootstrap-based technique for covariate selection suitable for both the PLS and GPLS frameworks. It consists in bootstrapping pairs $(\mathbf{y}_i, \mathbf{x}_{i\bullet})$ and extracting successively the optimal number of components for each bootstrap sample by using the previously mentioned bootstrap-based criterion. Through this development, our goal is to better approximate the uncertainty related to regression coefficients by removing the condition of extracting a fixed K-dimensional subspace for each bootstrap sample, leading to more reliable CI. This new method both avoids the use of CV and features the same advantages than those previously mentioned and related to the technique introduced by Lazraq *et al.* (2003). We refer to this new dynamic method as BootYTdyn in the following.

We also succeed in adapting the bootstrap-based criterion introduced by Magnanensi *et al.* (2015) for the determination of a unique optimal number of components related to each preset value of η in both the SPLS and SGPLS frameworks. Thus, these adapted versions, by reducing the use of CV, improve the reliability of the hyper-parameters tuning. We will refer to both these adapted techniques as SPLS BootYT and SGPLS BootYT, respectively.

The following paper is organized as follows. In Section 7.2, we introduce our new dynamic bootstrap-based technique followed by the description of our adaptation of the BootYT stopping criterion to the SPLS and SGPLS methods. In Section 7.3, we detail our different simulation plans related to the PLS framework. We also summarize the results, depending notably on different noise levels inserted in $\mathbf{y}$. In Section 7.4, we treat a real microarray gene expression dataset with a binary response, which represents the original localization of the colon tumors, by benchmarking our new dynamic bootstrap-based approach for GPLS regressions. Lastly, we discuss results and summarize our conclusions in Section 7.5.

7.2 Bootstrap-based approaches for predictors' selection

7.2.1 A new dynamic bootstrap-based technique

As mentioned in Section 7.1, the number of extracted components is crucial for a reliable estimation of β_j^{PLS}, $1 \leqslant j \leqslant p$ (Wiklund *et al.*, 2007). We demonstrated (supplementary material, Annex G) that determining an optimal number of components on the original dataset and using it to perform PLS regressions on the builded bootstrap samples, as done by Lazraq *et al.* (2003), is not suitable.

In order to take into account these theoretical results, we developed a new dynamic bootstrap-based approach for variable selection relevant for both PLS and GPLS frameworks. This approach consists in estimating the optimal number of components for each bootstrap sample created during the algorithm. To obtain consistent results, a robust and resample-stable stopping criterion in components construction has to be used. Let β_j be a lightened notation for β_j^{PLS}. The algorithm of this new dynamic bootstrap-based method is designed as follows:

1. Let D^{ori} be the original dataset and R the total number of bootstrap samples D_r^b, $r \in [\![1, R]\!]$.
2. $\forall r \in [\![1, R]\!]$:
   - Extract the number of PLS components that is needed for D_r^b using a preset stopping criterion.
   - Compute the estimations $\hat{\beta}_j^r$, $\forall j \in [\![1, p]\!]$ by fitting the relevant PLS or GPLS model.
3. $\forall j \in [\![1, p]\!]$, construct a $(100 \times (1-\alpha))\%$ bilateral BC_a CI, noted:
$$\text{CI}_j = [\text{CI}_{j,1}, \text{CI}_{j,2}].$$
4. $\forall j \in [\![1, p]\!]$, **If** $0 \notin \text{CI}_j$ **then** retain $\mathbf{x_j}$ **else** delete $\mathbf{x_j}$.
5. Obtain the reduced model $\mathcal{M}_{\text{sel}}$ by only integrating the significant predictors into and extracting the number of components K_{sel} determined by the preset stopping criterion.

Note that in this work, we fixed $\alpha = 0.05$.

7.2.2 An adapted bootstrap-based Sparse PLS implementation

As mentioned by Boulesteix (2014), using q-CV based methods for tuning parameters induces potential important issues, notably concerning the variability of the results due to their dependency on the way the folds are randomly formed. However, as detailed in Section 7.1, the determination of both tuning parameters involved in the SPLS regression developed by Chun and Keleş (2010) is performed using q-CV MSE. Therefore, to improve the reliability of this determination, we adapted the bootstrap-based stopping criterion to this method leading to the following algorithm:

1. Let $\{\eta_1, \ldots, \eta_s\}$ be the set of preset values for the sparsity parameter and $\{k_1, \ldots, k_s\} = \{1, \ldots, 1\}$ the set of initial number of components for each η_i. Let $i = 1$.

2. Let $\hat{c}_j^{\eta_i},\ j \in [\![1, k_i]\!]$ be the estimated regression coefficients of $\mathbf{y}$ on $\mathbf{T}_{k_i} = (\mathbf{t}_1, \ldots, \mathbf{t}_{k_i}) \in \mathcal{M}_{n,k_i}(\mathbb{R})$. Obtain k_i BC_a CI for $\hat{c}_j^{\eta_i},\ j \in [\![1, k_i]\!]$, using the bootstrap-based stopping criterion, noted:

$$\mathrm{CI}_j^{k_i} = \left[\mathrm{CI}_{j,1}^{k_i}, \mathrm{CI}_{j,2}^{k_i}\right].$$

3. **If** $\exists j \in [\![1, k_i]\!] \,|\, 0 \in \mathrm{CI}_j^{k_i}$ **then** $K_{opt}^{\eta_i} = k_i - 1$ **else** $\{k_i = k_i + 1$ and return to step 2$\}$.

4. **While** $i \neq s$ **then** $\{i = i + 1$ and return to step 2$\}$.

5. Return the set of extracted numbers of components $\left\{K_{opt}^{\eta_1}, \ldots, K_{opt}^{\eta_s}\right\}$ related to each $\eta_i,\ 1 \leqslant i \leqslant s$.

6. Return the couple $\left(\eta_{opt}, K_{opt}^{\eta_{opt}}\right)$ having the lowest CV-based MSE.

Retesting all the components obtained after each incrementation of k_i is essential since the original predictors involved in the components construction change when k_i grows up (Chun and Keleş, 2010). This fact, combined with the aim to keep orthogonality between the components, lead the components themselves to change, so that the significance of each component has to be retested at each step.

While the original implementation compare $K^{\max} \times s$ models through CV MSE, with $K^{\max}$ the maximal number of components which is set by the user, this new bootstrap-based version only focus on s models, since one single number of components is determined for each preset value of η. An illustration of the stability benefit obtained by using this new implementation, based on the simulated dataset introduced in Section 7.3.2 with $sd(\epsilon) = 5$, is performed in Fig.7.1.

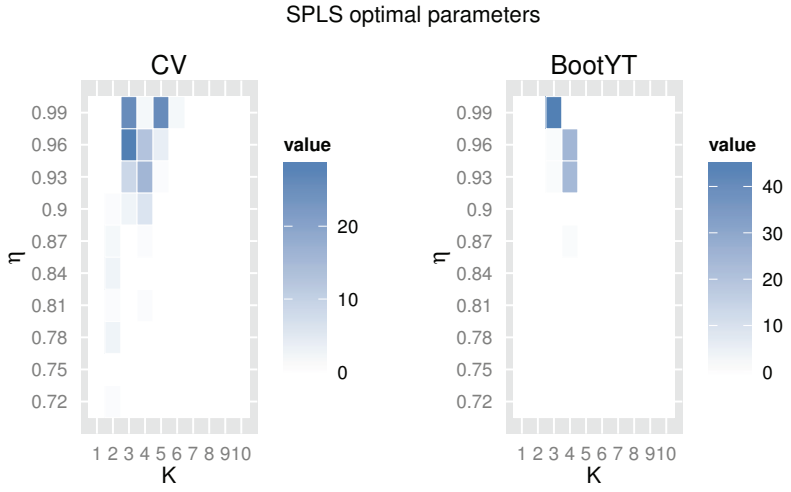


Figure 7.1: Repartition of 100 determinations of (η_{opt}, K_{opt}) using the original SPLS approach (Left) and the new bootstrap-based implementation (Right).

7.3 Simulations studies

7.3.1 Simulations for accuracy comparisons

These simulations are based on a simulation structure proposed by Chun and Keleş (2010, p.14) and modified in order to study high-dimensional settings. We consider the case where there are less observations than predictors *i.e.* $n < p$ and set $n = 100$ and $p = 200$ or $p = 1000$. Let q be the number of spurious predictors. While Chun and Keleş (2010, p.14) only consider a ratio q/p equal to 0.25 and 0.5, both the 0.05 and 0.95 ratio values have been added. Four independent and identically distributed hidden variables $\mathbf{h}_1, \ldots, \mathbf{h}_4$ following a $\mathcal{N}(0, 25I_n)$ distribution were computed. Then, columns of the covariate matrix $\mathbf{X}$ are generated by $\mathbf{x}_j = \mathbf{h}_l + \epsilon_j$ for $n_{j-1} \leqslant j \leqslant n_j$, where $j = 1, \ldots, 4$, $(n_0, \ldots, n_4) = (0, (p-q)/2, p-q, p-r, p)$, $r = 5$ when $p = 200$ and $r = 10$ when $p = 1000$, and $\epsilon_1, \ldots, \epsilon_p$ are drawn independently from a $\mathcal{N}(0, 0.1I_n)$. $\mathbf{y}$ is generated by $3\mathbf{h}_1 - 4\mathbf{h}_2 + f$, where f is normally distributed with mean 0 and variance such that the signal-to-noise ratio (SNR) equals 10.

Using this simulation plan, accuracy of both the SPLS technique using 10 fold-CV for the determination of tuning parameters (SPLS CV) and our new dynamic bootstrap method combined with the bootstrap-based stopping criterion (BootYTdyn) is compared. In order to do so, for each parameter setting, 50 determinations of the sparse support related to both methods were established. Lastly, mean values of accuracy over these 50 trials were calculated. Results are summarized in the following Table 7.1.

Table 7.1: Mean accuracy values (SNR).

p	200				1000			
q/p	0.05	0.25	0.5	0.95	0.05	0.25	0.5	0.95
SPLS CV	0.986	0.961	0.849	0.591	0.998	0.997	0.989	0.963
BootYTdyn	1.000	0.867	0.805	0.982	0.967	0.827	0.893	0.985

Based on these results, the SPLS CV features a better accuracy than the BootYTdyn for values of ratio q/p that are not close of 0 or 1. While both of them are feature good performances when this ratio is close of 0, *i.e.* when a major part of predictors are significant, the BootYTdyn outperforms the SPLS CV when only a small proportion of predictors is significant.

Nevertheless, through this simulation plan, covariates are collected into four groups. While within-group correlations between covariates are close to one, the between-group correlations are close to zero. This unrealistic scheme causes the irrelevance of the determination of an optimal support and seems better appropriate to the determination of the number of components. As an illustration, 50 additional samples in the case of $p = 1000$ and $q/p = 0.5$ have been simulated. We then calculated the predictive MSE (PMSE) based on four different supports S_1, S_2, S_3, S_4 where $S_1 = \{\mathbf{x}_j, 1 \leqslant j \leqslant p\}$, $S_2 = \{\mathbf{x}_j, 1 \leqslant j \leqslant (p-q)\}$, $S_3 = \{\mathbf{x}_1, \mathbf{x}_{251}\}$ and $S_4 = \{\mathbf{x}_1, \mathbf{x}_{251}\} \cup \{\mathbf{x}_j, (p-q)+1 \leqslant j \leqslant p\}$. Results are summarized in Table 7.2.

In the light of these observations, the aim of this simulation scheme rather lie in extracting an optimal number of components than an optimal support. We thus decided to use a real dataset as covariate matrix for a more global and relevant comparison.

Table 7.2: PMSE values for different supports.

	S_1	S_2	S_3	S_4
K=1	65.383	62.702	63.988	856.779
K=2	63.957	64.443	65.695	209.873
K=3	65.745	76.197	NA	61.801

7.3.2 Simulations for global comparisons

Datasets simulations

In this study, we used a real microarray gene expression dataset, which was established on fresh-frozen primary tumors samples, from a multicenter cohort, with stage I to IV of colon cancer. 566 samples fulfilled RNA quality requirement, constituting our database. These samples were split into a 443 samples discovery set and a 123 samples test set, well balanced for the main anatomoclinical characteristics. This database has already been studied by Marisa *et al.* (2013) and more precisions on it are also available in this article.

In order to save computational time, a first selection of 100 predictors was performed. Based on the original localization of the tumors as a response vector and on the full 566 samples, the 100 most differentially expressed probe sets were extracted. As mentioned in Section 7.1, this pre-processing is based on a Bayesian technique and results in our benchmark predictors matrix.

Then, based on correlation values, four positively correlated predictors were selected to form our set of significant covariates (supplementary material, Annex H). Thus, let $\mathbf{X}_{sel} = (\mathbf{x}_1, \mathbf{x}_{12}, \mathbf{x}_{15}, \mathbf{x}_{59}) \in \mathcal{M}_{n,4}(\mathbb{R})$ be the matrix composed of these selected predictors so that $\mathbf{y}$ is simulated as follows:

$$\mathbf{y} = \mathbf{X}_{sel}\beta + \epsilon \tag{7.8}$$

with $\beta = (3.559, 2.071, 1.440, 1.770)^T$, $\mathbb{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2 I_n$.

We performed simulations for six distinct levels of random noise standard deviation in order to follow the performances of the different methods. Both these standard deviations and their related SNR are summarized in Table 7.3.

Table 7.3: Noise standard deviation (SNR).

dataset 1	dataset 2	dataset 3	dataset 4	dataset 5	dataset 6
0.5 (810.603)	1 (202.651)	3 (22.517)	4 (12.666)	5 (8.106)	6.366 (5.000)

Benchmarked methods

Using these simulated datasets, eight methods were analyzed and compared.

1. **Q^2**. The bootstrap-based method, introduced by Lazraq *et al.* (2003), combined with the 10-fold CV-based Q^2 criterion (Tenenhaus, 1998, p.83) for previously selecting the number of components.

2. **BIC**. The bootstrap-based method, introduced by Lazraq *et al.* (2003), combined with the corrected BIC using the estimated degrees of freedom (DoF) (Krämer and Sugiyama, 2011) for previously selecting the number of components.

3. **BootYT**. The bootstrap-based method, introduced by Lazraq *et al.* (2003), combined with the bootstrap-based criterion (Magnanensi *et al.*, 2015) for previously selecting the number of components.

4. **BICdyn**. The new dynamic bootstrap-based method combined with the corrected BIC criterion for successive determinations of numbers of components.

5. **BootYTdyn**. The new dynamic bootstrap-based method combined with the bootstrap-based criterion for successive determinations of numbers of components.

6. **SPLS CV**. The original SPLS method using 10-fold CV for tuning parameters determination (Chun and Keleş, 2010).

7. **SPLS BootYT**. The new adapted SPLS version using the bootstrap-based criterion for components selection.

8. **Lasso**. The Lasso regression included as a benchmark (Efron *et al.*, 2004).

Simulation plan and notations

In order to perform reliable comparisons between these eight methods, each type of determination was performed a hundred times. Numbers of components, sparse supports or sparse tuning parameters are the main examples of these determinations. Results linked to the highest occurrence rates are then chosen for methods comparison. All bootstrap-based techniques were performed with $R = 1000$ bootstrap samples and each related CI was achieved with a first species risk $\alpha = 0.05$.

The global comparison is twofold. First, in order to compare accuracy and stability related to each technique, we focus on both different supports and models extracted by the different methods of variable selection. Indeed, in PLS framework, a specific model is resulting from both a set of predictors and a specific number of components. Due to the sparsity parameter of the SPLS approaches, the same support could be extracted but with a different number of components leading to different models. The Lasso regression can also extract the same support for several different sparsity parameters leading to different estimations of models' coefficients. Therefore, the following notations related to each specific variable selection technique have to be introduced.

- $\{\mathcal{S}_1, \ldots, \mathcal{S}_{\Gamma_1}\}$, the set of different extracted supports.
- $\{\mathcal{M}_1, \ldots, \mathcal{M}_{\Gamma_2}\}$, the set of different fitted models.
- $\mathcal{S}_{sel}$, the selected support *i.e.* featuring the highest occurrence rate.
- $\mathcal{M}_{sel}$, the selected model *i.e.* featuring the highest occurrence rate.

- $\%\mathcal{S}_{sel}$, occurrence rate of the selected support.
- $\%\mathcal{M}_{sel}$, occurrence rate of the selected model.
- K_{sel}, the number of components related to the selected model.

Second, in order to compare predictive abilities of selected models, 10-fold CV MSE related to each selected sparse model through PLS regression were computed a hundred times. The test set was also used in order to confirm results obtained by CV.

Note that, concerning the dynamic BIC-based method for $sd\,(\epsilon) = 0.5$ only 97 trials performed well. Lastly, due to equality of occurrence rates between the two mostly represented couples of tuning parameters, results of the SPLS CV for $\text{sd}(\epsilon) = 5$ were obtained on 150 trials.

Stability and accuracy results.

Both the means of accuracy values over the available trials in each case and stability results based on the extracted supports are summarized in Tables 7.4, 7.5 and 7.6. The numbers of components used for the original bootstrap-based approach (Lazraq *et al.*, 2003) are summarized in supplementary material (Annex I).

Table 7.4: Means of accuracy values.

	Q^2	BIC	BICdyn	BootYT	BootYTdyn	SPLS CV	SPLS BootYT	Lasso
$\text{sd}(\epsilon) = 0.5$	0.9331	0.9710	0.9882	0.9587	0.9718	0.9914	0.9960	0.9572
$\text{sd}(\epsilon) = 1$	0.9370	0.9503	0.9575	0.9557	0.9781	0.9915	1.0000	0.9689
$\text{sd}(\epsilon) = 3$	0.8730	0.9353	0.9576	0.9614	0.9837	0.9821	0.9799	0.9741
$\text{sd}(\epsilon) = 4$	0.8004	0.9289	0.9397	0.9692	0.9928	0.9771	0.9799	0.9686
$\text{sd}(\epsilon) = 5$	0.8327	0.9176	0.9444	0.9557	0.9876	0.9790	0.9841	0.9676
$\text{sd}(\epsilon) = 6.366$	0.8970	0.8755	0.9347	0.9625	0.9820	0.9745	0.9714	0.9731

Table 7.5: Number Γ_1 of different extracted supports ($\%\mathcal{S}_{sel}$).

	Q^2	BIC	BICdyn	BootYT	BootYTdyn	SPLS CV	SPLS BootYT	Lasso
$\text{sd}(\epsilon) = 0.5$	20 (17)	23 (10)	6 (73.2)	11 (30)	16 (35)	5 (56)	2 (90)	4 (48)
$\text{sd}(\epsilon) = 1$	18 (30)	8 (57)	7 (38)	11 (26)	17 (23)	6 (53)	1 (100)	5 (49)
$\text{sd}(\epsilon) = 3$	41 (19)	16 (47)	14 (39)	26 (16)	6 (44)	5 (34)	4 (90)	4 (58)
$\text{sd}(\epsilon) = 4$	96 (2)	6 (58)	11 (26)	18 (30)	6 (48)	12 (48)	4 (67)	4 (69)
$\text{sd}(\epsilon) = 5$	88 (3)	11 (48)	17 (33)	12 (22)	6 (54)	11 (25.33)	4 (45)	3 (64)
$\text{sd}(\epsilon) = 6.366$	47 (10)	25 (24)	9 (57)	10 (39)	5 (38)	10 (18)	4 (46)	3 (64)

Concerning the number of different models, results related to the Lasso regression are the same than that obtained concerning the number of different supports. Only the result for $\text{sd}(\epsilon) = 0.5$ differs since one trial concludes in a fifth value of sparsity parameter and the same support as the selected one. Therefore, only results concerning models established using the SPLS methods are summarized in the following Table 7.7. In the case of bootstrap-based techniques, supports and models are similar since no sparsity parameter is needed.

Table 7.6: Number of predictors in $\mathcal{S}_{sel}$ (K_{sel}).

	Q^2	BIC	BICdyn	BootYT	BootYTdyn	SPLS CV	SPLS BootYT	Lasso
$\text{sd}(\epsilon) = 0.5$	11 (5)	6 (4)	5 (5)	8 (5)	7 (6)	4 (4)	4 (4)	4 (4)
$\text{sd}(\epsilon) = 1$	10 (4)	9 (5)	9 (5)	9 (5)	6 (4)	4 (4)	4 (4)	5 (3)
$\text{sd}(\epsilon) = 3$	16 (4)	11 (6)	8 (6)	7 (4)	6 (4)	8 (4)	6 (4)	7 (4)
$\text{sd}(\epsilon) = 4$	24 (3)	11 (5)	10 (5)	6 (4)	5 (3)	6 (4)	6 (4)	7 (4)
$\text{sd}(\epsilon) = 5$	21 (3)	12 (5)	10 (5)	9 (4)	3 (3)	5 (3)	3 (3)	7 (4)
$\text{sd}(\epsilon) = 6.366$	15 (3)	16 (4)	11 (4)	7 (4)	3 (3)	4 (3)	5 (3)	7 (4)

Table 7.7: Results of SPLS models stability.

	$\#\,(\eta_{opt}, K_{opt})$ $(\%\,(\eta, K)_{sel})$		Γ_2 $(\%\mathcal{M}_{sel})$	
	SPLS CV	SPLS BootYT	SPLS CV	SPLS BootYT
$\text{sd}(\epsilon) = 0.5$	9 (46)	3 (81)	8 (55)	2 (90)
$\text{sd}(\epsilon) = 1$	13 (36)	2 (73)	9 (45)	1 (100)
$\text{sd}(\epsilon) = 3$	13 (17)	5 (59)	5 (34)	4 (90)
$\text{sd}(\epsilon) = 4$	15 (27)	6 (37)	12 (48)	4 (67)
$\text{sd}(\epsilon) = 5$	20 (19.33)	6 (45)	12 (25.33)	4 (45)
$\text{sd}(\epsilon) = 6.366$	16 (18)	4 (46)	11 (18)	4 (46)

Concerning the three bootstrap-based techniques and in the light of accuracy results (Table 7.4), the BootYT outperforms both the others excepted for the case where $\text{sd}(\epsilon) = 0.5$ for which the BIC has to be advised. This exception is confirmed through the comparison of both dynamic bootstrap-based methods, where the $\text{sd}(\epsilon) = 0.5$ case is the only one where using the BIC criterion represents the most relevant choice as well. This phenomenon matches with conclusion obtained by Magnanensi *et al.* (2015) in that the BIC criterion is well designed for small values of noise variance while the BootYT criterion outperforms for non-negligible levels of random noise. Using the Q^2 criterion for selecting the number of components appears to never be reliable option so that combining this criterion to our new dynamic approach was not performed. Accuracy values highlight that our new dynamic method always improves the original one. Concerning both versions of SPLS regressions, both of them featured similar values of accuracy. These techniques stand out to perform better than the others for the smallest levels of random noise variance while the BootYTdyn outperforms all the others for the highest values of noise variability.

Based on results introduced in Tables 7.5 and 7.7, the Q^2-based approach for predictors selection features important stability issues, providing between 18 and 96 different models on 100 simulations. Depending on both the noise variability and the criterion combined to our dynamic approach, the latter improves the original one by stabilizing the choice of a sparse supports. Cases where this is observed match with previous conclusions established when analyzing accuracy results, namely both the BIC-based dynamic approach for small values of noise variability and the BootYTdyn used for datasets with non-negligible noise variances, strengthening the fact that the BIC criterion is well designed for small values of noise variance while the BootYT criterion outperforms for non-negligible ones. Concerning SPLS methods, our new bootstrap-based version gains in stability in that it retains less different optimal couples (η_{opt}, K_{opt}) and also sparse models than the original does. Moreover,

since $\Gamma_1 = \Gamma_2$ for all studied datasets, it directly implies that it retains a unique optimal number of components for each sparse support. It so permits to choose the optimal model in a more reliable way than by using the CV-based technique. Lastly, the Lasso regression has an interesting stability ability leading the latter, the BootYTdyn and the SPLS BootYT to be advised in terms of stability.

The last descriptive statistic concerns the number of significant predictors retained (Table 7.6). The Q^2-based approach is the least sparse one by selecting the highest number of covariates. Both the BICdyn and BootYTdyn methods improve respectively the BIC and BootYT ones by featuring a better accuracy related to their selected support. Indeed, the four expected covariates are always included in these selected supports. Once again, the BootYT criterion has to be advised for datasets with non-negligible random noise variability compared to the corrected BIC criterion which has to be applied for small value of random variance. Indeed, both the BIC and BICdyn approach, by retaining a globally increasing number of predictors while the random noise standard deviation increases, let us suppose that it uses some predictors to model this random noise leading to over-fitting. On the contrary, while the BootYT approach selects a stable number of significant predictors, the BootYTdyn conclude in a decreasing number of significant predictors while the random noise standard deviation increases, which corresponds to an expected phenomenon. To confirm the over-fitting issues of the BIC criterion, the 10-fold CV-based MSE was reported in Fig.7.2

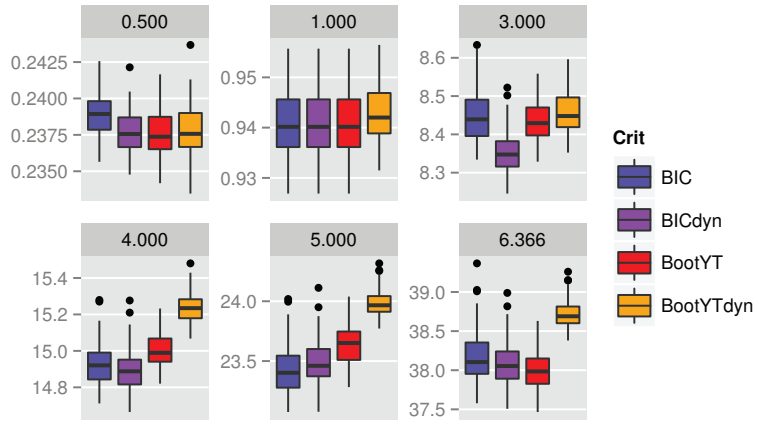


Figure 7.2: Boxplot of 10-CV MSE based on $\mathbf{y}$ with noise.

These results tend to confirm our assumption of over-fitting since, excepted for results related to the BootYTdyn technique, the others are linked to MSE which do not match with the theoretical random noise variances, reflecting a noise modeling phenomenon. Concerning the two SPLS methods, there is no relevant difference to mention, both of them concluding in similar numbers of selected predictors. Finally, similarly to BIC-based approaches, the Lasso regression extracts an increasing number of significant predictors.

As a first conclusion, we can reasonably conclude that, based on these first simulation

results, both the BootYTdyn and SPLS BootYT have to advised.

Complexity and predictive ability results.

To confirm and strengthen the conclusions we done in the previous Section 7.3.2, we will now focus on the predictive abilities of the models selected by the different approaches. We calculate a 100 times the 10-fold CV MSE based on the original simulated response values (without noise) of the different selected models. These results thus reflect the accuracy in predicting the original information by leaving out the random noise. We also compute the Predictive MSE (PMSE) based on the test set by using its simulated response $\mathbf{y}_{\text{test}}$ without including noise. Finally, we extract the DoF of each selected sparse models (DoF_{sel}) to compare their respective complexities.

Graphical representations of these statistics results related to the Q^2, BIC and BootYT techniques are performed in Fig.7.3.

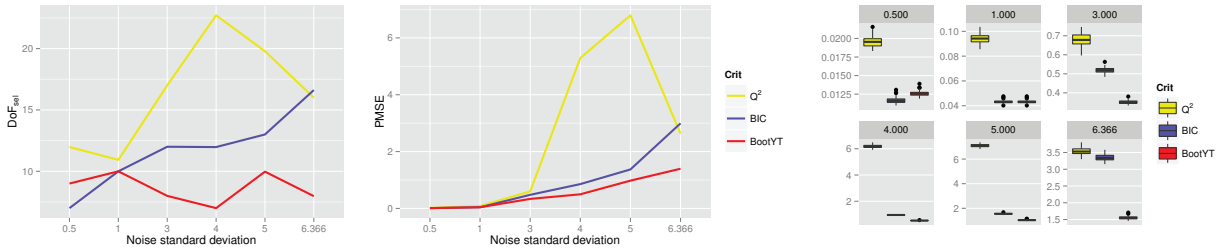


Figure 7.3: From Left to Right: DoF of the extracted sparse models, PMSE based on $\mathbf{y}_{\text{test}}$ without noise and boxplot of 10-CV MSE based on $\mathbf{y}$ without noise.

The evolution of estimated DoF highlights and confirms both BIC and Q^2 over-fitting issues. Indeed, while the random noise standard deviation is increasing, these two methods globally build sparse models with an increasing complexity. Thus, they model an increasing part of the inserted random noise, implying poor predictive abilities of their selected models compared to those obtained by applying the BootYT criterion. This is confirmed through higher values of PMSE and CV MSE, especially for datasets with non-negligible random noise variability. These results confirm the conclusions done in the previous section in that using the Q^2 criterion for selecting the number of PLS components has to be avoided and that the BootYT criterion outperforms the corrected BIC excepted for responses with negligible random noise levels. Therefore, only the BIC and BootYT are retained for further comparisons.

Results mapped out in Fig.7.4 highlight that the BootYTdyn approach is the only one modeling with decreasing DoF, ensuring a reduction of the complexity suitable to avoid predictive issues. Thus, the BootYTdyn select models with the lowest PMSE and 10-CV MSE.

In the light of these last results, only the BootYTyn approach is retained for further comparisons. Comparing both the two SPLS implementations using their predictive abilities lead to advise the SPLS BootYT one since models selected through this bootstrap-adapted

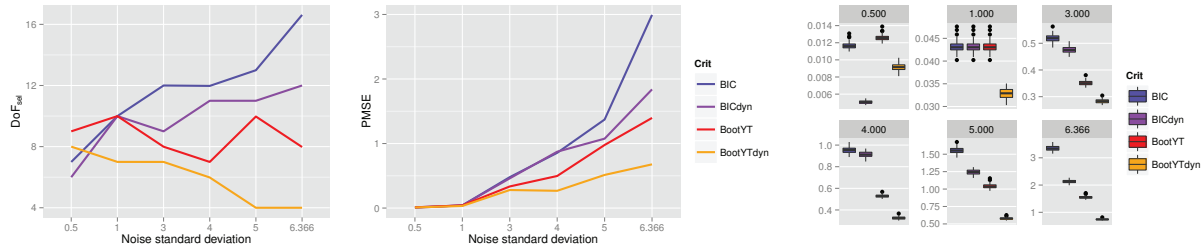


Figure 7.4: From Left to Right: DoF of the extracted sparse models, PMSE based on $\mathbf{y}_{\text{test}}$ without noise and boxplot of 10-CV MSE based on $\mathbf{y}$ without noise.

SPLS technique feature comparable if not lower both PMSE and 10-CV MSE (Fig.7.5). Let us precise that, in order to ensure a relevant comparison, we used ordinary PLS regressions with both the support and the number of components selected by the SPLS methods, and not the SPLS methods with selected tuning parameters, for computing the 10-CV MSE.

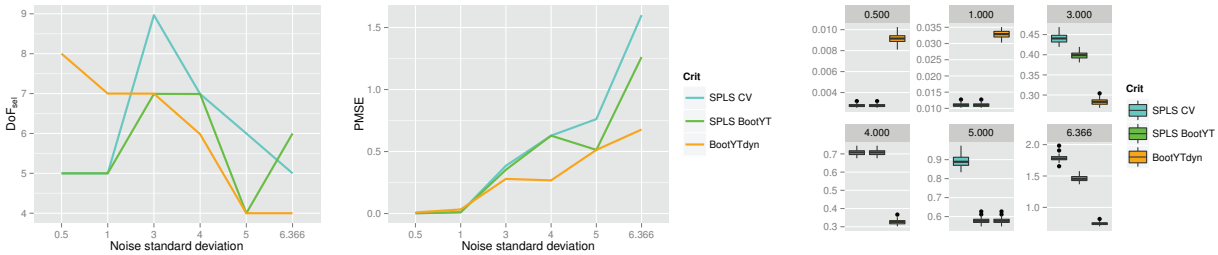


Figure 7.5: From Left to Right: DoF of the extracted sparse models, PMSE based on $\mathbf{y}_{\text{test}}$ without noise and boxplot of 10-CV MSE based on $\mathbf{y}$ without noise.

For datasets characterized by a low level of random noise variability in $\mathbf{y}$, the SPLS BootYT builds models inducing the smallest CV MSE and PMSE. By focusing on non-negligible random variability, the BootYTdyn features the smallest predictive errors traducing thus the high level of robustness of this approach against the random noise. This robustness is, as previously mentioned, due to its decreasing number of significant predictors and also components, leading thus to decreasing DoF *i.e.* a loss of complexity.

Lastly, comparison between the two retained methods, namely the BootYTdyn and the SPLS BootYT, and the Lasso is performed. As for SPLS methods, in order to perform a relevant comparison, the supports extracted by the Lasso regression are used as sets of covariates for a PLS regression. The number of PLS components was then established by performing a hundred times its determination using the bootstrap-based stopping criterion, the number of components related the highest occurrence rate was selected. In such a way, results reported in Fig.7.6 concerning the 10-CV MSE express efficiently the predictive ability of the extracted supports for a PLS regression. In order to provide a clear picture of the impact of this choice, the PMSE obtained through the Lasso regression is also reported in Fig.7.6. Both these approaches are referred to as Lasso and Lasso.supp.

Except for negligible values of random noise variability in $\mathbf{y}$, the support extracted from

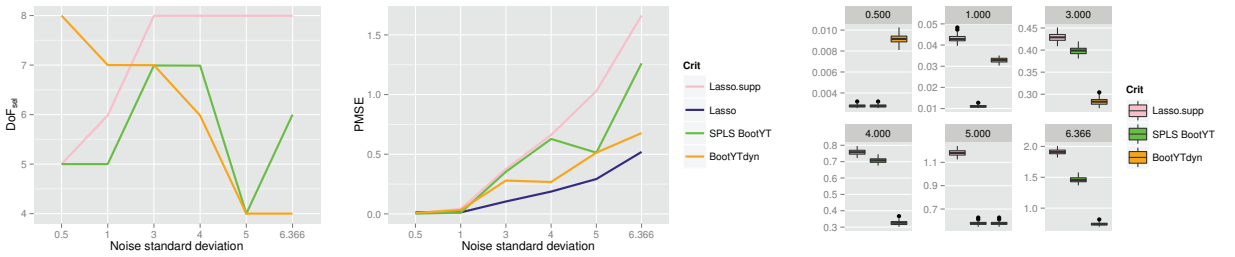


Figure 7.6: From Left to Right: DoF of the extracted sparse models, PMSE based on $\mathbf{y}_{\text{test}}$ without noise and boxplot of 10-CV MSE based on $\mathbf{y}$ without noise.

the Lasso regression and applied as explanatory variables for a PLS regression are related to both the highest 10-CV MSE and the highest PMSE. This is a direct consequence of the Lasso accuracy issues we mentioned in Section 7.3.2, notably the increasing number of extracted covariate while the random noise variability rises. However, performing the predictions using the model obtained by the original Lasso regression lead to the lowest PMSE values. This is due to the L^1 penalization which is applied on the vector of estimated parameters, permitting to correct the relative lack of accuracy of this technique.

Lastly, we summarize our conclusions in the following Table 7.8, by advising approaches depending if the initial aim was to select significant predictors or to obtain a sparse model with attractive predictive ability.

Table 7.8: Recommended approaches.

	Accuracy	Predictive ability
Low noise variability	SPLS BootYT	SPLS BootYT
High noise variability	BootYTdyn	Lasso\BootYTdyn

7.4 Real dataset application

In this part, we deal with the predictors matrix introduced in Section 7.3.2 and the original binary response vector. Thus, five approaches for variable selection adapted for GPLS framework are considered for comparison.

1. **BootYT**. The bootstrap-based method, introduced by Lazraq *et al.* (2003), combined with the bootstrap-based criterion (Magnanensi *et al.*, 2015) for previously selecting the number of components.

2. **BootYTdyn**. The new dynamic bootstrap-based method combined with the bootstrap-based criterion for successive determinations of numbers of components.

3. **SGPLS CV**. The original SGPLS method using 10-fold CV for tuning parameters determination (Chung and Keles, 2010).

4. **SGPLS BootYT**. The new adapted SGPLS version using the bootstrap-based criterion for components selection.

5. **RSGPLS**. An approach developed by Durif *et al.* (2015). It consists in adapting the SGPLS by introducing a Ridge penalty to ensure the convergence of parameter estimations and the stability concerning the hyper-parameter tuning. They also propose an adjustment of the L_1 constraint in order to further penalize the less significant predictors. Hyper-parameters are tuned through CV MSE.

6. **Lasso**. The adapted Lasso regression for logistic framework, available in the R package *glmnet*, as a benchmark.

Concerning bootstrap-based approaches, the incorporation of the PLS methodology into GLM developed by Bastien *et al.* (2005) was used. Due to non-convergence issues for parameters estimations on some bootstrap samples, some of these latter were excluded using a threshold for parameters estimations. Indeed, a model established on a bootstrap sample having at least one parameter estimation that is higher in absolute term than 10^4 times the one, in absolute value as well, estimated on the original dataset lead to the exclusion of the bootstrap sample. Thus, to ensure a sufficient number of relevant bootstrap samples, the preset number of computed ones was increased to R=4000. Concerning the Lasso regression, three different loss functions for the establishment of the sparsity parameter using 10-fold CV were used: the number of misclassified values, MSE and deviance values. We respectively refer to them as Lasso.Cl, Lasso.MSE and Lasso.Dev. The set of compared values for the sparsity parameter is preset by the "glmnet" package. For the SGPLS and RSGPLS approaches, the number of components K varies from 1 to 10, the sparsity parameter η varies from 0.04 to 0.99 by 0.05 and the ridge parameter involved in the RSGPLS technique has to be selected between 31 points that are $\log_{10}$-linearly spaced in the range $[10^{-2}; 10^{3}]$, as proceed by Durif *et al.* (2015).

Each method was performed one hundred times in order to obtain relevant results. However, due to the high observed stability of results extracted with the BootYTdyn approach and in order to save computational time, this technique was only performed twenty times instead of one hundred.

The different results of stability concerning the tuning parameters determinations, except the bootstrap-based methods ones, are reported in Fig.7.7 and Table 7.9.

Table 7.9: Number of different selected sets tuning parameters over the 100 determinations (Occurrence rate of the retained set of tuning parameters).

SGPLS CV	SGPLS BootYT	RSGPLS	Lasso.Cl	Lasso.MSE	Lasso.Dev
44 (9)	26 (16)	64 (5)	33 (8)	6 (44)	5 (48)

Based on results sum up in Table 7.9, the Lasso methodology using CV-based MSE or deviance values for the determination of its hyper-parameter is the most stable one. This could be explained by the fact that only one single hyper-parameter is involved while the others techniques are based on two or three tuning parameters. Using the number of misclassified

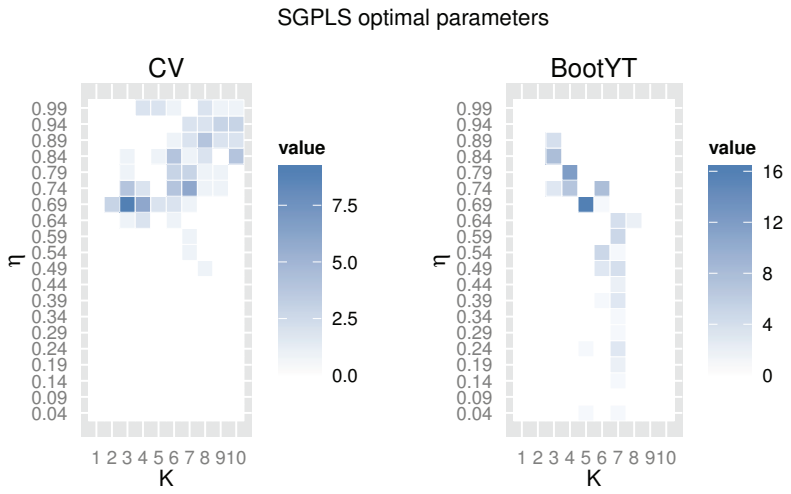


Figure 7.7: Repartition of selected sets of tuning parameters over the 100 determinations for both the SGPLS CV (Left) and the SGPLS BootYT (Right) approaches.

values for CV has to be avoided for a stable determination of the sparsity parameter. Our SGPLS adaptation with the BootYT criterion improves the reliability in selecting the set of tuning parameters, as previously observed in PLS frameworks (Sections 7.2.2 and 7.3.2). Note that concerning the RSGPLS, three different sets of optimal parameters were extracted with maximal occurrence rate equal to five, all of them selecting the same set of selected predictors. Thus, the set of parameters which leads to the smallest number of misclassified values on the training dataset was retained. As already mentioned in Section 7.3.2, extracting the same support does not necessarily lead to the same model when sparsity or ridge parameters are involved. Therefore, the numbers of both different sparse supports and different models retained are respectively summarized in Tables 7.10 and 7.11.

Table 7.10: Number Γ_1 of different extracted supports ($\%\mathcal{S}_{sel}$).

SGPLS CV	SGPLS BootYT	RSGPLS	Lasso.Cl	Lasso.MSE	Lasso.Dev	BootYT	BootYTdyn
44 (9)	14 (16)	26 (40)	26 (13)	5 (83)	4 (83)	4 (35)	2 (80)

Table 7.11: Number Γ_2 of different extracted sparse models ($\%\mathcal{M}_{sel}$).

SGPLS CV	SGPLS BootYT	RSGPLS	Lasso.Cl	Lasso.MSE	Lasso.Dev	BootYT	BootYTdyn
44 (9)	16 (16)	60 (5)	33 (8)	6 (44)	5 (48)	4 (35)	2 (80)

In the light of these results, the BootYTdyn and both the MSE- and deviance based Lasso techniques are the most stable ones in extracting supports and models. This could be notably due to the fact than all of them depend on one single tuning parameter. Even if this hyper-parameter for the BootYTdyn approach, *i.e.* the number of components, has to be determined R times, the high stability of the bootstrap-based stopping criterion introduced by Magnanensi *et al.* (2015) allows this approach to feature an attractive stability in selecting the sparse support. As for the PLS framework, our new bootstrap-based SGPLS implementation

improves significantly the stability of this approach. The lack of stability of the Lasso based on misclassified values is directly induced by the discrete form of this discriminating statistic. This issue was also observed and mentioned by Magnanensi *et al.* (2015).

Then, the different selected supports are displayed in Fig.7.8. Note that both the MSE- and deviance-based Lasso regressions select the same support and the same model leading to the Lasso.MSE.Dev notation in the following.

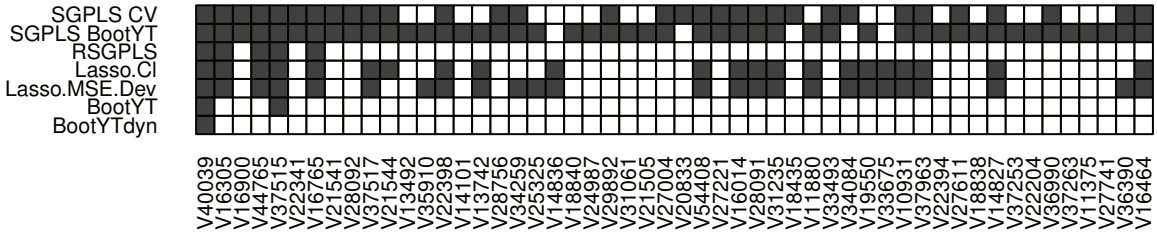


Figure 7.8: Summary of selected supports.

In the following, two additional independent public datasets mentioned by Marisa *et al.* (2013) as being comparable with our original dataset were included for the comparison of predictive abilities. The latter are named GSE18088 (n=53) (Gröne *et al.*, 2011) and GSE14333 (n=243) (Jorissen *et al.*, 2009). Both the MSE and numbers of misclassified (MC) value of the selected models were computed on both the training and test parts of the original dataset as well as on the two additional datasets. These results are summarized in Table 7.12. Indeed, since independent datasets are available, we decided to follow the indication of Van Wieringen *et al.* (2009, p.1596) in that "the true evaluation of a predictorS̆s performance is to be done on independent data".

Table 7.12: Summary of models fitting and predictive abilities.

	MC				MSE			
GSE	39582trai	39582test	18088	14333	39582trai	39582test	18088	14333
SGPLS CV	48	19	10	45	0.1372	0.1374	0.1743	0.1520
SGPLS BootYT	40	15	12	46	0.1295	0.1328	0.1848	0.1601
RSGPLS	53	13	17	45	0.1005	0.0912	0.1909	0.1276
Lasso.Cl	50	11	12	44	0.0928	0.0827	0.1283	0.1425
Lasso.MSE.Dev	46	15	12	47	0.0875	0.0828	0.1242	0.1406
BootYT	73	23	13	111	0.1272	0.1256	0.1817	0.2924
BootYTdyn	92	25	13	35	0.1507	0.1446	0.1695	0.1325

In this real dataset study, the BootYTdyn retains one single predictor, which is also retained by all the other methods. Thus, as expected, this sparsest support induced the highest values of both MSE and numbers of misclassified observations based on the training subset of the original dataset. It also provided the highest results based on the test subset of the original dataset, which is to be expected since, as explained in Section 7.3.2, these two parts are well balanced for the main anatomoclinical characteristics. Thus, this particularity inserts a bias in the evaluation of models predictive abilities by making comparable MSE and misclassified values based on both the training and testing subsets. Results in Table 7.12

confirmed this property and also strengthen the usefulness of independent additional datasets for reliably comparing the predictive abilities. While a well designed model for predictive purpose will provide similar PMSE values on comparable independent additional datasets than MSE obtained on the training dataset, an over-fitted one would be related to higher PMSE values due to its dependance on training data. Thus, the differences, noted Δ MSE, between the MSE obtained on the training subset of the original dataset and those obtained on the three test datasets are displayed in Fig.7.9.

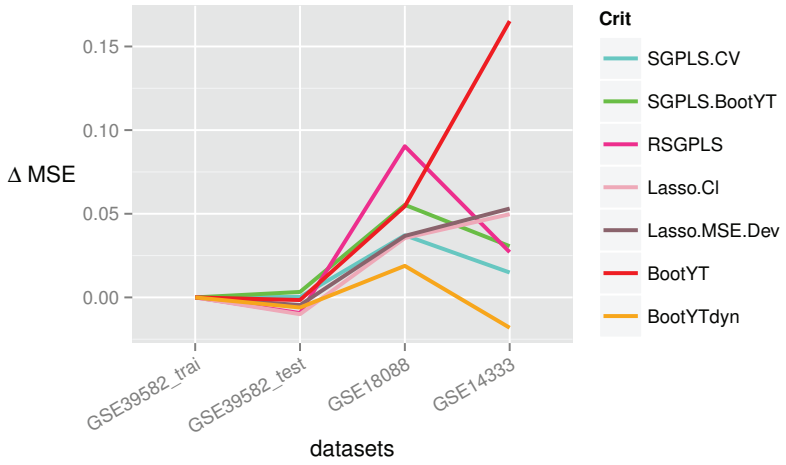


Figure 7.9: Differences between the MSE computed on the training subset of the original dataset and PMSE obtained on test sets.

Our new dynamic bootstrap-based approach is the only one exhibiting this property of MSE stability while all the other methods provided higher PMSE values on the two additional datasets than on the original one. This attractive feature lead the BootYTdyn to conclude in only 48 (16%) misclassified values on these two additional datasets, which represent the lowest result between all the studied approaches. Thus, we can reasonably assume that our new dynamic method has permitted to remove all the useless predictors in the sense that they are not relevant for improving the predictive ability.

Lastly, the single extracted probe set, named 230784_at, is already known to be specific to the original localisation in the distal colon. The sign of the regression coefficient obtained with the BootYTdyn method is coherent with this state-of-the-art result and thus strengthens our conclusion.

7.5 Discussion

In this article, we developed a new dynamic bootstrap-based technique for variable selection suitable for both PLS and GPLS frameworks and proposed a bootstrap-based version of the SPLS and SGPLS method for selecting the number of components. While the first one permit to completely avoid the use of CV, the second one permits to select the set of tuning parameters in a more reliable way.

In the state-of-the-art approaches, the use of CV-based techniques for the determination of involved hyper-parameters is current and could lead to important stability issues, observed notably by Boulesteix (2014) and Magnanensi *et al.* (2015) and confirmed on our studies. Sun *et al.* (2013) also worked on this subject and proposed a methodology for selecting the tuning parameters of penalized regressions in order to stabilize the variable selection. Even if their method is not applicable for the determination of the number of components for a PLS regression, it would be now interesting to adapt it to both the SPLS CV and our new SPLS BootYT implementation in order to observe a potential gain in stability. However, our new dynamic bootstrap-based technique represents a suitable method for avoiding CV-related issues since the single hyper-parameter is successively determined by a bootstrap-based criterion. This new technique improves an original bootstap-based methodology introduced by Lazraq *et al.* (2003) in that it permits to approximate the distribution of covariates' regression coefficients by removing the condition of a working subspace of fixed dimension K. Theoretical results have been established that strengthen the usefulness of building subspaces spanned by a dynamic number of components for performing PLS regressions on the bootstrap samples.

In PLS framework, the conclusion based on our simulations is twofold. First, for datasets with negligible random noise in $\mathbf{y}$ the SPLS BootYT has to be advised. Indeed, both in terms of accuracy and predictive abilities it outperforms all the other compared techniques. Second, for datasets with non-negligible random noise in $\mathbf{y}$, which represents a more realistic case, our new dynamic bootstrap-based has to advised. As for the SPLS BootYT for negligible random noise variability frameworks, the BootYTdyn outperform all the others for each studied characteristics. Furthermore, it is the only one which concludes in a decreasing number of significant predictors while the random noise variability is rising, that represents an expected behavior.

Results obtained on our classification study using real datasets match with the previous conclusions. Indeed, the BootYTdyn is the only one which lead to expected PMSE values on two additional independent datasets, that let us suppose that over-fitting issues were avoided and that these PMSE are induced by noise or information which can not be modeled by using only gene expressions. Furthermore, the extracted probe set is already known to be linked to the localization in the distal colon that strengthen the reliability of this new dynamic approach.

Lastly, our new bootstrap-based SPLS implementation improves the stability of this methodology. Indeed, in all studied cases, both the SPLS CV and SGPLS CV conclude in higher number of different sets of hyper-parameters than our bootstrap-based versions do, leading to higher numbers of different supports and models as well.

Simulations have to be done to ensure the results obtained on real datasets for logistic framework (Section 7.4). Testing the performances of these new approaches for responses following other distributions has also to be done. However, based on all the results obtained in these studies, our new dynamic method turns out to be the most efficient one compared to state-of-the-art approaches for datasets with non-negligible noise variability, a situation which is frequent in daily practice.

Quatrième partie

Perspectives et conclusion

Chapitre 8

Vers des améliorations des techniques développées

Les méthodes que nous avons développées et testées durant ce travail de thèse se sont avérées intéressantes à plusieurs points de vue. Cependant, des développements futurs sont encore à effectuer afin de généraliser ces procédés et les rendre utilisables en routine sur les différents cadres d'application possible. Nous allons ici évoquer quelques pistes d'amélioration qui seront effectuées dans un avenir proche.

8.1 Gérer la non-convergence des estimations

Dans le cadre des modèles de régressions linéaires généralisées, le problème de convergence des algorithmes d'estimation est bien connu. Ainsi les questions de l'existence et de l'unicité des estimations issues de la technique du maximum de vraisemblance, notamment les estimations liée aux modèles linéaires généralisés qui ont été introduit par Nelder and Wedderburn (1972), ont fait l'objet d'étude dès les années 1970 avec, entre autres, les articles de Anderson (1972), Anderson (1974) ou encore de Wedderburn (1976). Le problème de non-convergence des estimations des paramètres de régression se retrouve essentiellement dans le cadre de modèles liés à une réponse à distribution discrète, notamment dans les cas de classification multi-groupes et plus particulièrement dans le cadre de la régression logistique. Anderson (1972) introduit ainsi la notion de données *complètement séparées*, propriété problématique pour l'estimation par maximum de vraisemblance. Cependant, il faudra attendre l'article de Silvapulle (1981) pour obtenir un premier théorème explicitant des conditions suffisantes et nécessaires à l'existence et la convergence des estimations des paramètres de régression logistique. La condition première de ce théorème repose sur la nécessité de la non-séparation des données. Albert and Anderson (1984) finiront par définir proprement trois structures de données dans le cadre de la régression logistique, chacune liée à un théorème sur l'existence et l'unicité des estimations. Rappelons succinctement leurs résultats fondamentaux.

Definition 8.1.1. *Soit G le nombre de groupes représentés dans $\mathbf{y}$ et $\beta = (\beta_1, \ldots, \beta_{G-1})^T$ avec $\beta_g^T = (\beta_{g,0}, \ldots, \beta_{g,p})$ les coefficients de la régression logistique liée au groupe g pour $g = 1, \ldots, G-1$ et $\beta_G = 0_p$. Posons $q = (p+1)(G-1)$ de telle sorte que β soit ainsi un*

vecteur $\mathbb{R}^q$. *Soit* $\mathbf{x}_{i.}$ *la* $i^{ème}$ *ligne de la matrice* $\mathbf{X}$ *et* E_g *l'ensemble des indices des lignes de* $\mathbf{X}$ *liées au groupe* g.

- *On dit qu'il y a séparation complète des données si il existe un vecteur* $\beta \in \mathbb{R}^q$ *tel que pour tout* $i \in E_g$ *et* $g, h = 1, \ldots, G$ *avec* $g \neq h$ *on a :*

$$(\beta_g - \beta_h)^T \mathbf{x}_{i.} > 0 \tag{8.1}$$

  En d'autres termes, il existe un vecteur β *qui alloue parfaitement chaque observation à son groupe.*

- *On dit qu'il y a séparation quasi-complète des données si il existe un vecteur* $\beta \in \mathbb{R}^q$ *tel que pour tout* $i \in E_g$ *et* $g, h = 1, \ldots, G$ *avec* $g \neq h$ *on a :*

$$(\beta_g - \beta_h)^T \mathbf{x}_{i.} \geqslant 0 \tag{8.2}$$

  avec égalité pour au moins un triplet (i, g, h).

- *S'il n'existe pas de vecteur dans* $\beta \in \mathbb{R}^q$ *permettant une séparation complète ou quasi-complète des données, on dit alors qu'il y a chevauchement dans le sens où pour tout vecteur de* $\beta \in \mathbb{R}^q$ *il existe alors un triplet* (i, g, h) *avec* $i \in E_g$ *et* $g, h = 1, \ldots, G$ *avec* $g \neq h$ *tel que :*

$$(\beta_g - \beta_h)^T \mathbf{x}_{i.} < 0 \tag{8.3}$$

Les théorèmes liés à ces définitions sont alors les suivants.

Théorème 8.1.1. *(Albert and Anderson, 1984)*
Soit $L(\mathbf{X}, \beta)$ *la vraisemblance du modèle considéré. Alors :*

1. *S'il y a séparation complète des données, l'estimation* $\widehat{\beta}$ *par maximum de vraisemblance n'existe pas et* $\max_{\beta \in \mathbb{R}^q} L(\mathbf{X}, \beta) = 1$.

2. *S'il y a séparation quasi-complète des données, l'estimation* $\widehat{\beta}$ *par maximum de vraisemblance n'existe pas et* $\max_{\beta \in \mathbb{R}^q} L(\mathbf{X}, \beta) < 1$.

3. *S'il y a chevauchement alors l'estimation* $\widehat{\beta}$ *par maximum de vraisemblance existe et est unique.*

Ce problème de séparation est un cas auquel nous avons été confronté lors de l'étude de certaines bases de données, notamment celles liées à des études d'allélotypage. En effet, la matrice des régresseurs étant binaire, l'étude d'une réponse elle-même binaire, qui plus est via l'utilisation de techniques bootstrap, conduit souvent à l'obtention d'échantillons bootstrap présentant ce type de structure problématique. Des solutions ont été développées afin de palier à ces types de problème. L'intégration d'une pénalité de type *Ridge* a été proposée par Le Cessie and Van Houwelingen (1992) afin d'améliorer ces estimations. Cette technique sera d'ailleurs adaptée à la PLS, donnant ainsi lieu à la *Ridge* PLS, et à la *Sparse* PLS par, respectivement, Fort and Lambert-Lacroix (2005) et Durif *et al.* (2015). Une autre méthode,

appelée procédure de Firth, a été proposée par Firth (1993). Cette méthode consiste à réduire le biais des estimations à travers une modification de la fonction de score. Cette méthode a notamment été adaptée au cadre de la PLS logistique par Ding and Gentleman (2005) et à la *Sparse* PLS par Chung and Keles (2010). Enfin, Moulton and Zeger (1991) proposent l'utilisation du *one-step* bootstrap afin de palier à ces problèmes d'estimation lors de l'utilisation du bootstrap dans le cadre de régressions linéaires généralisées, le *one-step* bootstrap consistant à n'effectuer qu'une seule étape de l'algorithme de Newton-Raphson lors de l'estimation des paramètres par maximum de vraisemblance.

Il sera donc essentiel, et intéressant à la fois, d'effectuer les adaptations de ces solutions à nos développements afin, tout d'abord, de limiter les problèmes de non-convergence des procédures d'estimation des paramètres de régression mais également de comparer les performances de ces différentes adaptations entre elles et de déterminer, éventuellement, une adaptation à préférer.

8.2 Adaptation au *wild* Bootstrap

Le bootstrap par paires, procédure que nous avons utilisée dans le cadre de cette thèse, possède l'avantage de pouvoir traiter les cas d'hétéroscédasticité et d'être facilement implémentable. Cependant, une seconde procédure, généralisant le bootstrap des résidus en présence d'hétéroscédasticité, a également été proposée par Liu *et al.* (1988) suivant les premières indications énoncées par Wu (1986).

Le principe est le suivant. A partir du vecteur des résidus $\widehat{\epsilon} = (\widehat{\epsilon}_1, \ldots, \widehat{\epsilon}_n)$ obtenu suite à la régression de $\mathbf{y}$ sur $\mathbf{X}$, on va chercher à créer un nouveau vecteur $u = (u_1, \ldots, u_n)$ qui va servir à l'obtention d'une nouvelle réponse. Au lieu de tirer aléatoirement avec remise dans $\widehat{\epsilon}$ comme dans le cadre du bootstrap résiduel classique pour obtenir un nouvel échantillon (3.3.2), le vecteur u va être créé de la façon suivante :

$$\forall i \in [\![1, n]\!], \; u_i = f_i\left(\widehat{\epsilon}_i\right) \upsilon_i \tag{8.4}$$

avec f_i une transformation des résidus fixé par l'utilisateur, par exemple consistant en les diviser par $1 - h_i$ où h_i est le $i^{\text{ème}}$ élément de la diagonale de la matrice de projection, et les υ_i obtenus comme réalisations indépendantes d'une variable aléatoire tel que :

$$\mathbb{E}\left(\upsilon_i\right) = 0 \quad \text{et} \quad \text{Var}\left(\upsilon_i\right) = \mathbb{E}\left(\upsilon_i^2\right) = 1 \tag{8.5}$$

Un des choix de distribution particulièrement répandu (Davidson and Flachaire, 2008), proposé par Mammen (1993), est le suivant :

$$\mathcal{F}_1 : \quad \upsilon_i = \begin{cases} -\left(\sqrt{5}-1\right)/2, & \text{avec probabilité } \pi = \left(\sqrt{5}+1\right)/2\sqrt{5} \\ \left(\sqrt{5}+1\right)/2, & \text{avec probabilité } 1-\pi \end{cases} \tag{8.6}$$

Liu *et al.* (1988) propose également la distribution suivante :

$$\mathcal{F}_2 : \quad \upsilon_i = \begin{cases} 1, & \text{avec probabilité } 1/2 \\ -1, & \text{avec probabilité } 1/2 \end{cases} \tag{8.7}$$

Une fois le vecteur u obtenu, il reste alors à obtenir un nouvel échantillon bootstrap de la forme $(\mathbf{y}_w, \mathbf{X})$ avec :

$$\mathbf{y}_w = \widehat{\beta}\mathbf{X} + u \tag{8.8}$$

Nous remarquons bien ici que, comme dans le cadre du bootstrap résiduel classique, la matrice $\mathbf{X}$ est à nouveau considérée comme étant fixe. L'algorithme d'obtention d'échantillons bootstrap dans le cadre du *wild* bootstrap est donc le suivant :

1. Effectuer la régression de $\mathbf{y}$ sur $\mathbf{X}$, obtenant ainsi l'équation suivante :

$$\mathbf{y} = \mathbf{X}\widehat{\beta} + \widehat{\epsilon} \tag{8.9}$$

   avec $\widehat{\epsilon}$ le vecteur des résidus associé.

2. Pour $b = 1, \ldots, B$:

   - Déterminer un vecteur réponse $\mathbf{y}_w^{(b)}$ de la façon suivante :

$$\mathbf{y}_w^{(b)} = \mathbf{X}\widehat{\beta} + u^{(b)} \tag{8.10}$$

     avec $u^{(b)}$ obtenu suivant 8.4.

   - Effectuer la régression $\mathbf{y}_w^{(b)}$ sur $\mathbf{X}$:

$$\mathbf{y}_w^{(b)} = \mathbf{X}\widehat{\beta}^{(b)} + \widehat{\epsilon}_w^{(b)} \tag{8.11}$$

3. Récupérer l'ensemble $\left\{\widehat{\beta}^{(1)}, \ldots, \widehat{\beta}^{(B)}\right\}$.

Cette procédure sera notamment étudiée et comparée par Mammen (1993) et Davidson and Flachaire (2008). On y trouve ainsi des exemples de cas où elle se révèle être plus performante que le bootstrap par paires. Il peut donc être intéressant de comparer nos résultats obtenus durant ce travail de thèse avec des résultats issus de l'adaptation du *wild* bootstrap aux différentes méthodologies que nous avons développées.

8.3 Amélioration des codes utilisés

Le bootstrap est une technique basée sur l'*intensive computing*, autrement dit, il s'agit d'un procédé nécessitant une quantité importante de calcul informatique, les opérations de base pouvant être répétées des milliers de fois. Par exemple, notre développement de procédure dynamique pour la détermination de variable significative implique d'effectuer un grand nombre de régressions (généralisées) PLS. En effet, en supposant que l'on effectue 1000 itérations bootstrap concernant le bootstrap par paires $(\mathbf{y}, \mathbf{X})$ ainsi que 1000 itérations bootstrap pour tester le nombre de composantes significatives pour chacun des 1000 échantillons bootstrap créés précédemment et en supposant une moyenne de 5 composantes significatives, cela revient à effectuer approximativement 6×10^6 régressions (généralisées) PLS, pour ne parler que de cette opération.

Ainsi, une optimisation rigoureuse du code est essentielle afin de pouvoir traiter de grandes bases de données en des temps convenables. Pour ce faire, il nous faut développer des fonctions optimisées, n'effectuant et ne retournant que les éléments dont la procédure a besoin. Ce n'est actuellement pas le cas puisque les codes développés durant ce travail de thèse reposent en partie sur des fonctions existantes dans différent packages R, notamment le package *R plsRglm* (Bertrand *et al.*, 2014). Ces fonctions ne sont naturellement pas optimisées pour nos développements et effectuent ainsi souvent une quantité non-négligeable de calculs et d'opérations non nécessaires pour ces nouvelles procédures.

Un deuxième axe de travail, qui a déjà été débuté, consiste naturellement en l'élaboration de codes effectuant les calculs nécessaires en parallèle. En effet, de part sa nature, le bootstrap implique d'effectuer, indépendamment, de nombreuses fois les mêmes calculs. Il est donc parfaitement adaptable à du calcul en parallèle, accélérant considérablement les temps de calculs nécessaires à nos nouveaux procédés développés lors de cette thèse. Afin de fournir une idée du temps de calcul nécessaire lié ce nouveau procédé dynamique de détermination de variables significatives, nous retournons dans le tableau 8.1 des exemples de ces temps de calcul que nous avons obtenus lors des études menées et détaillées dans le chapitre 7. Précisons que ces temps représentent l'ensemble de l'opération, à savoir non seulement le temps de calcul de notre nouveau procédé mais également le temps de lecture de la base de données originale, la sélection des 100 prédicteurs retenue et le chargement des packages nécessaires. Il s'agit donc des temps liés aux différents codes tels qu'ils sont actuellement développés. Notons $R_{(\mathbf{y},\mathbf{X})}$ et $R_{(\mathbf{y},\mathbf{T})}$ les nombres d'itérations bootstrap effectuées concernant respectivement la phase de bootstrap sur les paires $(\mathbf{y},\mathbf{X})$ pour la détermination des prédicteurs significatifs et $(\mathbf{y},\mathbf{T})$ pour la détermination du nombre de composantes. Notons respectivement $\overline{K}_{\text{boot}}$ et $\widehat{\text{Var}}\left(K_{\text{boot}}\right)$ le nombre moyen de composantes extraites et la variance estimée de l'échantillon K_{boot}. Les processeurs utilisés pour ces calculs constituent un nœud de 24 cœurs montés sur deux sockets de type Intel Xeon E5-2680 v3 2.50GHz. Plus de détails sur ces aspects techniques sont disponibles en ligne à l'adresse suivante, http ://www.cemosis.fr/platform/atlas/.

TABLE 8.1 – Exemples de temps de calcul, effectués sur 24 cœurs en parallèle

Type	$R_{(\mathbf{y},\mathbf{X})}$	$R_{(\mathbf{y},\mathbf{T})}$	$\overline{K}_{\text{boot}}$	$\widehat{\text{Var}}\left(K_{\text{boot}}\right)$	Temps
PLS	1000	1000	15.024	4.161	42min 02s
PLS-Logistique	4000	1000	11.539	4.390	8h 29min 04s

Chapitre 9

Conclusion

Ce travail de thèse, travail qui aura duré trois ans, avait pour objectif d'améliorer l'utilisation de la régression PLS dans le cadre biologique. En effet, l'application de la régression PLS à des fins de traitement statistique d'études géniques, est devenu un procédé de référence (Boulesteix and Strimmer, 2007). Cependant, concrètement, sa mise en oeuvre et son interprétation restent délicates pour des professionnels du monde médical et ce pour plusieurs raisons, raisons que nous avons abordé durant cette rédaction de thèse. Précisons tout de fois que ces difficultés ne sont pas spécifiquement liées aux mondes biologique et médical, mais de façon plus large, à tous les praticiens de cette méthode. L'objectif principal aura donc été de développer des méthodes plus stables et plus robustes, que ce soit pour la détermination d'un nombre optimal de composantes ou celle de prédicteurs significatifs. En effet, il était, jusqu'à présent, difficilement concevable pour un utilisateur de cette méthode d'obtenir des conclusions fortes et pertinentes lorsqu'en exécutant plusieurs fois les mêmes critères ou procédés sur une même base de données, ceux-ci retournent des résultats différents. Voilà le coeur du problème que nous avons essayé de traiter dans ce travail de thèse, l'obtention de procédés liés à une certaine stabilité des résultats.

Ces trois années auront permis, d'après nous, l'élaboration de travaux intéressants et concluants.

Intéressants, ces travaux l'auront été sur plusieurs points. Tout d'abord, ils nous ont permis de confronter les méthodes actuelles à la présence de bruit dans des bases de données et, ainsi, évaluer leur robustesse face à cette caractéristique omniprésente dans la réalité. Les résultats ont été surprenants mais concordent avec des observations présentes dans la littérature. Nous pensons tout particulièrement à la non-robustesse du critère du Q^2, l'entraînant rapidement à sous-estimer le nombre optimal de composantes, sous-estimation observée par Tenenhaus (1998) sur la base de données des « Processionnaires du Pin ». Deuxièmement, il a été intéressant de se confronter aux techniques basées sur la VC ainsi qu'à leurs applications sur des jeux de données. En effet, très souvent, ces résultats sont quasi-inexploitables tant la variabilité des conclusions issues de ces critères est forte. Nous avons pu observer cela notamment sur la question du choix du nombre de composantes (voir par exemple la Fig. 5.18) ou encore sur la détermination des hyperparamètres de la *Sparse* PLS (voir par exemple la Fig. 7.7). Enfin, il a été particulièrement enthousiasment d'observer les premières améliorations

apportées par l'adaptation de techniques bootstrap à ce procédé de régression PLS ainsi que les perspectives futures que cela apporte, perspectives que nous avons brièvement évoquées au chapitre précédent et qui peuvent être complétées notamment par l'adaptation de ces méthodes à la PLS multivariée (PLS2) par exemple.

Enfin, ces travaux ont été concluants puisque sur la majorité des propriétés observées et étudiées (robustesse au bruit, variabilité des résultats, universalité de la méthode, performances prédictives...), nos différents développements basés sur le bootstrap se sont avérés globalement, voir unanimement, liés à des améliorations significatives comparativement aux procédés existants. Ces résultats laissent ainsi entrevoir l'opportunité, nous l'espérons, à des praticiens non spécialistes de la méthode et des statistiques au sens large, de pouvoir procéder à ce type d'analyse avec une plus grande confiance dans la méthode et de meilleures interprétations des résultats obtenus.

D'un point de vue plus personnel, l'ensemble de ce travail de thèse, du début à cette fin, aura été une expérience fortement enrichissante pour moi et ce sur plusieurs aspects. Scientifiquement, la démarche et les méthodes de travail dans le domaine de la recherche universitaire m'étaient entièrement inconnues. L'adaptation et l'apprentissage de telles méthodes n'ont pas toujours été évidents mais je pense que maintenant acquises, elles m'apportent un vrai atout quant à ma capacité à effectuer des recherches scientifiques indépendamment du cadre de la thèse. Ce fait est également dû à un autre aspect de ce travail de thèse, à savoir le développement de ma capacité à travailler de façon autonome. Cette caractéristique est à mettre au crédit d'une confiance réciproque qui s'est installée entre les différents acteurs de cette thèse. De plus, gérer une thématique de recherche sur trois années est un réel défi à mes yeux tant il est possible de se perdre dans la quantité impressionnante de recherches, de développements, d'articles et de publications de tout type auxquels j'ai pu être confronté. Quantité qui fait également preuve de l'important dynamisme autour de questions telles que celle que nous avons tenté de traiter, et plus globalement du milieu de la régression PLS et de ses applications, biologiques et médicales notamment.

Humainement, il aura été extrêmement enrichissant de pouvoir intégrer une équipe essentiellement constituée de personnels biologistes et médicaux, d'être confronté à leurs problématiques ainsi qu'à leurs attentes du monde des mathématiques. De formation intégralement mathématicienne, j'ai pu me rendre compte à quel point, à travers l'utilisation de deux « dictionnaires »de vocabulaire complètement différents possédant chacun leurs termes propres, ces deux mondes pouvaient mutuellement s'apporter. Il s'agit certainement d'une des optiques les plus enthousiasmantes à mes yeux, tant les avancés nées de l'« alliance »de ces deux domaines me paraissent prometteuses pour l'avenir. Cela passera certainement par la multiplication de ce type de travaux, par l'organisation de rencontres entre ces deux univers, par la nécessité pour chacun de faire un effort de compréhension et d'ouverture sur les problématiques et contraintes de l'autre. Mais, si cela est fait dans le respect des spécialités et spécialistes respectifs, le résultat ne pourra être que d'une importance capitale, notamment concernant les avancés possibles dans la compréhension et le traitement de pathologies aussi complexes et dévastatrices que le cancer.

Cinquième partie

Annexes

Annexe A

Processus de simulation des jeux de données pour la détermination du nombre de composantes PLS

Table A.1: Algorithm of data simulation process for PLS and generalized PLS regression

n = number of subjects
p_l = number of predictors in the simulated dataset.
Let $\mathbf{T} = (t_{k\rho})_{\substack{1\leqslant k\leqslant 4 \\ 1\leqslant \rho\leqslant p_l}} \in \mathcal{M}_{4,p_l}(\mathbb{R})$ represents 4 fixed orthogonal components.
for $j = 0.01, \ldots, \eta$ **by** γ **do**
for $h = 0.01, \ldots, \xi$ **by** λ **do**
$\sigma_R = (\sigma_1, \ldots, \sigma_5) = (10, 8, 6, h, j)$
$\tau_G = (\tau_1, \ldots, \tau_5) = (0.250, 0.125, 0.050, 0.0125, 0.005)$
for $l = 1, \ldots, 100$ **do**
for $i = 1, \ldots, n$ **do**
Let r_k be a realisation of $R_k \sim \mathcal{N}(0, \sigma_k^2)$ for $k = 1, \ldots, 5$
Let g_k be a realisation of $G_k \sim \mathcal{N}(0, \tau_k^2)$ for $k = 1, \ldots, 5$
Let $\epsilon_{i\rho}$ values be p_l realisations of $\epsilon \sim \mathcal{N}(0, 10^{-4})$
Let δ_i be a realisation of $\delta \sim \mathcal{N}(0, 10^{-3})$

$$x_{i\rho} = \begin{cases} \sum_{k=1}^{4} r_k.t_{k\rho} + \epsilon_{i\rho},\ \rho = 1, \ldots, p_l, & \text{PLS and PLS-logistic cases} \\ \sum_{k=1}^{4} r_k.t_{k\rho} + \epsilon_{i\rho},\ \rho = 1, \ldots, p_l,\ -1 \leqslant \sum_{k=1}^{3} r_k \leqslant 5, & \text{PLS-Poisson case} \end{cases}$$

$$\theta_i = ((r_1 + g_1)/2) + ((r_2 + g_2)/2) + ((r_3 + g_3)/2) + ((r_5 + g_5)/2) + \delta_i$$

$$y_i = \begin{cases} \theta_i, \text{PLS case} \\ \mathcal{B}(inv.logit(\theta_i)), \text{PLS-logistic case} \\ \mathcal{P}(exp(\theta_i)), \text{PLS-Poisson case} \end{cases}$$

Apply the different criteria, depending on the case, on the dataset.
Extract the retained number of components for each of them.
end for
end for
end for
end for

Table A.2: Simulation parameters values.

Parameters	PLS		PLS-logistic		PLS-Poisson	
	$n > p$	$n < p$	$n > p$	$n < p$	$n > p$	$n < p$
η	20.01	20.01	15.51	9.51	7.01	5.01
γ	0.50	0.50	0.50	0.50	0.20; 0.50	0.20; 0.50
ξ	30.01	6.01	9.51	9.51	9.51	9.51
λ	0.20;1.00	1.00	0.50	0.50	0.50	0.50

Annexe B

Démonstrations des propositions du Chapitre 5

B.1 Démonstration Proposition 5.2.1

Proof. Let $\mathbf{y}_{(k-1)}$ and $\mathbf{X}_{(k-1)}$ be respectively the deflated response vector and the deflated predictors matrix on $k-1$ components and suppose that $\forall k \in [\![1, K]\!], \exists i \in [\![1, p]\!], \mathbf{x}_{i,(k-1)}^T \mathbf{y}_{(k-1)} \neq 0$, then:

$$\begin{aligned}
c_k &= \mathbf{y}_{(k-1)}^T \mathbf{t}_k / \mathbf{t}_k^T \mathbf{t}_k \\
&= \mathbf{y}_{(k-1)}^T \mathbf{X}_{(k-1)} \mathbf{w}_k / \left\| \mathbf{t}_k \right\|^2 \\
&= \mathbf{y}_{(k-1)}^T \mathbf{X}_{(k-1)} \mathbf{X}_{(k-1)}^T \mathbf{y}_{(k-1)} / \left\| \mathbf{X}_{(k-1)}^T \mathbf{y}_{(k-1)} \right\| . \left\| \mathbf{t}_k \right\|^2 \\
&= \left(\mathbf{X}_{(k-1)}^T \mathbf{y}_{(k-1)} \right)^T \mathbf{X}_{(k-1)}^T \mathbf{y}_{(k-1)} / \left\| \mathbf{X}_{(k-1)}^T \mathbf{y}_{(k-1)} \right\| . \left\| \mathbf{t}_k \right\|^2 \\
&= \left\| \mathbf{X}_{(k-1)}^T \mathbf{y}_{(k-1)} \right\|^2 / \left\| \mathbf{X}_{(k-1)}^T \mathbf{y}_{(k-1)} \right\| . \left\| \mathbf{t}_k \right\|^2 \\
&= \left\| \mathbf{X}_{(k-1)}^T \mathbf{y}_{(k-1)} \right\| / \left\| \mathbf{t}_k \right\|^2 \\
&> 0
\end{aligned}$$

Consequently and by induction, $\mathbb{P}\left(c_k \leqslant 0 \,|\mathbf{X}\right) = \mathbb{P}\left(c_k \leqslant 0 \,\middle|\mathbf{X}_{(k-1)}\right) = 0$.

□

B.2 Démonstration Proposition 5.6.1

Proof. Let $n = pq+r$ be the Euclidean division of n by q. Then, a partition of the dataset for CV is composed of r $(p+1)$-element subsets and $(q-r)$ p-element subsets, noted A_k, $1 \leqslant k \leqslant q$. We set $q_0 = 0$ and let $E = \{q_1, \ldots, q_r\}$, $1 \leqslant q_j \leqslant q$, $\forall j \in [\![1, r]\!]$ be a subset of $\{1, \ldots, q\}$ containing the ordered set of indices of the $(p+1)$-element subsets so that:

$$Card\,(A_k) = \begin{cases} p+1, & \forall k \in E \\ p, & \forall k \in E^c \end{cases}$$

Let us first determine the number of distinct partitions of $\{1, \ldots, n\}$ which can be written as:

$$\{A_1, \ldots, A_{q_1-1}, A_{q_1}, A_{q_1+1}, \ldots, A_{q_r-1}, A_{q_r}, A_{q_r+1}, \ldots, A_q\} \tag{B.1}$$

To lighten the formulas, let us set $\omega = \{q_j \in E \,|q_j - q_{j-1} \neq 1\}$ and $m_j = q_j - q_{j-1} - 1$. Then, by knowing that, for any set containing n elements, the number of distinct p-element subsets of it that can be formed is given by $\binom{n}{p}$, we obtain that the number of distinct partitions of

the form (B.1) is equal to:

$$
\begin{aligned}
f(n,q) = \prod_{j=1}^{r} \Bigg[\Bigg[\mathbb{1}_{\{q_j \in \omega\}} \prod_{i=1}^{m_j} \binom{n - (q_{j-1} - (j-1) + (i-1))\, p - (j-1)\,(p+1)}{p} \\
+ \mathbb{1}_{\{q_j \notin \omega\}} \Bigg] \times \binom{n - ((q_j - 1) - (j-1))\, p - (j-1)\,(p+1)}{p+1} \Bigg] \\
\times \Bigg[\mathbb{1}_{\{q_r \neq q\}} \prod_{i=1}^{q-q_r} \binom{n - (q_r - r + (i-1))\, p - r\,(p+1)}{p} + \mathbb{1}_{\{q_r = q\}} \Bigg] \\
= \prod_{j=1}^{r} \Bigg[\Bigg[\mathbb{1}_{\{q_j \in \omega\}} \prod_{i=1}^{m_j} \binom{n - (q_{j-1} + i - 1)\, p - (j-1)}{p} + \mathbb{1}_{\{q_j \notin \omega\}} \Bigg] \\
\times \binom{n - (q_j - 1)\, p - (j-1)}{p+1} \Bigg] \\
\times \Bigg[\mathbb{1}_{\{q_r \neq q\}} \prod_{i=1}^{q-q_r} \binom{n - (q_r + (i-1))\, p - r}{p} + \mathbb{1}_{\{q_r = q\}} \Bigg] \\
= \prod_{j=1}^{r} \Bigg[\Bigg[\mathbb{1}_{\{q_j \in \omega\}} \prod_{i=1}^{m_j} \frac{(n - (q_{j-1} + i - 1)\, p - (j-1))!}{p!\,(n - (q_{j-1} + (i+1) - 1)\, p - (j-1))!} + \mathbb{1}_{\{q_j \notin \omega\}} \Bigg] \\
\times \frac{(n - (q_j - 1)\, p - (j-1))!}{(p+1)!\,(n - ((q_j + 1) - 1)\, p - ((j+1) - 1))!} \Bigg] \\
\times \Bigg[\mathbb{1}_{\{q_r \neq q\}} \prod_{i=1}^{q-q_r} \frac{(n - (q_r + (i-1))\, p - r)!}{p!\,(n - (q_r + (i+1) - 1)\, p - r)!} + \mathbb{1}_{\{q_r = q\}} \Bigg]
\end{aligned}
$$

Each second factorial term in the denominator being equal to the numerator of the following product term, we obtain that:

$$
\begin{aligned}
f(n,q) = \prod_{j=1}^{r} & \left[\left[\mathbb{1}_{\{q_j \in \omega\}} \frac{(n - q_{j-1}p - (j-1))!}{(p!)^{m_j} \, (n - (q_j - 1)\, p - (j-1))!} + \mathbb{1}_{\{q_j \notin \omega\}} \right] \right. \\
& \left. \times \frac{(n - (q_j - 1)\, p - (j-1))!}{(p+1)!\, (n - ((q_j + 1) - 1)\, p - ((j+1) - 1))!} \right] \\
& \times \left[\mathbb{1}_{\{q_r \neq q\}} \frac{(n - q_r p - r)!}{(p!)^{q - q_r} \, (n - qp - r)!} + \mathbb{1}_{\{q_r = q\}} \right] \\
= \prod_{j=1}^{r} & \left[\mathbb{1}_{\{q_j \in \omega\}} \frac{(n - q_{j-1}p - (j-1))!}{(p!)^{m_j} \, (p+1)!\, (n - q_j p - j)!} \right. \\
& \left. + \mathbb{1}_{\{q_j \notin \omega\}} \frac{(n - (q_j - 1)\, p - (j-1))!}{(p+1)!\, (n - q_j p - j)!} \right] \\
& \times \left[\mathbb{1}_{\{q_r \neq q\}} \frac{(n - q_r p - r)!}{(p!)^{q - q_r}} + \mathbb{1}_{\{q_r = q\}} \right]
\end{aligned}
$$

However, by definition, if $q_j \notin \omega$ then $q_j - 1 = q_{j-1}$ and $m_j = 0$. Furthermore, if $q_r = q$ then $\frac{(n - q_r p - r)!}{(p!)^{q - q_r}} = 1$, so:

$$
\begin{aligned}
f(n,q) = \prod_{j=1}^{r} & \left[\mathbb{1}_{\{q_j \in \omega\}} \frac{(n - q_{j-1}p - (j-1))!}{(p!)^{m_j} \, (p+1)!\, (n - q_j p - j)!} \right. \\
& \left. + \mathbb{1}_{\{q_j \notin \omega\}} \frac{(n - q_{j-1}p - (j-1))!}{(p!)^{m_j} \, (p+1)!\, (n - q_j p - j)!} \right] \\
& \times \left[\mathbb{1}_{\{q_r \neq q\}} \frac{(n - q_r p - r)!}{(p!)^{q - q_r}} + \mathbb{1}_{\{q_r = q\}} \frac{(n - q_r p - r)!}{(p!)^{q - q_r}} \right] \\
= & \prod_{j=1}^{r} \left[\frac{(n - q_{j-1}p - (j-1))!}{(p!)^{m_j} \, (p+1)!\, (n - q_j p - j)!} \right] \times \frac{(n - q_r p - r)!}{(p!)^{q - q_r}} \\
= & \left(\frac{1}{p!} \right)^{\sum_{j=1}^{r} m_j + q - q_r} \times \left(\frac{1}{(p+1)!} \right)^{r} \times n! \\
= & \left(\frac{1}{p!} \right)^{qr} \times \left(\frac{1}{(p+1)!} \right)^{r} \times n!
\end{aligned}
$$

Finally, since the order of parts creation within each of the two classes of subsets ((p+1) and p-element subsets) is not useful in distinguishing one partition from another in a CV framework, we had to divide our result by the number of distinct permutations it exists in each class, so:

$$
f(n,q) = \frac{n!}{r!\, (q-r)!} \times \left(\frac{1}{(p+1)!} \right)^{r} \times \left(\frac{1}{p!} \right)^{q-r} \tag{B.2}
$$

This result is therefore not dependant on E. □

Annexe C

Résultats obtenus par l'AIC et le BIC dans le cadre des simulations PLS-LR de la Section 5.5.1

Results obtained with the non-corrected AIC and BIC criteria are displayed in Fig.C.1.

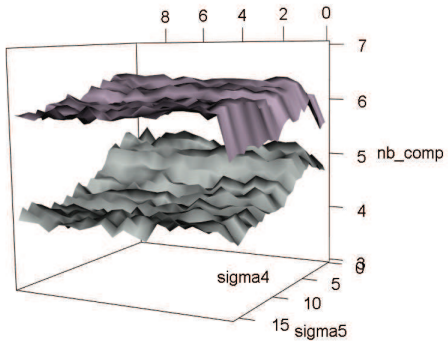


Figure C.1: PLS-LR, $n > p$, sets of NSD values (D); Evolution of means of selected numbers of components (nb_comp) over 100 datasets per couple (σ_4, σ_5), related to both the AIC (top) and BIC (bottom) criteria

Note that we only compared AIC and BIC values linked to the k-components models with $k \leqslant 7$ and retained the one which realize the minimum of the studied criterion.

Annexe D

Contrainte instaurée pour la simulation de données dans le cadre PLS-Poisson à la Section 5.5.2

Due to the specificities of the Poisson distribution and its inverse link function exp, we added the following constraint:

$$-1 \leqslant \sum_{k=1}^{3} r_k \leqslant 5,$$

where r_k is a realization of a random variable $R_k \sim \mathcal{N}(0, \sigma_k^2)$.

Indeed, the usual simulation process algorithmically performs realizations, noted θ_i, $1 \leqslant i \leqslant n$, of a random variable θ so that each response $y_i = \mathcal{P}(\exp(\theta_i))$. The distribution of θ is directly link to the sum of r_1, r_2 and r_3, among others, so that, without constraints on these parameters, this process leads to zero-inflated models combined with a heavy-tailed distribution. Let Z be a simplified approximation of θ. Adding this constraint leads to the following theoretical results. Let:

$$Z = 0.5\left(\sum_{k=1}^{3} R_k + R_5\right) \tag{D.1}$$

$$\Rightarrow \begin{cases} \mathbb{E}\left(Z | -1 \leqslant \sum_{k=1}^{3} R_k \leqslant 5\right) = 0.985 \\ \mathbb{V}\mathrm{ar}\left(Z | -1 \leqslant \sum_{k=1}^{3} R_k \leqslant 5\right) = 0.745 \\ \mathrm{sd}\left(\sum_{k=1}^{3} R_k | -1 \leqslant \sum_{k=1}^{3} R_k \leqslant 5\right) = 1.727 \end{cases} \tag{D.2}$$

We simulated two similar datasets, containing 1000 observations. These datasets were simulated respectively with and without constraint. As expected, without constraint no less than 490 out of the 1000 response values are equal to 0 and the maximal simulated one is equal to 111074934. By adding this constraint, the obtained response vector only contains 147 responses equal to 0 and 23 as a maximal value. The two types of distributions (with and without constraint) linked to θ and $\mathbf{y}$ are graphically displayed in Fig.D.1

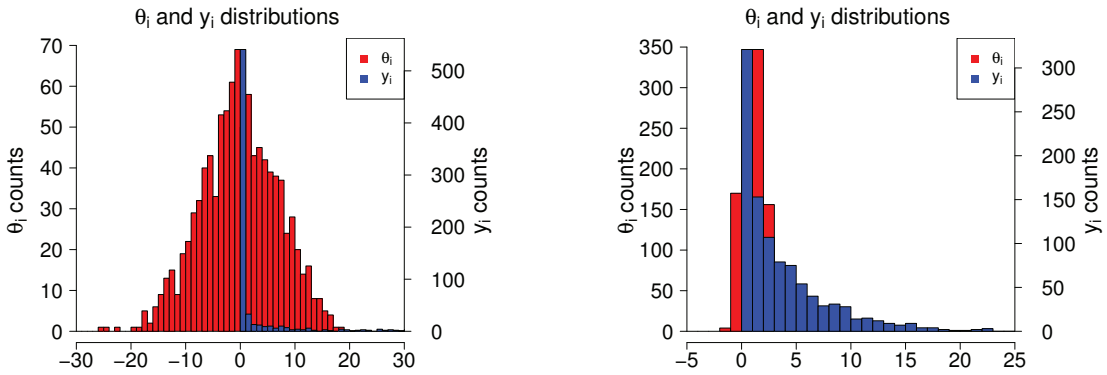


Figure D.1: Left: Distributions without constraint; Right: Distributions with constraint

Annexe E

Poster présenté à la conférence internationale PLS14

E.1 Résumé du poster publié dans le *Book of Abstract* de la conférence

A new bootstrap based stopping criterion in PLS components construction.

Frédéric Bertrand[b], Jérémy Magnanensi[a,b*], Myriam Maumy-Bertrand[b] and Nicolas Meyer[a]

[a]Faculté de Médecine - Laboratoire de Biostatistique - EA3430
[b]Institut de Recherche Mathematique Avancée - LabEx IRMIA - UMR7501
Université de Strasbourg, CNRS, Strasbourg, France

Keywords: Bootstrap, PLS, PLS-GLM, component

Introduction

The extraction of the optimal number of PLS components is a real challenge. Indeed, this step serves to find out the real dimension of the link between the response $\mathbf{y}$, which will be a $\mathbb{R}^{n\times 1}$ vector in our case, and the predictors matrix $\mathbf{X}$. Considering k as the optimal number of components, concluding in $k_1 < k$ components leads to a loss of information so that links between some predictors and $\mathbf{y}$ will not be correctly modelled. On the other hand, concluding in $k_2 > k$ components means that some "noise" or useless information in $\mathbf{X}$ will be used to explain $\mathbf{y}$, so that it leads to an over-complex model that could fit the data well but will have a poor predictive ability.
This complexity or sensibility of the model could be measured with the development of the degrees of freedom (DoF), which was done by N.Krämer and M.Sugiyama[1]. They also adapted the AIC and BIC criterion with these DoF and applied them to the selection of the optimal number of PLS components.
An extension to GLM of the PLS regression, noted *PLS-GLM*, has been developed by Bastien *et.al*[2]. In this case, fewer criteria can be used since DoF are not established yet and some criteria, like Q^2, are not suited.
Our aim was to find out an "universal" criterion, in that it could work either in PLS and PLS-GLM cases. Furthermore, we wanted to obtain some criterion that could be used as a classical test in that an error risk level α could be fixed. So, since bootstrap techniques are used to empirically model some statistics distributions, it also became possible to create confidence intervals in order to test regression coefficients[3].
In our case, we adapted the so-called bootstrapping pairs technique in order to test the significance of the successive PLS components and compare this criterion with the most used ones: the Q^2 obtained by leave-one-out cross-validation (CV) , the Q^2 obtained by 5-fold CV and the adapted AIC and BIC criterion. This choice of 5-fold CV is due to results on the better q to use in q-fold CV[4]. We also compare, in a logistic case, our bootstrap approach to two others criteria, the 5-fold CV missclassified value and a criterion developed by Bastien *et.al*[2], noted *p_val*, which define the non-significance of a new component $\mathbf{t_h}$ as not any significant predictors within it.
To compare all these criteria, whether in PLS or PLS-logistic case, we used some important data simulations for different levels of noise added in $\mathbf{X}$ and $\mathbf{y}$, noted respectively sigma4 and sigma5 on Fig.1, either when $n > p$ and $n < p$, with n the number of subjects and p the number of predictors. The simulation algorithm is available on: `http://www-irma.u-strasbg.fr/~magnanen/Algo_simul.pdf`. Our study show a better stability of our criterion and a globally better predictive accuracy compared to the others criteria, either in PLS or in PLS-logistic.

1 The adapted bootstrap method.

This method relies on the assumption that, once the components are build, they are considered as being independent of the response. Indeed, once this base is fixed, we want to test the significance of a new component $\mathbf{t_k}$, not by

*Corresponding author. E-mail: 7, Rue Rene Descartes 67084 Strasbourg Cedex, France. magnanensi@math.unistra.fr

simulating the real distribution of the coefficient linked to this component, which would be a positive one, but rather by approaching the conditional distribution of these coefficients given $\mathbf{T} = (\mathbf{t_1}, \ldots, \mathbf{t_k})$. First, let $j = 1$.

1. Compute the j firsts components $\mathbf{t_j}$.
2. Bootstrapping pair $(\mathbf{y}, \mathbf{T})$, returning R bootstrap samples noted $(\mathbf{y}, \mathbf{T})^{b_1}, \ldots, (\mathbf{y}, \mathbf{T})^{b_R}$.
3. For each couple $(\mathbf{y}, \mathbf{T})^{b_i}$, $i = 1, \ldots, R$, do the (generalized) linear regression $\mathbf{y} = \sum_{h=1}^{j} \left(\hat{c}_h^{b_i} . \mathbf{t_h} \right) + \hat{\varepsilon}_j^{b_i}$.
4. Since $c_j > 0$, construct an unilateral IC $= \left[IC_1^j, +\infty \right[$ of level $\alpha = 0.95$ for c_j with BCa technique.
5. While $IC_1^j > 0$, **do** j=j+1, and return to step 1. Else, the final extracted number of component is $k = j - 1$.

2 Results

Criteria predictive accuracies were compared. For that, results obtained in the $n < p$ case, where $n = 20$, were used, considering it as a training sample and 80 supplementary data were simulated as a test sample. Then, we measured their predictive performance by computing the normalized mean squares test error for each of the criterion and also by computing the number of predictive missclassified values in the PLS-logistic case. Finally, these means were compared with Student tests (see Fig.1).

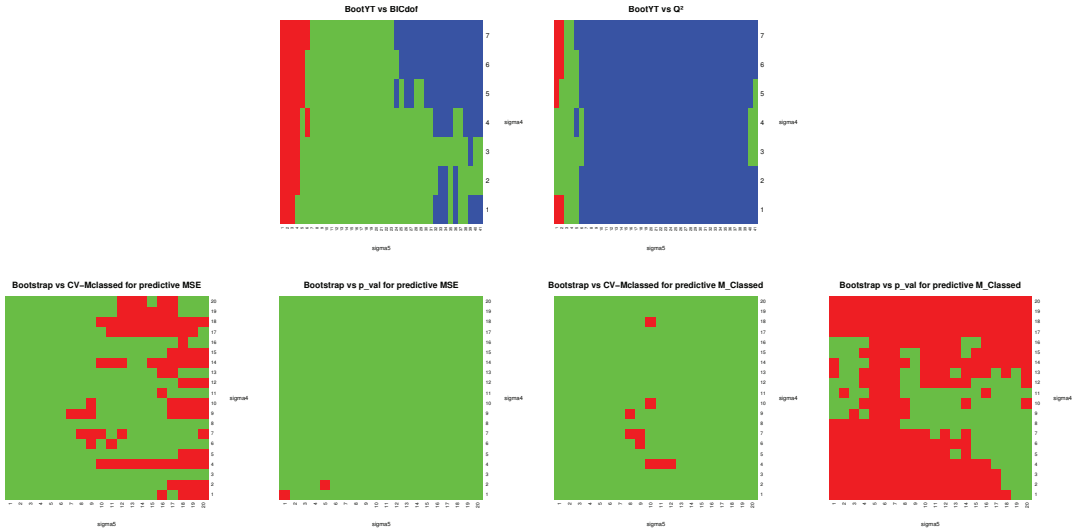


Figure 1. *Upper*: PLS case; Left: $BICdof$ better than $BootYT$ (red), no significant difference (green), $BootYT$ better than $BICdof$ (blue). Right: Q^2 better than $BootYT$ (red), no significant difference (green), $BootYT$ better than Q^2 (blue). *Under*: PLS-logistic case; $BootYT$ better (red), no significant difference (green).

References

[1] N. Krämer and M. Sugiyama, "The degrees of freedom of partial least squares regression," *Journal of the American Statistical Association* **106**(494), 2011.

[2] P. Bastien, V. E. Vinzi, and M. Tenenhaus, "Pls generalised linear regression," *Computational Statistics & Data Analysis* **48**(1), pp. 17–46, 2005.

[3] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*, vol. 57, Chapman & Hall/CRC, 1993.

[4] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI*, **14**, pp. 1137–1145, 1995.

E.2 Poster

Nous fournissons ici le poster présenté lors de la 8$^{\text{th}}$ *International Conference on Partial Least Squares and Related Methods* qui s'est déroulée à Paris du 26 au 28 Mai 2014.

A new bootstrap based stopping criterion in PLS components construction

1/ Introduction

The extraction of the optimal number of PLS components, which are noted $\mathbf{T} = (\mathbf{t_1}, \ldots, \mathbf{t_k})$ is the central and main important point in order to well perform a PLS regression (*RPLS*). Indeed, it models the real dimension of the link between the response $\mathbf{y}$, which will be a $\mathbb{R}^{n\times 1}$ vector, and the predictors matrix $\mathbf{X}$.

Considering k as the optimal number of components, the final regression model is the following one:

$$g(\theta) = \sum_{h=1}^{k} c_h \mathbf{t_h}$$
$$= \sum_{h=1}^{k} c_h \left(\sum_{j=1}^{p} w_{hj}^{*} \mathbf{x_j} \right)$$

with θ the conditional expectancy of $\mathbf{y}$ in case of a continuous distribution or the probability vector of a discrete law with a finite support. The link function g is chosen according to the distribution of $\mathbf{y}$ to best fit the model to the data.

2/ Graphically reported existing criteria

PLS framework

BICdof

BIC criterion corrected with the adapted degrees of freedom (Krämer & Sugiyama, 2011).

Q²

based on 5 Cross-Validation (*5-CV*)

$Q_h^2 = 1 - \frac{PRESS_h}{RSS_{h-1}}$

PLS-GLM framework

p_val

Bastien *et al.* (2005) define a stopping rule as the non-significance of a new component $\mathbf{t_h}$ if there is no any significant predictor in it (use of an asymptotic Wald test).

CV—MClassed

Through a *5-CV*, it returns the number of predicted miss-classed values. The selected model will be the one which returns the minimum of them.

3/ Motivations

Some defaults of existing stopping criteria:

- arbitrary limit values
- dependency on asymptotic law
- dependency on the value of q in q-*CV*
- applicable in some precise cases

The aim was also to develop:

- an universal method
- a method which test the significance of a new component for a fixed level α
- a method which deals with heteroscedasticity

➔ **Adaptation of the so-called non-parametric bootstrapping pairs technique.**

4/ Bootstrap based stopping criterion

Assumption

This method relies on the assumption that, once the components are build, they are considered as classical predictors. Indeed, in order to test the significance of a new component, we could not simulate the real distribution of the coefficient linked to this component, which would be a positive one, but we will rather approach the conditional distribution of these coefficients given $(\mathbf{t_1}, \ldots, \mathbf{t_h})$. In a word, we focus on: $\mathbb{P}(c_h > 0 | (\mathbf{t_1}, \ldots, \mathbf{t_h}))$.

Algorithmic criterion

Let $j = 1$ and $R = 500$

1. Compute the j firsts components $(\mathbf{t_1}, \ldots, \mathbf{t_j})$
2. Bootstrapping pair $(\mathbf{y}, \mathbf{T})$, returning R bootstrap samples noted $(\mathbf{y}, \mathbf{T})^{b_1}, \ldots, (\mathbf{y}, \mathbf{T})^{b_R}$
3. For each couple $(\mathbf{y}, \mathbf{T})^{b_i}$, do the (generalized) linear regression:

$$\mathbf{y}^{b_i} = \sum_{h=1}^{j} \left(\hat{c}_h^{b_i} . \mathbf{t_h}^{b_i} \right) + \hat{\epsilon}_j^{b_i}$$

4. Since $c_j > 0$, construct an unilateral BC_a CI: $\mathrm{CI} = \left[CI_1^j, +\infty \right[$
5. While $CI_1^j > 0$, do j=j+1, return to step 1 Else, $k = j - 1$

5/ Simulation algorithm

Data simulation process for *RPLS* and *RPLS-GLM* analysis

n = number of subjects
p_l = number of predictors in the simulated dataset.
Let $\mathbf{T} = (t_{k\rho})_{\substack{1\leq k\leq 4 \\ 1\leq \rho\leq p_l}}$ be a $\mathbb{R}^{4\times p_l}$ matrix, representing 4 fixed orthogonal components in $\mathbf{X}$.
for $j = 0.01, \ldots, \eta$ **by** γ **do**
 for $h = 0.01, \ldots, \xi$ **by** λ **do**
 $\sigma = (\sigma_1, \ldots, \sigma_5) = (10, 8, 6, h, j)$
 $\tau = (\tau_1, \ldots, \tau_5) = (0.25, 0.125, 0.05, 0.0125, 0.005)$
 for $l = 1, \ldots, 100$ **do**
 for $i = 1, \ldots, n$ **do**
 Let r_k be a realisation of $R_k \sim \mathcal{N}(0, \sigma_k^2)$ for $k = 1, \ldots, 5$
 Let f_k be a realisation of $F_k \sim \mathcal{N}(0, \tau_k^2)$ for $k = 1, \ldots, 5$
 Let $\epsilon_{i\rho}$ values be p_l realisations of $\epsilon \sim \mathcal{N}(0, 10^{-4})$
 Let δ_i be a realisation of $\delta \sim \mathcal{N}(0, 10^{-3})$

$$x_{i\rho} = \sum_{k=1}^{4} r_k . t_{k\rho} + \epsilon_{i\rho},\ \rho = 1, \ldots, p_l$$

$$\theta_i = ((r_1 + f_1)/2) + ((r_2 + f_2)/2) + ((r_3 + f_3)/2) + ((r_5 + f_5)/2) + \delta_i$$

$$y_i = \begin{cases} \theta_i & \text{, PLS case} \\ \mathcal{B}(inv.logit(\theta_i)) & \text{, PLS-logistic case} \end{cases}$$

 Apply the different criteria, depending on the case, on the dataset.
 Extract the retained number of component for each of them.
 end for
 end for
end for
end for

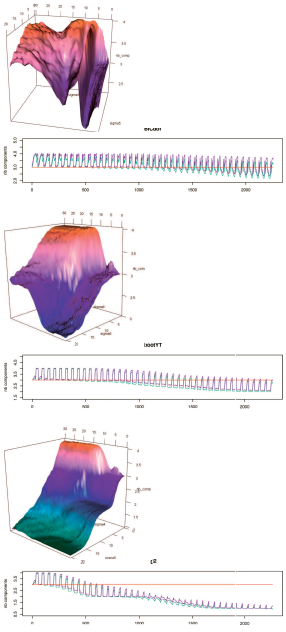
Simulations have been realised with *R* under two different cases, leading to different values of simulation parameters:

1. $n < p : n = 20,\ p_l \in \{25, \ldots, 50\}$ + 80 subjects for testing predictive ability

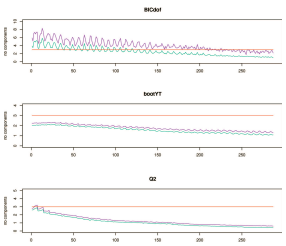
2. $n > p : n = 200,\ p_l \in \{7, \ldots, 50\}$

6/ Results

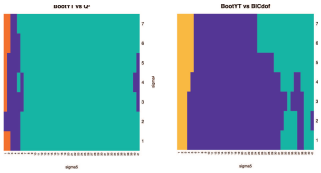
PLS $n > p$



PLS $n < p$

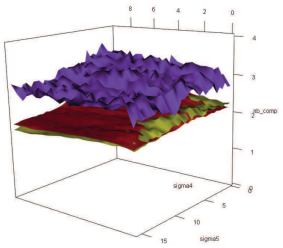


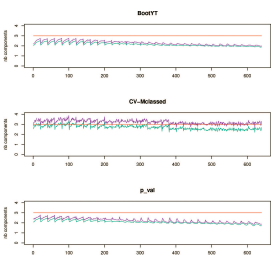
t-tests results on predictive MSE $n < p$



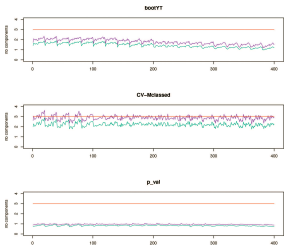
BootYT better
BICdof better
Q² better
No significant difference

PLS-Logistic $n > p$

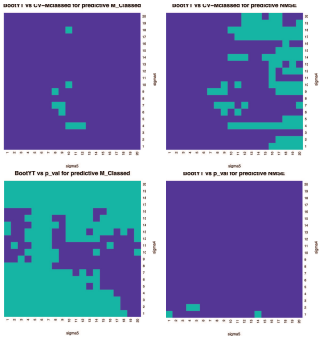




PLS-Logistic $n < p$



t-tests results $n < p$



References

- Bastien, P., Vinzi, V.E., and Tenenhaus, M. (2005). *Pls generalised linear regression. Computational Statistics & Data Analysis,* 48(1), 17–46.
- Krämer, N. and Sugiyama, M. (2011). *The degrees of freedom of partial least squares regression. Journal of the American Statistical Association,* 106(494), 697-705.

Acknowledgements

8th International Conference on Partial Least Squares and Related Methods
26-28 May 2014 Paris, France

Authors

Jérémy Magnanensi[1,2]
Frédéric Bertrand[1]
Myriam Maumy-Bertrand[1]
Nicolas Meyer[2]

[1] Institut de Recherche Mathématique Avancée LabEx IRMIA
[2] Faculté de Médecine Laboratoire de Biostatistique EA3430
—
Université de Strasbourg, Strasbourg, France

Annexe F

Article en cours de parution dans les actes de la conférence PLS14

A New Bootstrap-based Stopping Criterion in PLS Components Construction

Jérémy Magnanensi, Myriam Maumy-Bertrand, Nicolas Meyer and Frédéric Bertrand

Abstract We develop a new universal stopping criterion in components construction, in the sense that it is suitable both for Partial Least Squares Regressions (PLSR) and its extension to Generalized Linear Regressions (PLSGLR). This criterion is based on a bootstrap method and has to be computed algorithmically. It allows testing each successive component on a significant level α. In order to assess its performances and robustness with respect to different noise levels, we perform intensive datasets simulations, with a preset and known number of components to extract, both in the case $N > P$ (N being the number of subjects and P the number of original predictors), and for datasets with $N < P$. We then use t-tests to compare the predictive performance of our approach to some others classical criteria. Our conclusion is that our criterion presents better performances, both in PLSR and PLS-Logistic Regressions (PLS-LR) frameworks.

Jérémy Magnanensi
Institut de Recherche Mathématique Avancée, UMR 7501, LabEx IRMIA, Université de Strasbourg et CNRS, 7, Rue Rene Descartes 67084 Strasbourg Cedex, France
Laboratoire de Biostatistique et Informatique Médicale, Faculté de Médecine, EA3430, Université de Strasbourg, 4, Rue Kirschleger 67085 Strasbourg Cedex, France, e-mail: magnanensi@math.unistra.fr

Myriam Maumy-Bertrand
Institut de Recherche Mathématique Avancée, UMR 7501, Université de Strasbourg et CNRS, 7, Rue Rene Descartes 67084 Strasbourg Cedex, France, e-mail: mmaumy@math.unistra.fr

Nicolas Meyer
Laboratoire de Biostatistique et Informatique Médicale, Faculté de Médecine, EA3430, Université de Strasbourg, 4, Rue Kirschleger 67085 Strasbourg Cedex, France, e-mail: nmeyer@unistra.fr

Frédéric Bertrand
Institut de Recherche Mathématique Avancée, UMR 7501, Université de Strasbourg et CNRS, 7, Rue Rene Descartes 67084 Strasbourg Cedex, France, e-mail: fbertrand@math.unistra.fr

1 Introduction

Performing usual linear regressions between an univariate response $\mathbf{y} = (y_1, \ldots, y_N) \in \mathbb{R}^{N\times 1}$ and highly correlated predictors $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_P) \in \mathbb{R}^{N\times P}$, with N the number of subjects and P the number of predictors, or on datasets including more predictors than subjects, is not suitable or even possible. However, with the huge technological and computer science advances, providing consistent analysis of such datasets has become a major challenge, especially in domains such as medicine, biology or chemistry. To deal with them, statistical methods have been developed, especially the PLS Regression (PLSR) which was introduced by Wold *et al.* (1983) and Wold *et al.* (1984) and described precisely by Höskuldsson (1988) and Wold *et al.* (2001).

PLSR consists in building $K \leqslant \mathrm{rk}(\mathbf{X})$ orthogonal "latent" variables $\mathbf{T}_K = (\mathbf{t}_1, \ldots, \mathbf{t}_K)$, also called components, in such a way that $\mathbf{T}_K$ describes optimally the common information space between $\mathbf{X}$ and $\mathbf{y}$. Thus, these components are build up as linear combinations of the predictors, in order to maximise the covariances $\mathrm{Cov}(\mathbf{y}, \mathbf{t}_h)$ so that:

$$\mathbf{t}_h = \mathbf{X}\mathbf{w}_h^* = \sum_{j=1}^{P} w_{jh}^* \mathbf{x}_j,\ 1 \leqslant h \leqslant K \tag{1}$$

where $\mathbf{w}_h^* = (w_{1h}{}^*, \ldots, w_{Ph}{}^*)^T$ is the vector of predictors weights in the h^{th} component (Wold *et al.*, 2001) and $(.)^T$ represents the transpose.

Let K be the number of components. The final regression model is:

$$\mathbf{y} = \sum_{h=1}^{K} c_h \mathbf{t}_h + \varepsilon = \sum_{h=1}^{K} c_h \left(\sum_{j=1}^{P} w_{jh}^* \mathbf{x}_j \right) + \varepsilon, \tag{2}$$

with $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_N)^T$ the N by 1 error vector, verifying $\mathbb{E}(\varepsilon\,|\mathbf{T}_K) = 0_N$, $\mathbb{V}\mathrm{ar}(\varepsilon\,|\mathbf{T}_K) = \sigma_\varepsilon^2 \times Id_N$ and $(c_1, \ldots, c_K)$ the coefficients of regression of $\mathbf{y}$ on $\mathbf{T}_K$.

An extension to Generalized Linear Regression models, noted PLSGLR, has been developed by Bastien *et al.* (2005), with the aim of taking into account the specific distribution of $\mathbf{y}$. In this context, the regression model is the following one:

$$g(\theta) = \sum_{h=1}^{K} c_h \left(\sum_{j=1}^{P} w_{jh}^* \mathbf{x}_j \right), \tag{3}$$

with θ the conditional expected value of $\mathbf{y}$ for a continuous distribution or the probability vector of a discrete law with a finite support. The link function g depends on the distribution of $\mathbf{y}$.

The determination of the optimal number of components K, which is equal to the exact dimension of the link between $\mathbf{X}$ and $\mathbf{y}$, is crucial to obtain correct estimations of the original predictors coefficients. Indeed, concluding $K_1 < K$ leads to a loss of information so that links between some predictors and $\mathbf{y}$ will not be correctly modelled. Concluding $K_2 > K$ involves that useless information in $\mathbf{X}$ will be used to model knowledge in $\mathbf{y}$, which leads to over-fitting.

2 Criteria Compared Through Simulations

2.1 Existing Criteria Used for Comparison.

- In PLSR:
  1. $\mathbf{Q^2}$. This criterion is obtained by Cross-Validation (CV) with q, the number of parts the dataset is divided, chosen equal to five (5-CV), according to results obtained by Kohavi (1995) and Hastie *et al.* (2009). For a new component $\mathbf{t}_h$, Tenenhaus (1998) considers that it improves significantly the prediction if:
  $$\sqrt{PRESS_h} \leqslant 0.95\sqrt{RSS_{h-1}} \iff Q_h^2 \geqslant 0.0975.$$
  2. **BICdof**. Krämer and Sugiyama (2011) define a *dof* correction in the PLSR framework (without missing data) and apply it to the BIC criterion. We used the *R* package *plsdof*, based on Krämer and Sugiyama (2011) work, to obtain values of this corrected BIC and selected the model that realizes the first local minimum of this BICdof criterion.
- In PLSGLR:
  1. $\mathbf{CV-MClassed}$. This criterion could only be used for PLS-Logistic Regressions (PLS-LR). Through a 5-CV, it determines for each model the number of predicted missclassified values. The selected model is the one linked to the minimal value of this criterion.
  2. **p_val**. Bastien *et al.* (2005) define a new component $\mathbf{t}_h$ as non-significant if there is no significant predictor within it. An asymptotic Wald test is used to conclude to the significance of the different predictors.

2.2 Bootstrap Based Criterion

All the criteria described just above have major flaws including arbitrary bounds dependency, results based on asymptotic laws or derived from q-CV which naturally depends on the value of q and on the way the group will be randomly drawn.
For this purpose, we adapted non-parametric bootstrap techniques in order to test directly, with some confidence level $(1-\alpha)$, the significance of the different coefficients c_h by extracting confidence intervals (CI) for each of them.

The significance of a new component $\mathbf{t}_H$ can not be tested by simulating the usual conditional distribution given $\mathbf{X}$ of its regression coefficient linked to $\mathbf{y}$ since it would be a positive one. Since $\mathbf{t}_H$ maximizes $\mathrm{Cov}(\mathbf{y}, \mathbf{t}_H | \mathbf{T}_{H-1})$, we approached the conditional distribution given $\mathbf{T}_{H-1}$ to test each new component.

We define the significance of a new component as resulting from its significance for both $\mathbf{y}$ and $\mathbf{X}$, so that the extracted number of components K is defined as the

last one which is significant for both of them.

Bootstrapping pairs was introduced by Freedman (1981). This technique relies on the assumption that the originals pairs $(y_i, \mathbf{t}_{i\bullet})$, where $\mathbf{t}_{i\bullet}$ represents the i^{th} row of $\mathbf{T}_H$, are randomly sampled from some unknown $(H+1)$-dimensional distribution. This technique was developed to treat the so called correlation models, in which predictors are considered as random and ε may be related to them.
In order to adapt it to PLSR and PLSGLR frameworks, we designed the following double bootstrapping pairs algorithmic implementation, with $R = 500$, which will be graphically reported as **BootYT**. To avoid confusions between the number of predictors and the coefficients of the regressions of $\mathbf{X}$ on $\mathbf{T}_H$, we set M as the total number of predictors.

- Bootstrapping $(\mathbf{X}, \mathbf{T}_H)$: let $H = 1$ and $l = 1, \ldots, M$.
  1. Compute the H first components $(\mathbf{t}_1, \ldots, \mathbf{t}_H)$.
  2. Bootstrap pair $(\mathbf{X}, \mathbf{T}_H)$, returning R bootstrap samples $(\mathbf{X}, \mathbf{T}_H)^{b_r}$, $1 \leqslant r \leqslant R$.
  3. For each $(\mathbf{X}, \mathbf{T}_H)^{b_r}$, do M least squares regressions $\mathbf{x}_l^{b_r} = \sum_{h=1}^{H} \left(\hat{p_{lh}}^{b_r} . \mathbf{t}_h^{b_r} \right) + \hat{\delta}_{lH}^{b_r}$.
  4. $\forall p_{lH}$, construct a $(100 \times (1-\alpha))\%$ bilateral BC_a CI, noted $\mathrm{CI}_l = \left[\mathrm{CI}_{l,1}^H, \mathrm{CI}_{l,2}^H \right]$.
  5. **If** $\exists l \in \{1, \ldots, M\}$, $0 \notin \mathrm{CI}_l$, **then** $H = H + 1$ and return to step 1. **Else**, $K_{max} = H - 1$.

- Bootstrapping $(\mathbf{y}, \mathbf{T}_H)$: let $H = 1$. Note that for PLSGLR, a generalized regression is performed at step 3.
  1. Compute the H first components $(\mathbf{t}_1, \ldots, \mathbf{t}_H)$.
  2. Bootstrap pair $(\mathbf{y}, \mathbf{T}_H)$, returning R bootstrap samples $(\mathbf{y}, \mathbf{T}_H)^{b_r}$, $1 \leqslant r \leqslant R$.
  3. For each pair $(\mathbf{y}, \mathbf{T}_H)^{b_r}$, do the LS regression $\mathbf{y}^{b_r} = \sum_{h=1}^{H} \left(\hat{c}_h^{b_r} . \mathbf{t}_h^{b_r} \right) + \hat{\varepsilon}_H^{b_r}$.
  4. Since $c_H > 0$, construct a $(100 \times (1-\alpha))\%$ unilateral BC_a CI $= \left[\mathrm{CI}_1^H, +\infty \right[$ for c_H.
  5. **While** $\mathrm{CI}_1^H > 0$ and $H \leqslant K_{max}$, **do** $H = H + 1$, and return to step 1. **Else**, the final extracted number of components is $K = H - 1$.

2.3 Simulation Plan

To compare these different criteria, datasets simulations have been performed by adapting the *simul_data_UniYX* function, available in the *R* package *plsRglm* (Bertrand *et al.*, 2014).

Simulations were performed to obtain a three dimensions common space between $\mathbf{X}$ and $\mathbf{y}$, leading to an optimal number of components equal to three. They were performed under two different cases, both in PLSR and PLSGLR framework. The first one is the $N > P$ situation with $N = 200$ and $P \in \Omega_{200} = \{7, \ldots, 50\}$.

The second one is the $N < P$ situation where $N = 20$ and $P \in \Omega_{20} = \{25, \ldots, 50\}$. For each fixed couple (σ_4, σ_5), which respectively represents the standard deviation owned by the useless fourth component present in $\mathbf{X}$ and the random noise standard deviation in $\mathbf{y}$, we simulated 100 datasets with P_l predictors, $l = 1, \ldots, 100$, obtained by sampling with replacement in Ω_N.

3 PLSR Results

3.1 PLSR: Case $N > P$

Results are stored in three tables (one per criterion) of dimension 2255×100. The first 1230 rows correspond to results for fixed couples of values (σ_4, σ_5), with $\sigma_4 \in \{0.01, 0.21, \ldots, 5.81\}$ and $\sigma_5 \in \{0.01, 0.51, \ldots, 20.01\}$. The 1025 remaining rows correspond to results for $\sigma_4 \in \{6.01, 7.01, \ldots, 30.01\}$. Columns correspond to the 100 datasets simulated per couple.

We extract each row means and report them in Fig.1 as a function of σ_4 and σ_5. Each row variances were also extracted and reported in Fig.2.

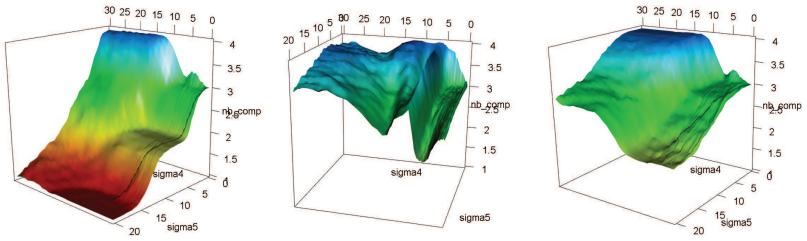


Fig. 1 Left: Q^2 row means. Centre: BICdof row means. Right: BootYT row means.

Considering the extracted number of components as a discriminant factor, we conclude that the Q^2 criterion is the less efficient criterion by being the most sensitive one to the increasing value of σ_5 so that it globally underestimates the number of components. Comparing BICdof and BootYT, or advertising one of them is quite difficult in this large N case. BICdof has a low computational runtime and is the less sensitive one to the increasing value of σ_5. However, referring to Fig.2, the variability of results linked to the BICdof is globally higher than the one linked to our new bootstrap based criterion, especially on datasets with large values of σ_4. BootYT is more robust than the BICdof to the increasing noise level in $\mathbf{X}$ and also directly applicable to the PLSGLR case. However, its computational runtime is clearly higher since, for each dataset, it requires $(K \times ((P_l + 1) \times R))$ least squares regressions.

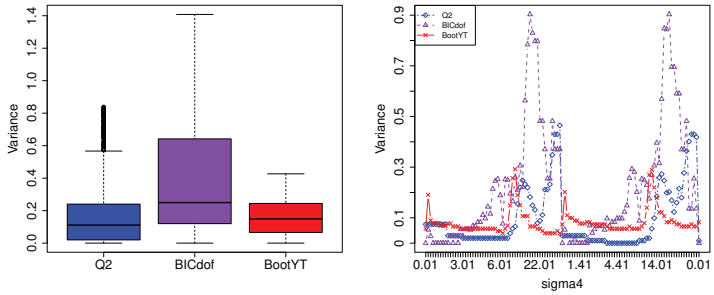


Fig. 2 Left: Boxplots of each row variances. Right: Evolution of variances for $\sigma_5 = \{5.01, 5.51\}$.

3.2 PLSR: Case $N < P$

This small training sample size allows us to consider high-dimensional settings and is very interesting since usually least squares regression could not be performed.

Results are stored in three tables of dimension 287×100, each row corresponds to results for fixed couples of values (σ_4, σ_5), with $\sigma_4 \in \{0.01, 1.01, \ldots, 6.01\}$ and $\sigma_5 \in \{0.01, 0.51, \ldots, 20.01\}$. Row means are represented as a function of σ_4 and σ_5 in Fig.3 and graphical representations of row variances were performed in Fig.4.

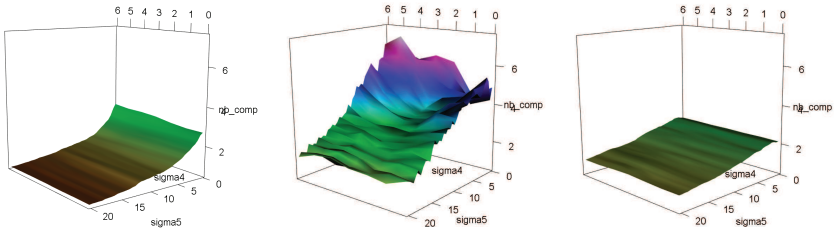


Fig. 3 Left: Q^2 row means. Centre: BICdof row means. Right:BootYT row means.

In this particular case, based on Fig.4, the BootYT criterion returns results with low variability for fixed couple (σ_4, σ_5) contrary to the BICdof criterion, which moreover is the most sensitive one to the increasing noise level in $\mathbf{y}$. Q^2 has a comparable attractive feature of stability but is less robust to noise level in $\mathbf{y}$ than our new bootstrap based criterion. So, by considering the number of extracted components as a discriminant factor, we conclude that the BootYT criterion is the best one to deal with these $N < P$ datasets.

However, we wanted to assess the predictive performances of each of these three criteria. Thus, for each of the 287 000 simulated datasets, we simulated 80 more

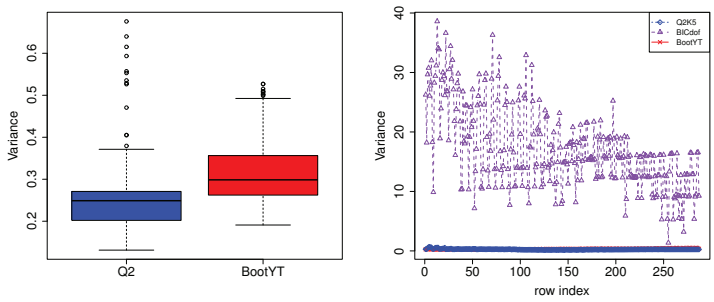


Fig. 4 Left: Boxplot of row variances. Right: Evolution of row variances.

observations as test points and computed testing Normalized Mean Square Error (NMSE). The normalisation was done by dividing the testing MSE of the obtained model with the MSE linked to the trivial one (constant model equal to the mean of the training data). Furthermore, as mentioned by Krämer and Sugiyama (2011, p.702), "the large test sample size ensures a reliable estimation of the test error."

In order to compare the predictive performances of the three criteria depending on noise levels, we treat these predictive results for each couple of values (σ_4, σ_5) by testing the equalities of NMSE means with asymptotic t-tests with Welch-Satterthwaite *dof* approximation (Welch, 1947). All these tests were performed on level $\alpha = 0.05$. Results of these t-tests are graphically reported in Fig. 5.

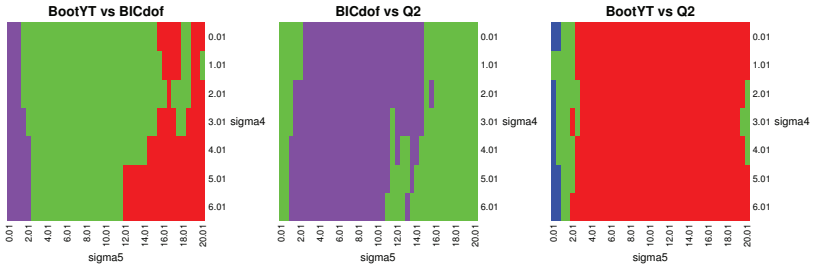


Fig. 5 t-tests results: BootYT better (red), BICdof better (purple), Q2K5 better (blue), no significant difference (green).

Concerning BootYT vs Q^2, the Q^2 has a better predictive ability for some very low values of σ_5. This result is not surprising since, in this case, the Q^2 criterion returns numbers of components closer to three than BootYT does (Fig. 3). However, tests results between the BICdof and the Q^2 criterion are not concluding to a significant better predictive performance of the Q^2 criterion for small values of σ_5 despite the BICdof globally overestimates the number of components in this case

(Fig. 3). In fact, due to the small values of σ_5, the 80 additional responses we simulated almost follow the same model than the first 20 ones. Thus, predictive NMSE react in the same way than the training ones *i.e.* the higher the extracted number of components is, the lower the predictive NMSE are. This fact lead us to only focus on the extracted number of components when $\sigma_5 \simeq 0$, leading the Q^2 criterion to be the best one.

Finally, in all others cases, the BootYT criterion returns models with, at least, comparable or better predictive abilities than the two others.

3.3 PLSR: Conclusion

In the $N > P$ case, the BootYT criterion offers a better robustness to noise in $\mathbf{y}$ than the Q^2. It is also more robust to the increasing noise level in $\mathbf{X}$ than the BICdof, which moreover has some variance issues for high values of σ_4. We also conclude the BootYT criterion as a good compromise between the two others criteria, owning their advantages without their drawbacks. Concerning the $N < P$ case, our bootstrap-based criterion is globally the best one since it is less sensitive than the others to the increasing noise level in $\mathbf{y}$ and is linked to low variance results, leading to global better predictive performances.

4 PLS-LR Results

In this framework, due to the specific distribution of $\mathbf{y}$ and link-function $g = inv.logit$, the increase of σ_5 does not lead to a linear increase of noise level in $\mathbf{y}$. The bijectivity of g insures the presence of three common components between $\mathbf{X}$ and $\mathbf{y}$.

4.1 PLS-LR: Case $N > P$

Results are stored in three tables of dimension 640×100, each row corresponds to results for fixed couples of values (σ_4, σ_5), with $\sigma_4 \in \{0.01, 0.51, 1.01, \ldots, 9.51\}$ and $\sigma_5 \in \{0.01, 0.51, 1.01, \ldots, 15.51\}$. We graphically report row means as a function of σ_4 and σ_5 as well as boxplots of row variances in Fig.6.

Based on these graphics, the CV-MClassed performs well in estimating the optimal number of components in average. However, this good property has to be nuanced by the high variances linked to its results and which lead this criterion to be used with caution. The BootYT and p_val criteria return similar results in this asymptotic case. Both of them slightly underestimate the optimal number of components but with the advantage of low variances of their results.

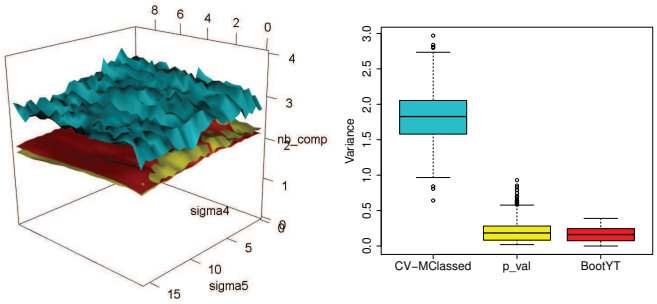


Fig. 6 Left: Row means surfaces, from top to bottom: CV-MClassed, BootYT, p_val. Right: Boxplots of row variances.

4.2 PLS-LR: Case $N < P$

Results are stored in three tables of dimension 400×100, each row corresponds to results for fixed couples of values (σ_4, σ_5), with $\sigma_4 \in \{0.01, 0.51, 1.01, \ldots, 9.51\}$ and $\sigma_5 \in \{0.01, 0.51, 1.01, \ldots, 9.51\}$. We set the maximal value of σ_5 to 9.51, and not to 15.51 as for the $N > P$ case, in order to save computational runtime since an increasing value of σ_5 does not really affect the choice of the number of extracted components.

We graphically report row means as a function of σ_4 and σ_5 as well as boxplots of row variances in Fig.7.

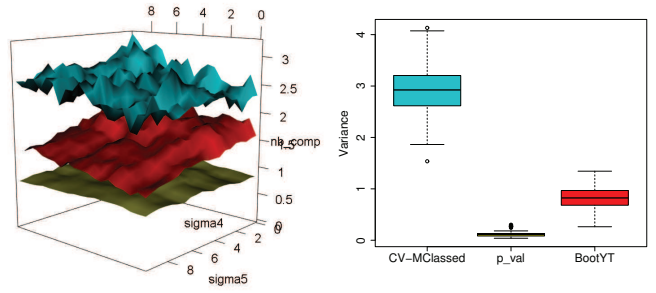


Fig. 7 Left: Row means surfaces, from top to bottom: CV-MClassed, BootYT, p_val. Right: Boxplots of row variances.

The CV-MClassed criterion conserves the same property of well estimating in average and issue of variability as in the $N > P$ framework. Concerning the two others criteria, we observed a higher underestimating issue linked to the p_val criterion

than for the BootYT one. Furthermore, they both had low variability in results they return.

In order to test their predictive performances, we simulated 80 more observations for each simulated datasets (40 000), and computed the predictive NMSE linked to each models established by the three criteria. Furthermore, since the binary response obtained by the model is equal to 1 if the estimated response is over 0.5, 0 if not, returning higher NMSE does not necessarily lead to higher number of missclassified values. Thus, we also computed the number of predictive missclassified values (*M_classed*) for each of these three criteria. Then, *t*-tests were computed for each fixed values of (σ_4, σ_5). Results of these tests are graphically reported in Fig.8.

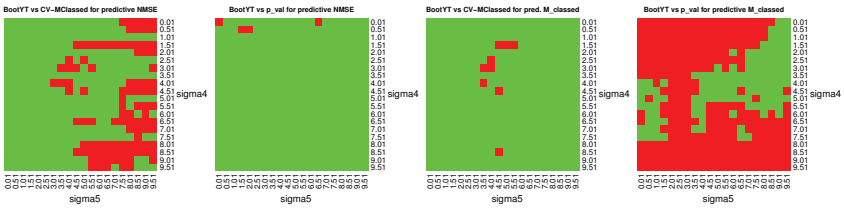


Fig. 8 *t*-tests results: BootYT better (red), no significant difference (green).

The bootstrap-based criterion is never less efficient than the other criteria. If there is globally no significant differences between bootstrapping pairs or the p_val criterion concerning the predictive NMSE, BootYT is better than this criterion concerning the predictive missclassified values. Then, there is few cases where bootstrapping pairs is significantly better than the CV-MClassed criterion concerning the predictive number of missclassified values. But, concerning the predictive NMSE, the BootYT criterion is better than this last one by returning significant smallest NMSE values, especially for high σ_5 values.

The boostrap-based criterion is also the best one by having, at least, similar predictive performances compared to the two others.

4.3 PLS-LR: Conclusion

Through these simulations, we can reasonably assume that the bootstrap-based criterion is globally more efficient than the other ones. In the $N > P$ case, it offers a similar stability compared to the p_val criterion. However, it globally underestimates the optimal number of components when the CV-MClassed criterion retains it on average but with high variabilities. Concerning the $N < P$ case, the BootYT criterion has better predictive performances than the two others studied criteria in terms of predictive NMSE and predictive missclassified values. It also keeps a quite low variability, which is really important for a future routine implementation.

5 Discussion

Our new bootstrap based criterion requires huge computational runtime, so that an optimization of the algorithm seems necessary. Furthermore, the development of corrected *dof* in PLSGLR framework would also permit to develop a corrected BIC formulation in this framework. This corrected BIC could provide an interesting alternative to the boostrap-based criterion since it could save an important computational runtime conditionally to the fact that it would have at least similar properties to those we conclude in Section 3.

However, this new criterion represents a reliable, consistent and universal stopping criterion in order to select the optimal number of *PLS* components. It also allows users to test the significance of a new component with a preset risk level α.

In the $N > P$ PLSR framework, our simulations confirm the BICdof as being an appropriate and well-designed criterion. However, our new bootstrap-based criterion is an appropriate alternative in the $N < P$ case, since the BICdof criterion suffers from overestimating issues for models with low random noise levels in $\mathbf{y}$ and returns results linked to high variances. Furthermore, both BICdof and Q^2 criteria are more sensitive than the bootstrap-based criterion to the increasing noise level in $\mathbf{y}$.

Concerning the PLSGLR framework, our simulations results lead to advertise this new bootstrap-based criterion. Indeed, in this PLS-LR case, we show that depending on the statistic we used (testing NMSE or predictive number of missclassified values) to test its predictive ability, the bootstrap-based is never significantly worse than both the CV-MClassed and p_val criteria.

References

Bastien, P., Vinzi, V. E., and Tenenhaus, M. (2005). PLS Generalised Linear Regression. *Computational Statistics & Data Analysis*, **48**(1), 17–46.

Bertrand, F., Magnanensi, J., Maumy-Bertrand, M., and Meyer, N. (2014). *Partial Least Squares Regression for Generalized Linear Models*. Book of abstracts, User2014!, Los Angeles, page 150.

Freedman, D. A. (1981). Bootstrapping Regression Models. *The Annals of Statistics*, **9**(6), 1218–1228.

Hastie, T., Tibshirani, R., and Friedman, J. J. H. (2009). *The Elements of Statistical Learning, second edition*, volume 1. Springer New York.

Höskuldsson, A. (1988). Pls Regression Methods. *Journal of Chemometrics*, **2**(3), 211–228.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the 14th international joint conference on Artificial intelligence-Volume 2*, pages 1137–1143. Morgan Kaufmann Publishers Inc.

Krämer, N. and Sugiyama, M. (2011). The Degrees of Freedom of Partial Least Squares Regression. *Journal of the American Statistical Association*, **106**(494), 697–705.

Tenenhaus, M. (1998). *La Régression PLS, Théorie et pratique*. Editions Technip.

Welch, B. L. (1947). The Generalization of Student's Problem when Several Different Population Variances are Involved. *Biometrika*, **34**(1-2), 28–35.

Wold, S., Martens, H., and Wold, H. (1983). The Multivariate Calibration Problem in Chemistry Solved by the PLS Method. In *Matrix Pencils*, pages 286–293. Springer.

Wold, S., Ruhe, A., Wold, H., and Dunn, III, W. J. (1984). The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *j-SIAM-J-SCI-STAT-COMP*, **5**(3), 735–743.

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-Regression: a Basic Tool of Chemometrics. *Chemometrics and intelligent laboratory systems*, **58**(2), 109–130.

Annexe G

Résultats théoriques quant au nombre de composantes pour un échantillon bootstrap

We developed some results concerning the differences of significant number of components between an original dataset D^{ori} linked to a number of components $K_{D^{ori}}$ and an extracted bootstrap sample.

Let $D^b = \left(\mathbf{y}^b, \mathbf{X}^b\right) \in \mathcal{M}_{n,p+1}(\mathbb{R})$ be a resampled dataset with replacement so that $\mathbf{y}^b$ is centered. Let $m \leqslant n$ be the number of different subjects involved in D^b. Each different subject δ_i, $1 \leqslant i \leqslant m$ appears q_{δ_i} times, so that $\sum_{i=1}^{m} q_{\delta_i} = n$. We note K_{D^b} the number of components linked to this resampled dataset.
Let $D^{red} = \left(\mathbf{y}^{red}, \mathbf{X}^{red}\right) \in \mathcal{M}_{m,p+1}(\mathbb{R})$ be the dataset formed by the m different subject available in D^b and $K_{D^{red}}$ be the number of components linked to D^{red}. Note that $\mathbf{y}^{red}$ is not necessarily centered.

Proposition G.0.1. *Let $0 = q_{\delta_0} < q_{\delta_1} \leqslant q_{\delta_2} \leqslant \ldots \leqslant q_{\delta_m}$ and $\mathbf{w}_1^b$ be the first weight vectors of the PLS regression process on D^b. Let $\langle ., . \rangle$ be the associated scalar product of $\|.\|$, then:*

$$\mathbf{w}_1^b = \frac{\sum_{i=1}^{m} \left(q_{\delta_i} - q_{\delta_{i-1}}\right) \left(\left(\mathbf{X}_{i\leqslant l}^{red}\right)^T \mathbf{y}_{i\leqslant l}^{red}\right)}{\sqrt{\sum_{i=1}^{m} \left(q_{\delta_i} + q_{\delta_{i-1}}\right) \left(q_{\delta_i} - q_{\delta_{i-1}}\right) \left\|\left(\mathbf{X}_{i\leqslant l}^{red}\right)^T \mathbf{y}_{i\leqslant l}^{red}\right\|^2 + \gamma}} \tag{G.1}$$

with $\gamma = 2 \sum_{i<j} \left(q_{\delta_i} - q_{\delta_{i-1}}\right) \left(q_{\delta_j} - q_{\delta_{j-1}}\right) \left\langle \left(\mathbf{X}_{i\leqslant l\leqslant j-1}^{red}\right)^T \mathbf{y}_{i\leqslant l\leqslant j-1}^{red}, \left(\mathbf{X}_{l\geqslant j}^{red}\right)^T \mathbf{y}_{l\geqslant j}^{red} \right\rangle$

Proof. Let $D^b = \left(\mathbf{y}^b, \mathbf{X}^b\right) = \begin{pmatrix} y_1^b & x_{11}^b & \ldots & x_{1p}^b \\ \vdots & & \vdots & \\ y_n^b & x_{n1}^b & \ldots & x_{np}^b \end{pmatrix}$ be a bootstrap sample made of m different subjects $\{\delta_1, \ldots, \delta_m\}$, $m \leqslant n$. Let these subjects be linked to respectively $\{q_{\delta_1} \leqslant \ldots \leqslant q_{\delta_m}\}$ occurrences in D^b. To simplify the notations, we note $D^b = (\mathbf{y}, \mathbf{X})$. We suppose that $\mathbf{y}$ is

centered. Let $D^{red} = \left(\mathbf{y}^{red}, \mathbf{X}^{red}\right) \in \mathcal{M}_{m,p}(\mathbb{R})$ be the reduced matrix composed by only one occurrence of each different subject. Then,

$$
\begin{aligned}
\mathbf{w}_1^b &= \frac{\mathbf{X}^T\mathbf{y}}{\|\mathbf{X}^T\mathbf{y}\|} \\
&= \begin{pmatrix} \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ip}y_i \end{pmatrix} / \left\|\mathbf{X}^T\mathbf{y}\right\| \\
&= \begin{pmatrix} \sum_{i=1}^{m} q_{\delta_i} x_{\delta_i 1} y_{\delta_i} \\ \vdots \\ \sum_{i=1}^{m} q_{\delta_i} x_{\delta_i p} y_{\delta_i} \end{pmatrix} / \left\|\mathbf{X}^T\mathbf{y}\right\| \\
&= \left[\begin{pmatrix} q_{\delta_1} \sum_{i=1}^{m} x_{\delta_i 1} y_{\delta_i} \\ \vdots \\ q_{\delta_1} \sum_{i=1}^{m} x_{\delta_i p} y_{\delta_i} \end{pmatrix} + \begin{pmatrix} \sum_{i=2}^{m} \left(q_{\delta_i} - q_{\delta_1}\right) x_{\delta_i 1} y_{\delta_i} \\ \vdots \\ \sum_{i=2}^{m} \left(q_{\delta_i} - q_{\delta_1}\right) x_{\delta_i p} y_{\delta_i} \end{pmatrix} \right] / \left\|\mathbf{X}^T\mathbf{y}\right\| \\
&= q_{\delta_1} \frac{\left(\mathbf{X}^{red}\right)^T \mathbf{y}^{red}}{\|\mathbf{X}^T\mathbf{y}\|} + \left[\begin{pmatrix} \left(q_{\delta_2} - q_{\delta_1}\right) \sum_{i=2}^{m} x_{\delta_i 1} y_{\delta_i} \\ \vdots \\ \left(q_{\delta_2} - q_{\delta_1}\right) \sum_{i=2}^{m} x_{\delta_i p} y_{\delta_i} \end{pmatrix} + \begin{pmatrix} \sum_{i=3}^{m} \left(q_{\delta_i} - q_{\delta_2}\right) x_{\delta_i 1} y_{\delta_i} \\ \vdots \\ \sum_{i=3}^{m} \left(q_{\delta_i} - q_{\delta_2}\right) x_{\delta_i p} y_{\delta_i} \end{pmatrix} \right] / \left\|\mathbf{X}^T\mathbf{y}\right\|
\end{aligned}
$$

By induction and by defining $q_{\delta_0} = 0$, we obtain that:

$$
\mathbf{w}_1^b = \sum_{i=1}^{m} \left(q_{\delta_i} - q_{\delta_{i-1}}\right) \frac{\left(\mathbf{X}_{i\leqslant l}^{red}\right)^T \mathbf{y}_{i\leqslant l}^{red}}{\|\mathbf{X}^T\mathbf{y}\|} \tag{G.2}
$$

with $\mathbf{y}_{i\leqslant l}^{red} = \left(y_{\delta_i}, y_{\delta_{i+1}}, \ldots, y_{\delta_m}\right)^T$ and $\mathbf{X}_{i\leqslant l}^{red} = \begin{pmatrix} x_{\delta_i 1} & \ldots & x_{\delta_i p} \\ x_{\delta_{i+1} 1} & \ldots & x_{\delta_{i+1} p} \\ & \vdots & \\ x_{\delta_m 1} & \ldots & x_{\delta_m p} \end{pmatrix}$

Then, by considering this result, we obtain that:

$$
\begin{aligned}
\left\|\mathbf{X}^T\mathbf{y}\right\|^2 &= \left\|\sum_{i=1}^{m}\left(q_{\delta_i}-q_{\delta_{i-1}}\right)\left(\mathbf{X}_{i\leqslant l}^{red}\right)^T\mathbf{y}_{i\leqslant l}^{red}\right\|^2\\
&= \sum_{i=1}^{m}\left(q_{\delta_i}-q_{\delta_{i-1}}\right)^2\left\|\left(\mathbf{X}_{i\leqslant l}^{red}\right)^T\mathbf{y}_{i\leqslant l}^{red}\right\|^2\\
&\quad+2\sum_{i<j}\left(q_{\delta_i}-q_{\delta_{i-1}}\right)\left(q_{\delta_j}-q_{\delta_{j-1}}\right)\left(\left(\mathbf{X}_{i\leqslant l}^{red}\right)^T\mathbf{y}_{i\leqslant l}^{red}\right)^T\left(\left(\mathbf{X}_{j\leqslant l}^{red}\right)^T\mathbf{y}_{j\leqslant l}^{red}\right)\\
&= \sum_{i=1}^{m}\left(q_{\delta_i}-q_{\delta_{i-1}}\right)^2\left\|\left(\mathbf{X}_{i\leqslant l}^{red}\right)^T\mathbf{y}_{i\leqslant l}^{red}\right\|^2\\
&\quad+2\sum_{i<j}\left(q_{\delta_i}-q_{\delta_{i-1}}\right)\left(q_{\delta_j}-q_{\delta_{j-1}}\right)\left\|\left(\mathbf{X}_{j\leqslant l}^{red}\right)^T\mathbf{y}_{j\leqslant l}^{red}\right\|^2\\
&\quad+2\sum_{i<j}\left(q_{\delta_i}-q_{\delta_{i-1}}\right)\left(q_{\delta_j}-q_{\delta_{j-1}}\right)\left\langle\left(\mathbf{X}_{i\leqslant l\leqslant j-1}^{red}\right)^T\mathbf{y}_{i\leqslant l\leqslant j-1}^{red},\left(\mathbf{X}_{l\geqslant j}^{red}\right)^T\mathbf{y}_{l\geqslant j}^{red}\right\rangle\\
&= \sum_{i=1}^{m}\left\|\left(\mathbf{X}_{i\leqslant l}^{red}\right)^T\mathbf{y}_{i\leqslant l}^{red}\right\|^2\left(\left(q_{\delta_i}-q_{\delta_{i-1}}\right)^2+2\left(q_{\delta_i}-q_{\delta_{i-1}}\right)\sum_{s=1}^{i-1}\left(q_{\delta_s}-q_{\delta_{s-1}}\right)\right)\\
&\quad+2\sum_{i<j}\left(q_{\delta_i}-q_{\delta_{i-1}}\right)\left(q_{\delta_j}-q_{\delta_{j-1}}\right)\left\langle\left(\mathbf{X}_{i\leqslant l\leqslant j-1}^{red}\right)^T\mathbf{y}_{i\leqslant l\leqslant j-1}^{red},\left(\mathbf{X}_{l\geqslant j}^{red}\right)^T\mathbf{y}_{l\geqslant j}^{red}\right\rangle\\
&= \sum_{i=1}^{m}\left\|\left(\mathbf{X}_{i\leqslant l}^{red}\right)^T\mathbf{y}_{i\leqslant l}^{red}\right\|^2\left(\left(q_{\delta_i}-q_{\delta_{i-1}}\right)^2+2\left(q_{\delta_i}-q_{\delta_{i-1}}\right)q_{\delta_{i-1}}\right)\\
&\quad+2\sum_{i<j}\left(q_{\delta_i}-q_{\delta_{i-1}}\right)\left(q_{\delta_j}-q_{\delta_{j-1}}\right)\left\langle\left(\mathbf{X}_{i\leqslant l\leqslant j-1}^{red}\right)^T\mathbf{y}_{i\leqslant l\leqslant j-1}^{red},\left(\mathbf{X}_{l\geqslant j}^{red}\right)^T\mathbf{y}_{l\geqslant j}^{red}\right\rangle\\
&= \sum_{i=1}^{m}\left\|\left(\mathbf{X}_{i\leqslant l}^{red}\right)^T\mathbf{y}_{i\leqslant l}^{red}\right\|^2\left(\left(q_{\delta_i}-q_{\delta_{i-1}}+q_{\delta_{i-1}}\right)^2-q_{\delta_{i-1}}^2\right)\\
&\quad+2\sum_{i<j}\left(q_{\delta_i}-q_{\delta_{i-1}}\right)\left(q_{\delta_j}-q_{\delta_{j-1}}\right)\left\langle\left(\mathbf{X}_{i\leqslant l\leqslant j-1}^{red}\right)^T\mathbf{y}_{i\leqslant l\leqslant j-1}^{red},\left(\mathbf{X}_{l\geqslant j}^{red}\right)^T\mathbf{y}_{l\geqslant j}^{red}\right\rangle\\
&= \sum_{i=1}^{m}\left(q_{\delta_i}^2-q_{\delta_{i-1}}^2\right)\left\|\left(\mathbf{X}_{i\leqslant l}^{red}\right)^T\mathbf{y}_{i\leqslant l}^{red}\right\|^2\\
&\quad+2\sum_{i<j}\left(q_{\delta_i}-q_{\delta_{i-1}}\right)\left(q_{\delta_j}-q_{\delta_{j-1}}\right)\left\langle\left(\mathbf{X}_{i\leqslant l\leqslant j-1}^{red}\right)^T\mathbf{y}_{i\leqslant l\leqslant j-1}^{red},\left(\mathbf{X}_{l\geqslant j}^{red}\right)^T\mathbf{y}_{l\geqslant j}^{red}\right\rangle\\
&= \sum_{i=1}^{m}\left(q_{\delta_i}+q_{\delta_{i-1}}\right)\left(q_{\delta_i}-q_{\delta_{i-1}}\right)\left\|\left(\mathbf{X}_{i\leqslant l}^{red}\right)^T\mathbf{y}_{i\leqslant l}^{red}\right\|^2\\
&\quad+2\sum_{i<j}\left(q_{\delta_i}-q_{\delta_{i-1}}\right)\left(q_{\delta_j}-q_{\delta_{j-1}}\right)\left\langle\left(\mathbf{X}_{i\leqslant l\leqslant j-1}^{red}\right)^T\mathbf{y}_{i\leqslant l\leqslant j-1}^{red},\left(\mathbf{X}_{l\geqslant j}^{red}\right)^T\mathbf{y}_{l\geqslant j}^{red}\right\rangle
\end{aligned}
$$

□

In order to simplify the formulas and notations, we assume the following assumption.

Assumption G.0.1. *Assume that* $\frac{1}{n-m}\sum_{i=1}^{m}(q_{\delta_i}-1)\,y_{\delta_i}=\frac{1}{m}\sum_{i=1}^{m}y_{\delta_i}$.

Proposition G.0.2. *Under assumption G.0.1,* $\mathbf{y}^{red}$ *is centered so that*

$$\begin{aligned}\mathbf{w}_1^{red}&=\frac{cov\left(\mathbf{X}^{red},\mathbf{y}^{red}\right)}{\|cov(\mathbf{X}^{red},\mathbf{y}^{red})\|}\\&=\frac{\left(\mathbf{X}^{red}\right)^T\mathbf{y}^{red}}{\left\|\left(\mathbf{X}^{red}\right)^T\mathbf{y}^{red}\right\|}.\end{aligned}$$

Let $n=qm+r$ be the Euclidean division of n by m. We focus on and define balanced bootstrap samples as follows:

$$\begin{cases}q_{\delta_i}=q, & 1\leqslant i\leqslant m-r\\ q_{\delta_i}=q+1, & m-(r-1)\leqslant i\leqslant m\end{cases}$$

Let us begin with the case $r=0$. In this particular case, the assumption G.0.1 is satisfied and we obtain the following proposition.

Proposition G.0.3. *Let* D^b *be linked to a balanced plan with* $r=0$*, then:*

$$\forall k\in\left[1,rk\left(\mathbf{X}^{red}\right)\right],\hat{c}_k^b=\hat{c}_k^{red}\ \textit{and}\ \hat{Var}\left(\hat{c}_k^b\right)=\tfrac{m-DoF}{qm-DoF}\hat{Var}\left(\hat{c}_k^{red}\right)<\hat{Var}\left(\hat{c}_k^{red}\right)$$

Proof. Let $D^b=\left(\mathbf{y}^b,\mathbf{X}^b\right)=\begin{pmatrix}y_1^b & x_{11}^b & \dots & x_{1p}^b\\ \vdots & & \vdots & \\ y_n^b & x_{n1}^b & \dots & x_{np}^b\end{pmatrix}$ be a bootstrap sample made of m different subjects $\{\delta_1,\dots,\delta_m\}$, $m\leqslant n$. Let these subjects be linked to respectively $\{q_{\delta_1}\leqslant\dots\leqslant q_{\delta_m}\}$ occurrences in D^b. To simplify the notations, we note $D^b=(\mathbf{y},\mathbf{X})$. We suppose that $\mathbf{y}$ is centered. Let $D^{red}=\left(\mathbf{y}^{red},\mathbf{X}^{red}\right)\in\mathcal{M}_{m,p}(\mathbb{R})$ be the reduced matrix composed by only one occurrence of each different subject. Let $n=qm$ so that $q_{\delta_i}=q$, $\forall i\in\{1,\dots,m\}$ and $q_{\delta_0}=0$. Then,

$$\begin{aligned}\mathbf{w}_1^b&=\frac{\mathbf{X}^T\mathbf{y}}{\|\mathbf{X}^T\mathbf{y}\|}\\&=\frac{\sum_{i=1}^{m}\left(q_{\delta_i}-q_{\delta_{i-1}}\right)\left(\left(\mathbf{X}_{i\leqslant l}^{red}\right)^T\mathbf{y}_{i\leqslant l}^{red}\right)}{\sqrt{\sum_{i=1}^{m}\left(q_{\delta_i}+q_{\delta_{i-1}}\right)\left(q_{\delta_i}-q_{\delta_{i-1}}\right)\left\|\left(\mathbf{X}_{i\leqslant l}^{red}\right)^T\mathbf{y}_{i\leqslant l}^{red}\right\|^2+\gamma}}\quad\text{(Prop. G.0.1)}\\&=\frac{q\left(\mathbf{X}^{red}\right)^T\mathbf{y}^{red}}{\sqrt{q^2\left\|\left(\mathbf{X}^{red}\right)^T\mathbf{y}^{red}\right\|^2}}\\&=\frac{\left(\mathbf{X}^{red}\right)^T\mathbf{y}^{red}}{\left\|\left(\mathbf{X}^{red}\right)^T\mathbf{y}^{red}\right\|}\\&=\mathbf{w}_1^{red}\end{aligned}\tag{G.3}$$

Then, by reordering subjects in an appropriate manner, we obtain that:

$$\mathbf{X} = \left.\begin{pmatrix} \mathbf{X}^{red} \\ \mathbf{X}^{red} \\ \vdots \\ \mathbf{X}^{red} \end{pmatrix}\right\} q \text{ times} \tag{G.4}$$

Due to G.3 and G.4, we have that:

$$\begin{aligned} \mathbf{t}_1^b &= \mathbf{X}\mathbf{w}_1^b \\ &= \begin{pmatrix} \mathbf{X}^{red} \\ \mathbf{X}^{red} \\ \vdots \\ \mathbf{X}^{red} \end{pmatrix} \cdot \mathbf{w}_1^{red} \\ &= \begin{pmatrix} \mathbf{t}_1^{red} \\ \mathbf{t}_1^{red} \\ \vdots \\ \mathbf{t}_1^{red} \end{pmatrix} \end{aligned} \tag{G.5}$$

Finally, due to G.5, we obtain:

$$\begin{aligned} \hat{c}_1^b &= \left(\left(\mathbf{t}_1^b\right)^T \mathbf{t}_1^b\right)^{-1} \left(\mathbf{t}_1^b\right)^T \mathbf{y} \\ &= \left(\begin{pmatrix} \mathbf{t}_1^{red} \\ \mathbf{t}_1^{red} \\ \vdots \\ \mathbf{t}_1^{red} \end{pmatrix}^T \begin{pmatrix} \mathbf{t}_1^{red} \\ \mathbf{t}_1^{red} \\ \vdots \\ \mathbf{t}_1^{red} \end{pmatrix}\right)^{-1} \cdot \begin{pmatrix} \mathbf{t}_1^{red} \\ \mathbf{t}_1^{red} \\ \vdots \\ \mathbf{t}_1^{red} \end{pmatrix}^T \begin{pmatrix} \mathbf{y}^{red} \\ \mathbf{y}^{red} \\ \vdots \\ \mathbf{y}^{red} \end{pmatrix} \\ &= \frac{1}{q}\left(\left(\mathbf{t}_1^{red}\right)^T \mathbf{t}_1^{red}\right)^{-1} \cdot q \left(\mathbf{t}_1^{red}\right)^T \mathbf{y}^{red} \\ &= \hat{c}_1^{red} \end{aligned}$$

By induction due to the PLSR process, we obtain that $\forall k \in \left[1, rk\left(\mathbf{X}^{red}\right)\right], \hat{c}_k^b = \hat{c}_k^{red}$.

Concerning the estimated variances, since $\hat{\text{Var}}\left(\hat{c}_k^b\right) = \left(\hat{\sigma}_k^b\right)^2 \left(\left(\mathbf{t}_k^b\right)^T \mathbf{t}_k^b\right)^{-1}$ with $\left(\hat{\sigma}_k^b\right)^2 = \frac{1}{n-DoF} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$, the same process could be done, leading to the same simplifications. We so obtain that:

$$\left(\left(\mathbf{t}_k^b\right)^T \mathbf{t}_k^b\right)^{-1} = \frac{1}{q}\left(\left(\mathbf{t}_k^{red}\right)^T \mathbf{t}_k^{red}\right)^{-1}$$

and

$$\begin{aligned}\left(\hat{\sigma}_k^b\right)^2 &= \frac{1}{n-DoF}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2\\ &= \frac{q}{n-DoF}\sum_{i=1}^{m}\left(y_{\delta_i}-\hat{y}_{\delta_i}\right)^2\\ &= q\cdot\frac{m-DoF}{(n-DoF)(m-DoF)}\sum_{i=1}^{m}\left(y_{\delta_i}-\hat{y}_{\delta_i}\right)^2\\ &= q\cdot\frac{m-DoF}{n-DoF}\left(\hat{\sigma}_k^{red}\right)^2\end{aligned}$$

□

Corollary G.0.1. *If $r=0$ then:*

$$K_{D^b} \geqslant K_{D^{red}}$$

For a bootstrap sample, the average number of different subject is $m = 0.632\times n$ leading to the following Euclidean division of n, $n = 1\times m + 0.368\times n$. We so focus our study on the $q=1$ particular case and $r \neq 0$ so that $0 = q_{\delta_0} < 1 = q_{\delta_1} = \ldots = q_{\delta_{m-r}} < q_{\delta_{m-(r-1)}} = \ldots = q_{\delta_m} = 2$. We then obtain the following corollary of the proposition G.0.1:

Corollary G.0.2. *Assume the assumption G.0.1 is fulfilled, then:*

$$Corr\left(\mathbf{w}_1^b, \mathbf{w}_1^{red}\right) = \frac{2a+b+3\rho\sqrt{ab}}{\sqrt{\left(2a+b+3\rho\sqrt{ab}\right)^2+\left(1-\rho^2\right)ab}}, \tag{G.6}$$

$$\begin{aligned}\textit{with: } a &= Var\left(\left(\mathbf{X}_{m-r<l}^{red}\right)^T\mathbf{y}_{m-r<l}^{red}\right)\\ b &= Var\left(\left(\mathbf{X}_{m-r\geqslant l}^{red}\right)^T\mathbf{y}_{m-r\geqslant l}^{red}\right)\\ \rho &= Corr\left(\left(\mathbf{X}_{m-r<l}^{red}\right)^T\mathbf{y}_{m-r<l}^{red}, \left(\mathbf{X}_{m-r\geqslant l}^{red}\right)^T\mathbf{y}_{m-r\geqslant l}^{red}\right)\end{aligned}$$

Proof. Let $D^b = \left(\mathbf{y}^b, \mathbf{X}^b\right) = \begin{pmatrix} y_1^b & x_{11}^b & \ldots & x_{1p}^b \\ \vdots & & \vdots & \\ y_n^b & x_{n1}^b & \ldots & x_{np}^b \end{pmatrix}$ be a bootstrap sample made of m different subjects $\{\delta_1, \ldots, \delta_m\}$, $m \leqslant n$. Let these subjects be linked to respectively $\{q_{\delta_1} \leqslant \ldots \leqslant q_{\delta_m}\}$ occurrences in D^b. To simplify the notations, we note $D^b = (\mathbf{y}, \mathbf{X})$. We suppose that $\mathbf{y}$ is centered. Let $D^{red} = \left(\mathbf{y}^{red}, \mathbf{X}^{red}\right) \in \mathcal{M}_{m,p}(\mathbb{R})$ be the reduced matrix composed by only one occurrence of each different subject. Let $n = qm + r$ be the Euclidean division of n by m with $q=1$ and $r = n-m$. Then,

$$\begin{aligned}
\mathrm{Cov}\left(\mathbf{w}_1^b, \mathbf{w}_1^{red}\right) &= \mathrm{Corr}\left(\mathbf{w}_1^b, \mathbf{w}_1^{red}\right) \\
&= \mathrm{Corr}\left(\left(\mathbf{X}_1^{red}\right)^T \mathbf{y}^{red} + \left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}, \left(\mathbf{X}_1^{red}\right)^T \mathbf{y}^{red}\right) \\
&= \frac{\mathrm{Cov}\left(\left(\mathbf{X}_1^{red}\right)^T \mathbf{y}^{red} + \left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}, \left(\mathbf{X}_1^{red}\right)^T \mathbf{y}^{red}\right)}{\sqrt{\mathrm{Var}\left(\left(\mathbf{X}_1^{red}\right)^T \mathbf{y}^{red} + \left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}\right) \cdot \mathrm{Var}\left(\left(\mathbf{X}_1^{red}\right)^T \mathbf{y}^{red}\right)}}
\end{aligned}$$

Let us first develop the numerator.

$$\begin{aligned}
&\mathrm{Cov}\left(\left(\mathbf{X}_1^{red}\right)^T \mathbf{y}^{red} + \left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}, \left(\mathbf{X}_1^{red}\right)^T \mathbf{y}^{red}\right) \\
&\quad = \mathrm{Cov}\left(\left(\mathbf{X}_1^{red}\right)^T \mathbf{y}^{red}, \left(\mathbf{X}_1^{red}\right)^T \mathbf{y}^{red}\right) + \mathrm{Cov}\left(\left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}, \left(\mathbf{X}_1^{red}\right)^T \mathbf{y}^{red}\right) \\
&\quad = \mathrm{Var}\left(\left(\mathbf{X}_1^{red}\right)^T \mathbf{y}^{red}\right) \\
&\qquad + \mathrm{Cov}\left(\left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}, \left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red} + \left(\mathbf{X}_{1,m-r\geqslant l}^{red}\right)^T \mathbf{y}_{m-r\geqslant l}^{red}\right) \\
&\quad = \mathrm{Var}\left(\left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red} + \left(\mathbf{X}_{1,m-r\geqslant l}^{red}\right)^T \mathbf{y}_{m-r\geqslant l}^{red}\right) \\
&\qquad + \mathrm{Var}\left(\left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}\right) \\
&\qquad + \mathrm{Cov}\left(\left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}, \left(\mathbf{X}_{1,m-r\geqslant l}^{red}\right)^T \mathbf{y}_{m-r\geqslant l}^{red}\right) \\
&\quad = 2 \cdot \mathrm{Var}\left(\left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}\right) + \mathrm{Var}\left(\left(\mathbf{X}_{1,m-r\geqslant l}^{red}\right)^T \mathbf{y}_{m-r\geqslant l}^{red}\right) \\
&\qquad + 3 \cdot \mathrm{Cov}\left(\left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}, \left(\mathbf{X}_{1,m-r\geqslant l}^{red}\right)^T \mathbf{y}_{m-r\geqslant l}^{red}\right) \\
&\quad = 2a + b + 3\rho\sqrt{ab}
\end{aligned}$$

with: $a = \mathrm{Var}\left(\left(\mathbf{X}_{m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}\right)$

$b = \mathrm{Var}\left(\left(\mathbf{X}_{m-r\geqslant l}^{red}\right)^T \mathbf{y}_{m-r\geqslant l}^{red}\right)$

$\rho = \mathrm{Corr}\left(\left(\mathbf{X}_{m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}, \left(\mathbf{X}_{m-r\geqslant l}^{red}\right)^T \mathbf{y}_{m-r\geqslant l}^{red}\right)$

Then, let us develop the squared denominator:

$$\begin{aligned}
&\mathrm{Var}\left(\left(\mathbf{X}_1^{red}\right)^T \mathbf{y}^{red} + \left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}\right) \cdot \mathrm{Var}\left(\left(\mathbf{X}_1^{red}\right)^T \mathbf{y}^{red}\right) \\
&\quad = \mathrm{Var}^2\left(\left(\mathbf{X}_1^{red}\right)^T \mathbf{y}^{red}\right) + \mathrm{Var}\left(\left(\mathbf{X}_1^{red}\right)^T \mathbf{y}^{red}\right) \cdot \mathrm{Var}\left(\left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}\right) \\
&\qquad + 2\mathrm{Var}\left(\left(\mathbf{X}_1^{red}\right)^T \mathbf{y}^{red}\right) \cdot \mathrm{Cov}\left(\left(\mathbf{X}_1^{red}\right)^T \mathbf{y}^{red}, \left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}\right)
\end{aligned}$$

By replacing $\left(\mathbf{X}_1^{red}\right)^T \mathbf{y}^{red}$ by $\left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red} + \left(\mathbf{X}_{1,m-r\geqslant l}^{red}\right)^T \mathbf{y}_{m-r\geqslant l}^{red}$ and developing we obtain that:

$$
\begin{aligned}
&\mathrm{Var}\left(\left(\mathbf{X}_1^{red}\right)^T \mathbf{y}^{red} + \left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}\right) \cdot \mathrm{Var}\left(\left(\mathbf{X}_1^{red}\right)^T \mathbf{y}^{red}\right) \\
&\quad = 4\mathrm{Var}^2\left(\left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}\right) + \mathrm{Var}^2\left(\left(\mathbf{X}_{1,m-r\geqslant l}^{red}\right)^T \mathbf{y}_{m-r\geqslant l}^{red}\right) \\
&\qquad + 5\mathrm{Var}\left(\left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}\right) \mathrm{Var}\left(\left(\mathbf{X}_{1,m-r\geqslant l}^{red}\right)^T \mathbf{y}_{m-r\geqslant l}^{red}\right) \\
&\qquad\quad + 8\mathrm{Cov}^2\left(\left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}, \left(\mathbf{X}_{1,m-r\geqslant l}^{red}\right)^T \mathbf{y}_{m-r\geqslant l}^{red}\right) \\
&\qquad\quad + 2\mathrm{Cov}\left(\left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}, \left(\mathbf{X}_{1,m-r\geqslant l}^{red}\right)^T \mathbf{y}_{m-r\geqslant l}^{red}\right) \\
&\qquad\quad\quad \cdot \left[6\mathrm{Var}\left(\left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}\right) + 3\mathrm{Var}\left(\left(\mathbf{X}_{1,m-r\geqslant l}^{red}\right)^T \mathbf{y}_{m-r\geqslant l}^{red}\right)\right] \\
&\quad = \left[2\mathrm{Var}\left(\left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}\right) + \mathrm{Var}\left(\left(\mathbf{X}_{1,m-r\geqslant l}^{red}\right)^T \mathbf{y}_{m-r\geqslant l}^{red}\right)\right. \\
&\qquad \left. + 3\mathrm{Cov}\left(\left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}, \left(\mathbf{X}_{1,m-r\geqslant l}^{red}\right)^T \mathbf{y}_{m-r\geqslant l}^{red}\right)\right]^2 \\
&\qquad\quad + \mathrm{Var}\left(\left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}\right) \mathrm{Var}\left(\left(\mathbf{X}_{1,m-r\geqslant l}^{red}\right)^T \mathbf{y}_{m-r\geqslant l}^{red}\right) \\
&\qquad\quad\quad - \mathrm{Cov}^2\left(\left(\mathbf{X}_{1,m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}, \left(\mathbf{X}_{1,m-r\geqslant l}^{red}\right)^T \mathbf{y}_{m-r\geqslant l}^{red}\right) \\
&\quad = \left(2a + b + 3\rho\sqrt{ab}\right)^2 + \left(1-\rho^2\right) ab
\end{aligned}
$$

□

Let us denote:

$$
\begin{aligned}
\mathrm{f}: \mathbb{R}^{+*} \times \mathbb{R}^{+*} \times [-1,1] &\longrightarrow [-1,1] \\
(a,b,\rho) &\longmapsto \mathrm{Corr}\left(\mathbf{w}_1^b, \mathbf{w}_1^{red}\right)
\end{aligned}
$$

Then, we treat separately the cases $\rho \in \{-1,1\}$ and $\rho \in]-1,1[$. For $\rho \in \{-1,1\}$, we obtain the following results:

$$
\begin{aligned}
&f(a,b,1) = 1, \ \forall (a,b) \in \left(\mathbb{R}^{+*}\right)^2 \\
&f(a,b,-1) = \begin{cases} 1, & \forall a \in]0, b/4[\cup]b, +\infty[\\ 0, & \text{if } a \in \{b, b/4\} \\ -1, & \forall a \in]b/4, b[\end{cases}
\end{aligned}
$$

Concerning the second case, $\forall \rho \in]-1,1[$ f is differentiable on $\mathbb{R}^{+*} \times \mathbb{R}^{+*}$ and own a minimum for $a = b/2$. This minimum only depends on ρ and is equal to:

$$
g(\rho) = f(b/2, b, \rho) = \frac{2 + \frac{3\rho}{\sqrt{2}}}{\sqrt{\left(2 + \frac{3\rho}{\sqrt{2}}\right)^2 + \frac{1}{2}\left(1-\rho^2\right)}}
$$

A graphical representation of g is performed on Fig.G.1.

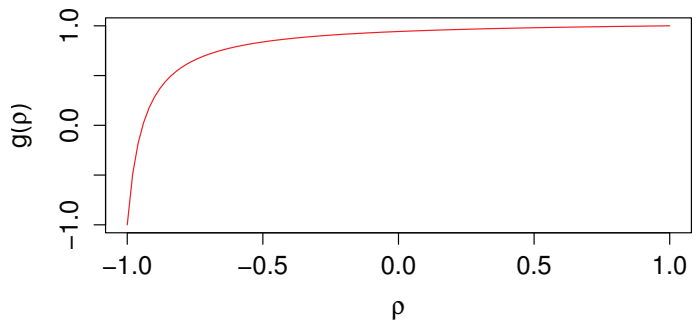


Figure G.1: Minimum of f for $\rho \in]-1,1[$.

Proposition G.0.1. *Let* $\mathbf{w}_1^* \in \mathbb{R}^p$ *so that* $\mathbf{w}_1^b = \mathbf{w}_1^{red} + \mathbf{w}_1^*$ *then:*

$$\|\mathbf{w}_1^*\| \in \left[0, \sqrt{2\left(1 - g\left(\rho\right)\right)}\right]$$

Corollary G.0.3. *Let* $\mathbf{v} = \mathbf{X}^b\mathbf{w}_1^*$ *so that* $\mathbf{t}_1^b = \left(\mathbf{t}_1^{red}, \mathbf{t}_{1,m-(r-1)\leqslant l}^{red}\right) + \mathbf{v}$*, then:*

$$\forall \rho \in [-1,1], \exists \zeta \in \left[0, \frac{\sqrt{2\left(1 - g\left(\rho\right)\right)}}{\alpha}\right], \|\mathbf{v}\| \leqslant \zeta \left\|\left(\mathbf{t}_1^{red}, \mathbf{t}_{1,m-(r-1)\leqslant l}^{red}\right)\right\|$$

with $\alpha \in]0,1]$ *such that* $\left\|\left(\mathbf{t}_1^{red}, \mathbf{t}_{1,m-(r-1)\leqslant l}^{red}\right)\right\| = \alpha \left\|\left\| \mathbf{X}_1^b \right\|\right\|$.

Proof. Let $\mathbf{w}_1^* \in \mathbb{R}^p$ so that $\mathbf{w}_1^b = \mathbf{w}_1^{red} + \mathbf{w}_1^*$. Then, we have that:

$$\left\|\mathbf{w}_1^*\right\|^2 = \left\|\mathbf{w}_1^b\right\|^2 + \left\|\mathbf{w}_1^{red}\right\|^2 - 2\left\|\mathbf{w}_1^b\right\| \cdot \left\|\mathbf{w}_1^{red}\right\| \cdot \text{Corr}\left(\mathbf{w}_1^b, \mathbf{w}_1^{red}\right) \tag{G.7}$$

Since $\left\|\mathbf{w}_1^b\right\| = \left\|\mathbf{w}_1^{red}\right\| = 1$, we obtain that:

$$\left\|\mathbf{w}_1^*\right\|^2 = 2 - 2\text{Corr}\left(\mathbf{w}_1^b, \mathbf{w}_1^{red}\right) = 2\left(1 - \text{Corr}\left(\mathbf{w}_1^b, \mathbf{w}_1^{red}\right)\right) \tag{G.8}$$

Since,

$$\text{Corr}\left(\mathbf{w}_1^b, \mathbf{w}_1^{red}\right) \geqslant g\left(\rho\right) = \frac{2 + \frac{3\rho}{\sqrt{2}}}{\sqrt{\left(2 + \frac{3\rho}{\sqrt{2}}\right)^2 + \frac{1}{2}\left(1 - \rho^2\right)}}$$

with $\rho = \text{Corr}\left(\left(\mathbf{X}_{m-r<l}^{red}\right)^T \mathbf{y}_{m-r<l}^{red}, \left(\mathbf{X}_{m-r\geqslant l}^{red}\right)^T \mathbf{y}_{m-r\geqslant l}^{red}\right)$, we obtain the following result:

$$\begin{aligned} &\left\|\mathbf{w}_1^*\right\|^2 \leqslant 2\left(1 - g\left(\rho\right)\right) \\ \Rightarrow &\left\|\mathbf{w}_1^*\right\| \in \left[0, \sqrt{2\left(1 - g\left(\rho\right)\right)}\right] \end{aligned} \tag{G.9}$$

Then, let us denote $\mathbf{v} = \mathbf{X}^b \mathbf{w}_1^*$ so that $\mathbf{t}_1^b = \left(\mathbf{t}_1^{red}, \mathbf{t}_{1,m-(r-1)\leqslant l}^{red}\right) + \mathbf{v}$. We obtain that:

$$\begin{aligned} \|\mathbf{v}\| &= \left\|\mathbf{X}^b \mathbf{w}_1^*\right\| \\ &\leqslant |||\mathbf{X}_1^b||| \cdot \|\mathbf{w}_1^*\| \\ &\leqslant \sqrt{2\left(1 - g\left(\rho\right)\right)} \cdot |||\mathbf{X}_1^b||| \end{aligned} \tag{G.10}$$

and

$$\begin{aligned} \left\|\left(\mathbf{t}_1^{red}, \mathbf{t}_{1,m-(r-1)\leqslant l}^{red}\right)\right\| &= \left\|\mathbf{X}^b \mathbf{w}_1^{red}\right\| \\ &\leqslant |||\mathbf{X}_1^b||| \cdot \left\|\mathbf{w}_1^{red}\right\| \\ &= |||\mathbf{X}_1^b||| \end{aligned} \tag{G.11}$$

See to G.11, we conclude that $\exists 0 < \alpha \leqslant 1$ so that $\left\|\left(\mathbf{t}_1^{red}, \mathbf{t}_{1,m-(r-1)\leqslant l}^{red}\right)\right\| = \alpha \, ||| \, \mathbf{X}_1^b|||$. Consequently, we obtain that:

$$\frac{\|\mathbf{v}\|}{\left\|\left(\mathbf{t}_1^{red}, \mathbf{t}_{1,m-(r-1)\leqslant l}^{red}\right)\right\|} \leqslant \frac{\sqrt{2\left(1 - g\left(\rho\right)\right)}}{\alpha} \tag{G.12}$$

□

Corollary G.0.4. *Let $\zeta \geqslant 0$ so that $\|\mathbf{v}\| \leqslant \zeta \left\|\left(\mathbf{t}_1^{red}, \mathbf{t}_{1,m-(r-1)\leqslant l}^{red}\right)\right\|$. Then,*

$$\frac{1}{(1+\zeta)^2}\left(\frac{\left(\mathbf{t}_1^{red,c}\right)^T \mathbf{y}^{red,c}}{\left(\mathbf{t}_1^{red,c}\right)^T \mathbf{t}_1^{red,c}} - \zeta \frac{\left\|\mathbf{y}^{red,c}\right\|}{\left\|\mathbf{t}_1^{red,c}\right\|}\right) \leqslant \hat{c}_1^b \leqslant \frac{1}{(1-\zeta)^2}\left(\frac{\left(\mathbf{t}_1^{red,c}\right)^T \mathbf{y}^{red,c}}{\left(\mathbf{t}_1^{red,c}\right)^T \mathbf{t}_1^{red,c}} + \zeta \frac{\left\|\mathbf{y}^{red,c}\right\|}{\left\|\mathbf{t}_1^{red,c}\right\|}\right)$$

with $\mathbf{t}_1^{red,c} = \left(\mathbf{t}_1^{red}, \mathbf{t}_{1,m-(r-1)\leqslant l}^{red}\right)$ *and* $\mathbf{y}^{red,c} = \left(\mathbf{y}^{red}, \mathbf{y}_{m-(r-1)\leqslant l}^{red}\right)$.

Proof. Let us denote $\mathbf{t}_1^{red,c} = \left(\mathbf{t}_1^{red}, \mathbf{t}_{1,m-(r-1)\leqslant l}^{red}\right)$ and $\mathbf{y}^{red,c} = \left(\mathbf{y}^{red}, \mathbf{y}_{m-(r-1)\leqslant l}^{red}\right) = \mathbf{y}^b$. Let $\zeta \geqslant 0$ so that $\|\mathbf{v}\| \leqslant \zeta \left\|\mathbf{t}_1^{red,c}\right\|$. Then,

$$\begin{aligned} \hat{c}_1^b &= \frac{\left(\mathbf{t}_1^b\right)^T \mathbf{y}^b}{\left(\mathbf{t}_1^b\right)^T \mathbf{t}_1^b} \\ &= \frac{\left(\mathbf{t}_1^{red,c} + \mathbf{v}\right)^T \mathbf{y}^{red,c}}{\left(\mathbf{t}_1^{red,c} + \mathbf{v}\right)^T \left(\mathbf{t}_1^{red,c} + \mathbf{v}\right)} \\ &= \frac{\left\langle \mathbf{t}_1^{red,c}, \mathbf{y}^{red,c} \right\rangle + \left\langle \mathbf{v}, \mathbf{y}^{red,c} \right\rangle}{\left\|\mathbf{t}_1^{red,c}\right\|^2 + 2\left\langle \mathbf{t}_1^{red,c}, \mathbf{v} \right\rangle + \|\mathbf{v}\|^2} \end{aligned}$$

Due to the Cauchy-Schwartz inequality, we have that $|\langle a,b\rangle| \leqslant \|a\| \cdot \|b\| \Leftrightarrow -\|a\| \cdot \|b\| \leqslant \langle a,b\rangle \leqslant \|a\| \cdot \|b\|$. We so obtain that:

$$\frac{\left\langle \mathbf{t}_1^{red,c}, \mathbf{y}^{red,c} \right\rangle - \|\mathbf{v}\| \cdot \left\|\mathbf{y}^{red,c}\right\|}{\left\|\mathbf{t}_1^{red,c}\right\|^2 + 2\left\|\mathbf{t}_1^{red,c}\right\| \cdot \|\mathbf{v}\| + \|\mathbf{v}\|^2} \leqslant \hat{c}_1^b \leqslant \frac{\left\langle \mathbf{t}_1^{red,c}, \mathbf{y}^{red,c} \right\rangle + \|\mathbf{v}\| \cdot \left\|\mathbf{y}^{red,c}\right\|}{\left\|\mathbf{t}_1^{red,c}\right\|^2 - 2\left\|\mathbf{t}_1^{red,c}\right\| \cdot \|\mathbf{v}\| + \|\mathbf{v}\|^2} \tag{G.13}$$

Since $\|\mathbf{v}\| \leqslant \zeta \left\|\mathbf{t}_1^{red,c}\right\|$, we obtain that:

$$\begin{aligned}
\frac{\left\langle \mathbf{t}_1^{red,c}, \mathbf{y}^{red,c} \right\rangle - \|\mathbf{v}\| \cdot \left\|\mathbf{y}^{red,c}\right\|}{\left\|\mathbf{t}_1^{red,c}\right\|^2 + 2\left\|\mathbf{t}_1^{red,c}\right\| \cdot \|\mathbf{v}\| + \|\mathbf{v}\|^2} &\geqslant \frac{\left\langle \mathbf{t}_1^{red,c}, \mathbf{y}^{red,c} \right\rangle - \zeta \left\|\mathbf{t}_1^{red,c}\right\| \cdot \left\|\mathbf{y}^{red,c}\right\|}{\left\|\mathbf{t}_1^{red,c}\right\|^2 + 2\left\|\mathbf{t}_1^{red,c}\right\| \cdot \zeta \left\|\mathbf{t}_1^{red,c}\right\| + \zeta^2 \left\|\mathbf{t}_1^{red,c}\right\|^2} \\
&= \frac{\left\langle \mathbf{t}_1^{red,c}, \mathbf{y}^{red,c} \right\rangle - \zeta \left\|\mathbf{t}_1^{red,c}\right\| \cdot \left\|\mathbf{y}^{red,c}\right\|}{(1+\zeta)^2 \left\|\mathbf{t}_1^{red,c}\right\|^2} \\
&= \frac{1}{(1+\zeta)^2} \cdot \left(\frac{\left(\mathbf{t}_1^{red,c}\right)^T \mathbf{y}^{red,c}}{\left(\mathbf{t}_1^{red,c}\right)^T \mathbf{t}_1^{red,c}} - \zeta \frac{\left\|\mathbf{y}^{red,c}\right\|}{\left\|\mathbf{t}_1^{red,c}\right\|} \right)
\end{aligned}$$

and

$$\begin{aligned}
\frac{\left\langle \mathbf{t}_1^{red,c}, \mathbf{y}^{red,c} \right\rangle + \|\mathbf{v}\| \cdot \left\|\mathbf{y}^{red,c}\right\|}{\left\|\mathbf{t}_1^{red,c}\right\|^2 - 2\left\|\mathbf{t}_1^{red,c}\right\| \cdot \|\mathbf{v}\| + \|\mathbf{v}\|^2} &\leqslant \frac{\left\langle \mathbf{t}_1^{red,c}, \mathbf{y}^{red,c} \right\rangle + \zeta \left\|\mathbf{t}_1^{red,c}\right\| \cdot \left\|\mathbf{y}^{red,c}\right\|}{\left\|\mathbf{t}_1^{red,c}\right\|^2 - 2\left\|\mathbf{t}_1^{red,c}\right\| \cdot \zeta \left\|\mathbf{t}_1^{red,c}\right\| + \zeta^2 \left\|\mathbf{t}_1^{red,c}\right\|^2} \\
&= \frac{\left\langle \mathbf{t}_1^{red,c}, \mathbf{y}^{red,c} \right\rangle + \zeta \left\|\mathbf{t}_1^{red,c}\right\| \cdot \left\|\mathbf{y}^{red,c}\right\|}{(1-\zeta)^2 \left\|\mathbf{t}_1^{red,c}\right\|^2} \\
&= \frac{1}{(1-\zeta)^2} \cdot \left(\frac{\left(\mathbf{t}_1^{red,c}\right)^T \mathbf{y}^{red,c}}{\left(\mathbf{t}_1^{red,c}\right)^T \mathbf{t}_1^{red,c}} + \zeta \frac{\left\|\mathbf{y}^{red,c}\right\|}{\left\|\mathbf{t}_1^{red,c}\right\|} \right)
\end{aligned}$$

By applying these inequalities to G.13, we finally obtain the following result:

$$\frac{1}{(1+\zeta)^2} \left(\frac{\left(\mathbf{t}_1^{red,c}\right)^T \mathbf{y}^{red,c}}{\left(\mathbf{t}_1^{red,c}\right)^T \mathbf{t}_1^{red,c}} - \zeta \frac{\left\|\mathbf{y}^{red,c}\right\|}{\left\|\mathbf{t}_1^{red,c}\right\|} \right) \leqslant \hat{c}_1^b \leqslant \frac{1}{(1-\zeta)^2} \left(\frac{\left(\mathbf{t}_1^{red,c}\right)^T \mathbf{y}^{red,c}}{\left(\mathbf{t}_1^{red,c}\right)^T \mathbf{t}_1^{red,c}} + \zeta \frac{\left\|\mathbf{y}^{red,c}\right\|}{\left\|\mathbf{t}_1^{red,c}\right\|} \right)$$

□

With respect to our simulation plan, we can reasonably assume that $\exists \varepsilon > 0$ so that:

$$\begin{aligned}
\hat{c}_1^{red} - \varepsilon &= \frac{\left(\mathbf{t}_1^{red}\right)^T \mathbf{y}^{red}}{\left(\mathbf{t}_1^{red}\right)^T \mathbf{t}_1^{red}} - \varepsilon \\
&\leqslant \frac{\left(\mathbf{t}_1^{red}\right)^T \mathbf{y}^{red} + \left(\mathbf{t}_{1,m-(r-1)\leqslant l}^{red}\right)^T \mathbf{y}_{m-(r-1)\leqslant l}^{red}}{\left(\mathbf{t}_1^{red}\right)^T \mathbf{t}_1^{red} + \left(\mathbf{t}_{1,m-(r-1)\leqslant l}^{red}\right)^T \mathbf{t}_{1,m-(r-1)\leqslant l}^{red}} \\
&= \frac{\left(\mathbf{t}_1^{red,c}\right)^T \mathbf{y}^{red,c}}{\left(\mathbf{t}_1^{red,c}\right)^T \mathbf{t}_1^{red,c}} \\
&\leqslant \frac{\left(\mathbf{t}_1^{red}\right)^T \mathbf{y}^{red}}{\left(\mathbf{t}_1^{red}\right)^T \mathbf{t}_1^{red}} + \varepsilon \\
&= \hat{c}_1^{red} + \varepsilon
\end{aligned}$$

By taking into account this property specific to our simulation process, we obtain the following result:

$$\frac{1}{(1+\zeta)^2}\left(\hat{c}_1^{red} - \varepsilon - \zeta\frac{\left\|\mathbf{y}^{red,c}\right\|}{\left\|\mathbf{t}_1^{red,c}\right\|}\right) \leqslant \hat{c}_1^b \leqslant \frac{1}{(1-\zeta)^2}\left(\hat{c}_1^{red} + \varepsilon + \zeta\frac{\left\|\mathbf{y}^{red,c}\right\|}{\left\|\mathbf{t}_1^{red,c}\right\|}\right)$$

Finally, we have that:

$$\lim_{\rho\to 1}\frac{\sqrt{2\left(1-g\left(\rho\right)\right)}}{\alpha} = 0 \Rightarrow \zeta \underset{\rho\to 1}{\longrightarrow} 0 \Rightarrow \hat{c}_1^{red} - \varepsilon \leqslant \hat{c}_1^b \leqslant \hat{c}_1^{red} + \varepsilon$$

Since our simulation process leads to high values of ρ, we can reasonably assume that $\hat{c}_1^b \simeq \hat{c}_1^{red}$. By induction due to the PLS regression process, we obtain similarly results concerning $\hat{c}_k^b$ and $\hat{c}_k^{red}$ for $k \geqslant 2$, as well as for the relative equality between $\mathbf{t}_k^b$ and $\mathbf{t}_k^{red,c}$.

Due to these approximate equalities, we obtain that $RSS_k^b \simeq RSS_k^{red} + \sum_{j=m-(r-1)}^{m}\left(y_j^{red} - \hat{y}_{j,k}^{red}\right)^2$

so that we can reasonably assume that:

$$
\begin{aligned}
\hat{\mathrm{Var}}\left(\hat{c}_k^b\right) &= \frac{RSS_k^b}{\left(n - DoF_k^b\right)\sum_{i=1}^{n}\left(t_{i,k}^b\right)^2} \\
&\simeq \frac{RSS_k^{red} + \sum_{j=m-(r-1)}^{m}\left(y_j^{red} - \hat{y}_{j,k}^{red}\right)^2}{\left((m+r) - DoF_k^b\right)\sum_{i=1}^{n}\left(t_{i,k}^{red,c}\right)^2} \\
&< \frac{RSS_k^{red} + \sum_{j=m-(r-1)}^{m}\left(y_j^{red} - \hat{y}_{j,k}^{red}\right)^2}{\left(m - DoF_k^{red}\right).\left(\sum_{i=1}^{m}\left(t_{i,k}^{red}\right)^2 + \sum_{j=m-(r-1)}^{m}\left(t_{j,k}^{red}\right)^2\right)} \\
&\simeq \frac{RSS_k^{red}}{\left(m - DoF_k^{red}\right)\sum_{i=1}^{m}\left(t_{i,k}^{red}\right)^2} \\
&= \hat{\mathrm{Var}}\left(\hat{c}_k^{red}\right)
\end{aligned}
$$

These results globally imply that $K_{D^b} \geqslant K_{D^{red}}$. Finally, by assuming that the reduced dataset contains sufficiently information to maintain the global common space dimensioning, *i.e.* $K_{D^{red}} = K_{D^{ori}}$, we obtain that $K_{D^b} \geqslant K_{D^{ori}}$.

We performed simulations to verify this property by extracting 50 times the number of components with the BootYT stopping criterion for $m \in [\![3, n]\!]$. We performed a graphical representation of their means and linked confidence intervals depending on the value of m in Fig.G.2. We also verify this property on the 1000 bootstrap samples created during the bootstrap process for predictor selection, applied on the simulated dataset with $sd(\epsilon) = 5$ on which we obtain $K_{D^{ori}} = 9$, by extracting for each of them the numbers of components. We graphically report these results in Fig.G.2.

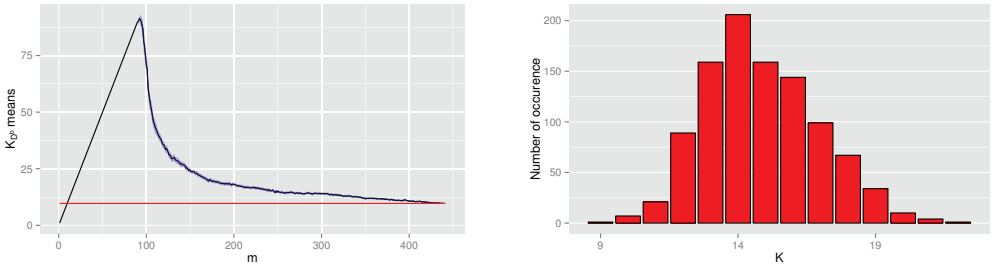


Figure G.2: Left: Evolution of extracted numbers of components depending on the value of m. Right: Distribution of extracted numbers of components on 1000 bootstrap samples, $K_{D^{ori}} = 9$.

For both experiences, the theoretical results are verified since the obtained numbers of components are never lower than the original one.

Annexe H

Résumé des corrélations entre prédicteurs pour le jeu de données présenté dans la Section 7.3.2

The matrix of predictors correlations and the four selected variables are displayed by using a heatmap in the following Fig.H.1.

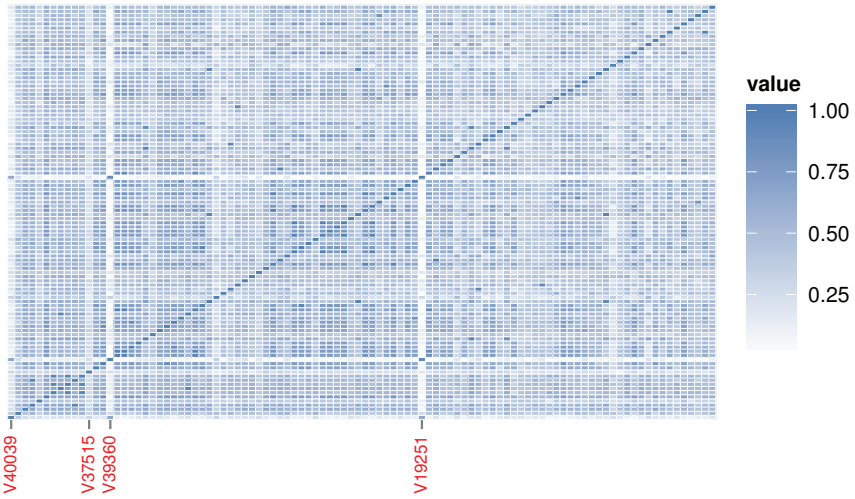


Figure H.1: Predictors absolute correlation heatmap.

Table H.1: Selected predictors correlations.

	$\mathbf{x}_1$	$\mathbf{x}_{12}$	$\mathbf{x}_{15}$	$\mathbf{x}_{59}$
$\mathbf{x}_1$	1	0.312	0.672	0.666
$\mathbf{x}_{12}$	0.312	1	0.256	0.203
$\mathbf{x}_{15}$	0.672	0.256	1	0.926
$\mathbf{x}_{59}$	0.666	0.203	0.926	1

Annexe I

Nombres de composantes obtenus par les différents critères étudiés sur les jeux de données simulés dans la Section 7.3.2

The determination of the number of components K_{opt} using the Q^2, the corrected BIC and the BootYT criteria is performed. Concerning the Q^2 and BootYT criteria, one hundred determinations are computed and results with the highest occurrence rate are selected. These results are displayed in Fig.I.1

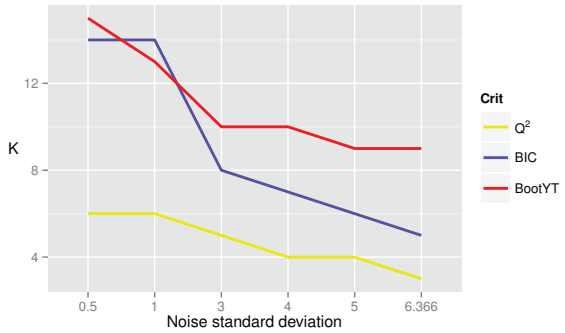


Figure I.1: Number of components depending on the used criterion.

These results globally match with a conclusion obtained by Magnanensi *et al.* (2015), in that the Q^2 returns lower numbers of components than the two others criteria. In almost all cases, it represents an underestimating issue. These numbers of components are then used in the original bootstrap-based approach introduced and described by Lazraq *et al.* (2003).

Annexe J

Poster présenté à la conférence internationale CMStatistics

Nous fournissons ici le poster présenté dans le cadre de la 8$^{\text{th}}$ *International Conference of the ERCIM WG on Computational and Methodological Statistics* qui s'est déroulée à Londres du 12 au 14 Décembre 2015.

New Developments of Sparse PLS regressions

Jérémy Magnanensi[1,2], Myriam Maumy-Bertrand[1], Nicolas Meyer[2], Frédéric Bertrand[1]
[1] Institut de Recherche Mathématique Avancée, UMR 7501, LabEx IRMIA
[2] Laboratoire de Biostatistique et Informatique Médicale, Faculté de Médecine, EA3430
Université de Strasbourg et CNRS, France

Context and aims

Methods based on the so-called Partial Least Squares (PLS) regression, which recently gained much attention in the analysis of high-dimensional genomic datasets, were developed since the early 2000s to perform variable selection. The Sparse PLS (SPLS) introduced by Chun and Keleş (2010) has become a benchmark. However, most of the developed techniques rely on some tuning parameters that are commonly determined by cross-validation (CV) based methods, which raise important stability issues Boulesteix (2014). Bootstrap-based techniques are known to be more stable than CV-based ones and represent an interesting alternative to improve predictors selection.

Thus, we developed a new dynamic bootstrapping pairs-based technique for variable selection, relevant for both PLS and its incorporation into generalized linear models (GPLS) (Bastien *et al.*, 2005) frameworks, characterized by successive establishments of dynamic dimensional working subspaces for each bootstrap sample. A recent robust bootstrap-based stopping criterion for PLS components construction (BootYT), introduced by Magnanensi *et al.* (2015) and characterized by a high level of stability, was used to successively achieved the determinations of this tuning parameter, leading to the following notation BootYTdyn. We also succeed in adapting the BootYT criterion for the determination of an unique optimal number of components related to each preset value of the sparsity parameter η in both the SPLS and its extension to classification problems (SGPLS) (Chung and Keles, 2010) frameworks, improving the stability in selecting the tunings parameters. We refer to these adapted implementations as SPLS BootYT and SGPLS BootYT while the original ones are noted SPLS CV and SGPLS CV.

Dynamic Bootstrap-based algorithm

1. Let D^{ori} be the original dataset and R the total number of bootstrap samples $D_r^b,\ r \in [\![1, R]\!]$.
2. $\forall r \in [\![1, R]\!]$:
   - Extract the number of components that is needed for D_r^b using the BootYT criterion.
   - Compute the estimations $\hat{\beta}_j^r,\ \forall j \in [\![1, p]\!]$ by fitting the relevant PLS or GPLS model.
3. $\forall j \in [\![1, p]\!]$, construct a $(100 \times (1-\alpha))\%$ bilateral BC_a CI, noted:
$$\mathrm{CI}_j = [\mathrm{CI}_{j,1}, \mathrm{CI}_{j,2}].$$
4. $\forall j \in [\![1, p]\!]$, **If** $0 \notin \mathrm{CI}_j$ **then** retain $\mathbf{x_j}$ **else** delete $\mathbf{x_j}$.
5. Obtain the reduced model $\mathcal{M}_{\mathrm{sel}}$ by only integrating the significant predictors into and extracting the number of components K_{sel} determined by BootYT criterion.

Note that in this work, we fixed $\alpha = 0.05$ and $R = 1000$ bootstrap samples.

SPLS BootYT algorithm

1. Let $\{\eta_1, \ldots, \eta_s\}$ be the set of preset values for the sparsity parameter and $\{k_1, \ldots, k_s\} = \{1, \ldots, 1\}$ the set of initial number of components for each η_i. Let $i = 1$.
2. Let $\hat{c}_j^{\eta_i},\ j \in [\![1, k_i]\!]$ be the estimated regression coefficients of $\mathbf{y}$ on $\mathbf{T}_{k_i} = (\mathbf{t}_1, \ldots, \mathbf{t}_{k_i}) \in \mathcal{M}_{n,k_i}(\mathbb{R})$. Obtain k_i BC_a CI for $\hat{c}_j^{\eta_i},\ j \in [\![1, k_i]\!]$, using the BootYT criterion, noted:
$$\mathrm{CI}_j^{k_i} = \left[\mathrm{CI}_{j,1}^{k_i}, \mathrm{CI}_{j,2}^{k_i}\right].$$
3. **If** $\exists j \in [\![1, k_i]\!] \,\big|\, 0 \in \mathrm{CI}_j^{k_i}$ **then** $K_{opt}^{\eta_i} = k_i - 1$ **else** $\{k_i = k_i + 1$ and return to step 2$\}$.
4. **While** $i \neq s$ **then** $\{i = i + 1$ and return to step 2$\}$.
5. Return the set of extracted numbers of components $\left\{K_{opt}^{\eta_1}, \ldots, K_{opt}^{\eta_s}\right\}$ related to each $\eta_i,\ 1 \leqslant i \leqslant s$.
6. Return the couple $\left(\eta_{opt}, K_{opt}^{\eta_{opt}}\right)$ having the lowest CV-based MSE.

Dataset and Simulations plan

A microarray gene expression dataset splited into a 443 samples discovery set (39582trai) and a 123 samples test set (39582test), well balanced for the main anatomoclinical characteristics, was used. 100 predictors were first selected by using a Bayesian pre-processing technique based on the binary response vector (original localization of the colon tumors), representing our dataset for comparison in GPLS framework. Two additional independent datasets (18088, n=53; 14333, n=243) were also used to perform a reliable comparison of the prediction ability. Concerning PLS case, response vectors have been simulated as follows:

$$\mathbf{y} = \mathbf{X}_{sel}\beta + \epsilon \tag{1}$$

with $\beta = (3.559, 2.071, 1.440, 1.770)^T$, $\mathbb{E}(\epsilon) = 0$, $\mathrm{Var}(\epsilon) = \sigma^2 I_n$ with $\sigma \in \{0.5, 1, 3, 4, 5, 6.366\}$ and $\mathbf{X}_{sel} = (\mathbf{x}_1, \mathbf{x}_{12}, \mathbf{x}_{15}, \mathbf{x}_{59}) \in \mathcal{M}_{n,4}(\mathbb{R})$ the matrix of the four selected positively correlated covariates. Each determination was performed one hundred times. Results linked to the highest occurrence rates were chosen for methods comparison.

PLS results

Table 1: Means of accuracy values over the 100 determinations. Highest values in red.

	SPLS CV	SPLS BootYT	BootYTdyn
$\mathrm{sd}(\epsilon) = 0.5$	0.9914	0.9960	0.9718
$\mathrm{sd}(\epsilon) = 1$	0.9915	1.0000	0.9781
$\mathrm{sd}(\epsilon) = 3$	0.9821	0.9799	0.9837
$\mathrm{sd}(\epsilon) = 4$	0.9771	0.9799	0.9928
$\mathrm{sd}(\epsilon) = 5$	0.9790	0.9841	0.9876
$\mathrm{sd}(\epsilon) = 6.366$	0.9745	0.9714	0.9820

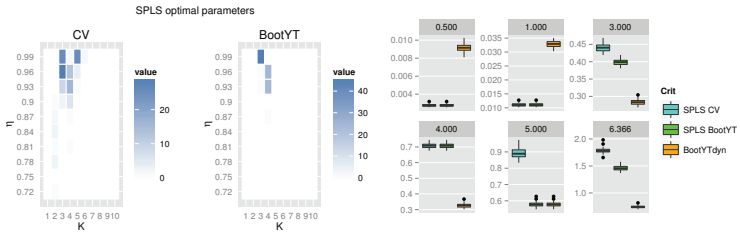


Figure 1: Left, Illustration of stability improvement: Repartition of 100 determinations of (η_{opt}, K_{opt}) obtained on the simulated dataset with $\mathrm{sd}(\epsilon) = 5$; Right, Predictive ability: 100 computations of 10-fold CV-based PMSE of the 3 retained models *i.e.* with highest occurrence rate over the 100 determinations.

GPLS results

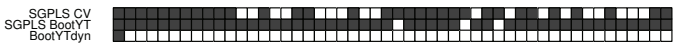


Figure 2: Summary of selected variables.

Table 2: MSE for the selected models (Number of misclassified values). Lowest values in red.

	39582trai	39582test	18088	14333	whole test data
SGPLS CV	0.1372 (48)	0.1374 (19)	0.1743 (10)	0.1520 (45)	0.1506 (74)
SGPLS BootYT	0.1295 (40)	0.1328 (15)	0.1848 (12)	0.1601 (46)	0.1553 (83)
BootYTdyn	0.1507 (92)	0.1446 (25)	0.1695 (13)	0.1325 (35)	0.1407 (73)

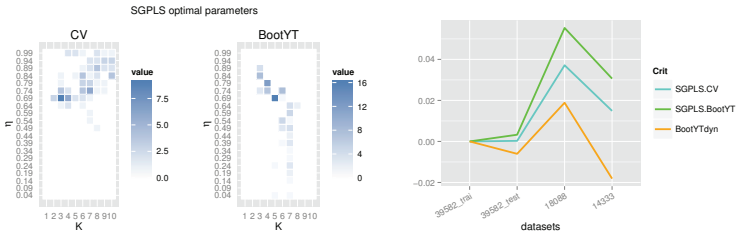


Figure 3: Left: Repartition of the 100 determinations of (η_{opt}, K_{opt}); Right, Illustration of over-fitting issue for SGPLS approaches: Differences between the MSE obtained on 39582trai and the PMSE obtained on the different test datasets.

References

Bastien, P., Vinzi, V. E., and Tenenhaus, M. (2005). PLS generalised linear regression. *Computational Statistics & Data Analysis*, **48**(1), 17–46.

Boulesteix, A.-L. (2014). Accuracy estimation for PLS and related methods via resampling-based procedures. In *PLS'14 Book of Abstracts*, pages 13–14.

Chun, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**(1), 3–25.

Chung, D. and Keles, S. (2010). Sparse partial least squares classification for high dimensional data. *Statistical Applications in Genetics and Molecular Biology*, **9**(1). Article 17.

Magnanensi, J., Bertrand, F., Maumy, M., and Meyer, N. (2015). A new Universal Resample Stable Bootstrap-based Stopping Criterion in PLS Components Construction. *arXiv preprint arXiv:1507.01404*.

Acknowledgments

UNIVERSITÉ DE STRASBOURG
IRMA
EA3430
LABEX IRMIA
ERCIM WG CMStatistics

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, **19**(6), 716–723.

Albert, A. and Anderson, J. (1984). On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, **71**(1), 1–10.

Allen, D. M. (1971). *The prediction sum of squares as a criterion for selecting predictor variables*. Technical report 23 - Department of Statistics. University of Kentucky.

Amato, S. and Vinzi, V. E. (2003). Bootstrap-based qÌĆ kh 2 for the selection of components and variables in pls regression. *Chemometrics and intelligent laboratory systems*, **68**(1), 5–16.

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika*, **59**(1), 19–35.

Anderson, J. A. (1974). Diagnosis by logistic discriminant function: further practical problems and results. *Applied Statistics*, **23**(3), 397–404.

Andrieux, J. (2008). Puces à ADN (CGH-array): application pour le diagnostic de déséquilibres cytogénétiques cryptiques. *Pathologie Biologie*, **56**(6), 368–374.

Astler, V. B. and Coller, F. A. (1954). The prognostic significance of direct extension of carcinoma of the colon and rectum. *Annals of Surgery*, **139**(6), 846.

Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nature Genetics*, **37**(4), 382–390.

Bastien, P., Vinzi, V. E., and Tenenhaus, M. (2005). PLS generalised linear regression. *Computational Statistics & Data Analysis*, **48**(1), 17–46.

Bertrand, F., Magnanensi, J., Maumy-Bertrand, M., and Meyer, N. (2014). *Partial Least Squares Regression for Generalized Linear Models*. Book of abstracts, User2014!, Los Angeles, page 150.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics*, **9**(6), 1196–1217.

Blazere, M. (2015). *Inférence en grande dimension pour des modèles structurels. Modèles linéaires généralisés parcimonieux, méthode PLS et polynômes orthogonaux et détection de communautés dans des graphes.* Ph.D. thesis, Toulouse, INSA.

Blazere, M., Gamboa, F., and Loubes, J.-M. (2014). A unified framework for the study of the PLS estimator's properties. *arXiv preprint arXiv:1411.0229*.

Boley, D. L. (1994). Krylov space methods on state-space control models. *Circuits, Systems and Signal Processing*, **13**(6), 733–758.

Boulesteix, A.-L. (2004). PLS dimension reduction for classification with microarray data. *Statistical Applications in Genetics and Molecular Biology*, **3**(1), 1–30.

Boulesteix, A.-L. (2014). Accuracy estimation for PLS and related methods via resampling-based procedures. In *PLS'14 Book of Abstracts*, pages 13–14.

Boulesteix, A.-L. and Strimmer, K. (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, **8**(1), 32–44.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press, 2000 N.W. Corporate Blvd, Boca Raton, Florida 33431, US.

Burris, J., Cook-Deegan, R., and Alberts, B. (1998). The human genome project after a decade: policy issues. *Nature Genetics*, **20**(4), 333–335.

Chang, T.-W. (1983). Binding of cells to matrixes of distinct antibodies coated on solid surface. *Journal of Immunological Methods*, **65**(1), 217–223.

Chun, H. and Keles, S. (2007). Sparse partial least squares regression with an application to genome scale transcription factor analysis. *Department of Statistics, University of Wisconsin, Madison*.

Chun, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**(1), 3–25.

Chung, D. and Keles, S. (2010). Sparse partial least squares classification for high dimensional data. *Statistical Applications in Genetics and Molecular Biology*, **9**(1). Article 17, DOI:10.2202/1544-6115.1492.

Collins, F. and Galas, D. (1993). A new five-year plan for the us human genome project. *Science*, **262**(5130), 43–46.

Davidson, R. and Flachaire, E. (2008). The wild bootstrap, tamed at last. *Journal of Econometrics*, **146**(1), 162–169.

De Jong, S. (1993a). PLS fits closer than PCR. *Journal of Chemometrics*, **7**(6), 551–557.

De Jong, S. (1993b). SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, **18**(3), 251–263.

De Jong, S. (1995). PLS shrinks. *Journal of Chemometrics*, **9**(4), 323–326.

Denham, M. C. (2000). Choosing the number of factors in partial least squares regression: estimating and minimizing the mean squared error of prediction. *Journal of Chemometrics*, **14**(4), 351–361.

Ding, B. and Gentleman, R. (2005). Classification Using Generalized Partial Least Squares. *Journal of Computational and Graphical Statistics*, **14**(2), 280–298.

Dray, S., Pettorelli, N., and Chessel, D. (2003). Multivariate analysis of incomplete mapped data. *Transactions in GIS*, **7**(3), 411–422.

Durif, G., Picard, F., and Lambert-Lacroix, S. (2015). Adaptive Sparse PLS for Logistic Regression. *arXiv preprint arXiv:1502.05933*.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, **7**(1), 1–26.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, **9**(2), 139–158.

Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika*, **72**(1), 45–58.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, **82**(397), 171–185.

Efron, B. (2004). The estimation of prediction error. *Journal of the American Statistical Association*, **99**(467), 619–632.

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*, volume 57. Chapman & Hall/CRC, 2000 N.W. Corporate Blvd, Boca Raton, Florida 33431, US.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., *et al.* (2004). Least angle regression. *The Annals of Statistics*, **32**(2), 407–499.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, **80**(1), 27–38.

Forina, M., Lanteri, S., Oliveros, M. C., and Millan, C. P. (2004). Selection of useful predictors in multivariate calibration. *Analytical and Bioanalytical Chemistry*, **380**(3), 397–418.

Fort, G. and Lambert-Lacroix, S. (2005). Classification using partial least squares with penalized logistic regression. *Bioinformatics*, **21**(7), 1104–1111.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**(2), 109–135.

Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics*, **9**(6), 1218–1228.

Freund, R. W., Golub, G. H., and Nachtigal, N. M. (1992). Iterative solution of linear systems. *Acta Numerica*, **1**, 57–100.

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

Gaussier, É. and Yvon, F. (2011). *Modèles statistiques pour l'accès à l'information textuelle*. Lavoisier, 14, Rue de Provigny, 94230 Cachan, France.

Gibson, G. and Muse, S. V. (2004). *Précis de génomique*. De Boeck Supérieur, Rue des Minimes 39, B-1000 Bruxelles, Belgique.

Gómez-Carracedo, M., Andrade, J., Rutledge, D., and Faber, N. (2007). Selecting the optimum number of partial least squares components for the calibration of attenuated total reflectance-mid-infrared spectra of undesigned kerosene samples. *Analytica Chimica Acta*, **585**(2), 253–265.

Gourvénec, S., Pierna, J. F., Massart, D., and Rutledge, D. (2003). An evaluation of the PoLiSh smoothed regression and the Monte Carlo cross-validation for the determination of the complexity of a PLS model. *Chemometrics and Intelligent Laboratory Systems*, **68**(1), 41–51.

Goutis, C. *et al.* (1996). Partial least squares algorithm yields shrinkage estimators. *The Annals of Statistics*, **24**(2), 816–824.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, **46**(2), 149–192.

Green, R. L. and Kalivas, J. H. (2002). Graphical diagnostics for regression model determinations with consideration of the bias/variance trade-off. *Chemometrics and Intelligent Laboratory Systems*, **60**(1), 173–188.

Gröne, J., Lenze, D., Jurinovic, V., Hummel, M., Seidel, H., Leder, G., Beckmann, G., Sommer, A., Grützmann, R., Pilarsky, C., *et al.* (2011). Molecular profiles and clinical outcome of stage UICC II colon cancer patients. *International Journal of Colorectal Disease*, **26**(7), 847–858.

Haaland, D. M. and Thomas, E. V. (1988). Partial least-squares methods for spectral analyses. 1. relation to other quantitative calibration methods and the extraction of qualitative information. *Analytical Chemistry*, **60**(11), 1193–1202.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, 175 Fifth Avenue, New York, NY 10010, USA.

Hastie, T., Tibshirani, R., and Friedman, J. J. H. (2009). *The elements of statistical learning, second edition*, volume 1. Springer New York, 233 Spring Street, New York, NY, 10013, USA.

Helland, I. S. (1988). On the structure of partial least squares regression. *Communications in statistics-Simulation and Computation*, **17**(2), 581–607.

Helland, I. S. (1990). Partial least squares regression and statistical models. *Scandinavian Journal of Statistics*, **17**(2), 97–114.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**(1), 55–67.

Holcomb, T. R., Hjalmarsson, H., Morari, M., and Tyler, M. L. (1997). Significance regression: a statistical approach to partial least squares. *Journal of Chemometrics*, **11**(4), 283–309.

Höskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics*, **2**(3), 211–228.

Höskuldsson, A. (1992). The H-principle in modelling with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems*, **14**(1), 139–153.

Höskuldsson, A. (1996). Dimension of linear models. *Chemometrics and Intelligent Laboratory Systems*, **32**(1), 37–55.

Höskuldsson, A. (2001). Variable and subset selection in PLS regression. *Chemometrics and Intelligent Laboratory Systems*, **55**(1), 23–38.

Hotelling, H. (1957). The relations of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology*, **10**(2), 69–79.

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Applied Statistics*, **31**(3), 300–303.

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, **12**(3), 531–547.

Jorissen, R. N., Gibbs, P., Christie, M., Prakash, S., Lipton, L., Desai, J., Kerr, D., Aaltonen, L. A., Arango, D., Kruhøffer, M., *et al.* (2009). Metastasis-associated gene expression changes predict poor outcomes in patients with Dukes stage B and C colorectal cancer. *Clinical Cancer Research*, **15**(24), 7642–7651.

Kallioniemi, A., Kallioniemi, O.-P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F., and Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**(5083), 818–821.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on Artificial intelligence-Volume 2*, pages 1137–1143, USA. Morgan Kaufmann Publishers Inc.

Komornik, V. (2002). *Précis d'analyse réelle: analyse fonctionnelle, intégrale de Lebesgue, espaces fonctionnels*. Ellipses, 32, rue Bargue, 75740 Paris Cedex 15, France.

Krämer, N. and Sugiyama, M. (2011). The degrees of freedom of partial least squares regression. *Journal of the American Statistical Association*, **106**(494), 697–705.

Krzanowski, W. (1987). Cross-validation in principal component analysis. *Biometrics*, **43**(3), 575–584.

Lazraq, A., Cleroux, R., and Gauchi, J.-P. (2003). Selecting both latent and explanatory variables in the PLS1 regression model. *Chemometrics and Intelligent Laboratory Systems*, **66**(2), 117–126.

Lê Cao, K.-A., Rossouw, D., Robert-Granié, C., and Besse, P. (2008). A sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, **7**(1). Article 35, DOI:10.2202/1544-6115.1390.

Le Cessie, S. and Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics*, **41**(1), 191–201.

Li, B., Morris, J., and Martin, E. B. (2002a). Model selection for partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, **64**(1), 79–89.

Li, Y.-C., Korol, A. B., Fahima, T., Beiles, A., and Nevo, E. (2002b). Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology*, **11**(12), 2453–2465.

Lindberg, W., Persson, J.-A., and Wold, S. (1983). Partial least-squares method for spectrofluorimetric analysis of mixtures of humic acid and lignin sulfonate. *Analytical Chemistry*, **55**(4), 643–648.

Liu, R. Y. *et al.* (1988). Bootstrap procedures under some non-iid models. *The Annals of Statistics*, **16**(4), 1696–1708.

Lorber, A., Wangen, L. E., and Kowalski, B. R. (1987). A theoretical foundation for the PLS algorithm. *Journal of Chemometrics*, **1**(1), 19–31.

Magnanensi, J., Bertrand, F., Maumy, M., and Meyer, N. (2015). A new Universal Resample Stable Bootstrap-based Stopping Criterion in PLS Components Construction. *arXiv preprint arXiv:1507.01404*.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **15**(4), 661–675.

Mammen, E. (1993). Bootstrap and Wild Bootstrap for High Dimensional Linear Models. *The Annals of Statistics*, **21**(1), 255–285.

Mammen, E. (2012). *When does bootstrap work?: asymptotic results and simulations*, volume 77. Springer Science & Business Media.

Manne, R. (1987). Analysis of two partial-least-squares algorithms for multivariate calibration. *Chemometrics and Intelligent Laboratory Systems*, **2**(1), 187–197.

Marbach, R. and Heise, H. (1990). Calibration modeling by partial least-squares and principal component regression and its optimization using an improved leverage correction for prediction testing. *Chemometrics and Intelligent Laboratory Systems*, **9**(1), 45–63.

Marisa, L., de Reyniès, A., Duval, A., Selves, J., Gaub, M. P., Vescovo, L., Etienne-Grimaldi, M.-C., Schiappa, R., Guenot, D., Ayadi, M., *et al.* (2013). Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS medicine*, **10**(5), e1001453.

Martens, H. (2001). Reliable and relevant modelling of real world data: a personal account of the development of PLS Regression. *Chemometrics and Intelligent Laboratory Systems*, **58**(2), 85–95.

Marx, B. D. (1996). Iteratively Reweighted Partial Least Squares Estimation for Generalized Linear Regression. *Technometrics*, **38**(4), 374–381.

Mehmood, T., Liland, K. H., Snipen, L., and Sæbø, S. (2012). A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, **118**, 62–69.

Messick, N. J., Kalivas, J. H., and Lang, P. M. (1997). Selecting factors for partial least squares. *Microchemical Journal*, **55**(2), 200–207.

Mevik, B.-H. and Cederkvist, H. R. (2004). Mean squared error of prediction (msep) estimates for principal component regression (pcr) and partial least squares regression (plsr). *Journal of Chemometrics*, **18**(9), 422–429.

Meyer, N., Maumy-Bertrand, M., and Bertrand, F. (2010). Comparaison de variantes de régressions logistiques PLS et de régression PLS sur variables qualitatives: application aux données d'allélotypage. *Journal de la Société Française de Statistique*, **151**(2), 1–18.

Miller, R. G. (1974). The jackknife-a review. *Biometrika*, **61**(1), 1–15.

Miller, R. G. *et al.* (1974). An unbalanced jackknife. *The Annals of Statistics*, **2**(5), 880–891.

Moulton, L. H. and Zeger, S. L. (1991). Bootstrapping generalized linear models. *Computational Statistics & Data Analysis*, **11**(1), 53–63.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., and Erlich, H. (1986). Specific enzymatic amplification of dna in vitro: The polymerase chain reaction. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 51, pages 263–273, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA. Cold Spring Harbor Laboratory Press.

Naidoo, R. and Chetty, R. (1998). The Application of Microsatellites in Molecular Pathology. *Pathology & Oncology Research*, **4**(4), 310–315.

Nelder, J. and Wedderburn, R. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, **135**(3), 370–384.

Nguyen, D. V. and Rocke, D. M. (2002a). Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics*, **18**(9), 1216–1226.

Nguyen, D. V. and Rocke, D. M. (2002b). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**(1), 39–50.

Nguyen, D. V. and Rocke, D. M. (2004). On partial least squares dimension reduction for microarray-based classification: a simulation study. *Computational Statistics & Data Analysis*, **46**(3), 407–425.

Osten, D. W. (1988). Selection of optimal regression models via cross-validation. *Journal of Chemometrics*, **2**(1), 39–48.

Panhard, X., Dominique, S., Gaub, M.-P., Ravery, V., Grandchamp, B., and Mentré, F. (2003). Construction of a global score quantifying allelic imbalance among biallelic SIDP markers in bladder cancer. *Statistics in Medicine*, **22**(24), 3771–3779.

Phatak, A. and de Hoog, F. (2002). Exploiting the connection between PLS, Lanczos methods and conjugate gradients: alternative proofs of some properties of PLS. *Journal of Chemometrics*, **16**(7), 361–367.

Phatak, A., Reilly, P., and Penlidis, A. (2002). The asymptotic variance of the univariate PLS estimator. *Linear Algebra and its Applications*, **354**(1), 245–253.

Picard, R. R. and Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, **79**(387), 575–583.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.-L., Chen, C., Zhai, Y., *et al.* (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, **20**(2), 207–211.

Quenouille, M. (1949). Approximate Tests of Correlation in Time-Series. *Journal of the Royal Statistical Society. Series B (Methodological)*, **11**(1), 68–84.

Quenouille, M. H. (1956). Notes on Bias in Estimation. *Biometrika*, **43**(3-4), 353–360.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rao, C. R. (2005). *Handbook of Statistics: Data Mining and Data Visualization*, volume 24. Elsevier, Radarweg 29, P.O. Box 211, 1000 AE Amsterdam, Netherlands.

Romera, R. (2010). Prediction intervals in Partial Least Squares regression via a new local linearization approach. *Chemometrics and Intelligent Laboratory Systems*, **103**(2), 122–128.

Rosipal, R. and Krämer, N. (2006). Overview and recent advances in partial least squares. In *Subspace, latent structure and feature selection*, pages 34–51. Springer, Springer-Verlag Berlin, Heidelberg, Deutschland.

Saad, Y. (2003). *Iterative methods for sparse linear systems*. SIAM, 3600 University City Science Center, Philadelphia, PA 19104, USA.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**(5235), 467–470.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American statistical Association*, **88**(422), 486–494.

Shao, J. and Tu, D. (2012). *The jackknife and bootstrap*. Springer Science & Business Media.

Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society. Series B (Methodological)*, **43**(3), 310–313.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics*, **9**(6), 1187–1195.

Skotheim, R. I., Diep, C. B., Kraggerud, S. M., Jakobsen, K. S., and Lothe, R. A. (2001). Evaluation of loss of heterozygosity/allelic imbalance scoring in tumor DNA. *Cancer Genetics and Cytogenetics*, **127**(1), 64–70.

Smyth, G. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**(1).

Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Döhner, H., Cremer, T., and Lichter, P. (1997). Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes, Chromosomes and Cancer*, **20**(4), 399–407.

Sun, W., Wang, J., and Fang, Y. (2013). Consistent selection of tuning parameters via variable selection stability. *The Journal of Machine Learning Research*, **14**(1), 3419–3440.

Tenenhaus, M. (1998). *La régression PLS, Théorie et pratique*. Editions Technip, 27, Rue Ginoux 75737 Paris Cedex 1.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), 267–288.

Tomlinson, I. P., Lambros, M. B., Roylance, R. R., and Cleton-Jansen, A.-M. (2002). Loss of heterozygosity analysis: practically and conceptually flawed? *Genes, Chromosomes and Cancer*, **34**(4), 349–353.

Tufféry, S. (2010). *Data mining et statistique décisionnelle: l'intelligence des données*. Editions Technip.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples. *The Annals of Mathematical Statistics*, **29**(2), 614.

Umetrics, A. (2005). User's guide to simca-p, simca-p+, version 10. 0. *Umeå, Sweden: Umetrics AB*.

Van der Voet, H. (1994). Comparing the predictive accuracy of models using a simple randomization test. *Chemometrics and Intelligent Laboratory Systems*, **25**(2), 313–323.

Van Wieringen, W. N., Kun, D., Hampel, R., and Boulesteix, A.-L. (2009). Survival prediction using gene expression data: a review and comparison. *Computational Statistics & Data Analysis*, **53**(5), 1590–1603.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.* (2001). The sequence of the human genome. *Science*, **291**(5507), 1304–1351.

Wakeling, I. N. and Morris, J. J. (1993). A test of significance for partial least squares regression. *Journal of Chemometrics*, **7**(4), 291–304.

Weber, J.-C., Meyer, N., Pencreach, E., Schneider, A., Guérin, E., Neuville, A., Stemmer, C., Brigand, C., Bachellier, P., Rohr, S., *et al.* (2007). Allelotyping analyses of synchronous primary and metastasis CIN colon cancers identified different subtypes. *International Journal of Cancer*, **120**(3), 524–532.

Wedderburn, R. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, **63**(1), 27–32.

Wehrens, R. and Linden, W. v. d. (1997). Bootstrapping principal component regression models. *Journal of Chemometrics*, **11**(2), 157–171.

Welch, B. L. (1947). The generalization of Student's problem when several different population variances are involved. *Biometrika*, **34**(1-2), 28–35.

Wiklund, S., Nilsson, D., Eriksson, L., Sjöström, M., Wold, S., and Faber, K. (2007). A randomization test for PLS component selection. *Journal of Chemometrics*, **21**(10-11), 427–439.

Wold, H. (1975). *Path models with latent variables: The NIPALS approach*. Acad. Press.

Wold, H. *et al.* (1966). Estimation of principal components and related models by iterative least squares. *Multivariate Analysis*, **1**, 391–420.

Wold, S. (1978). Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, **20**(4), 397–405.

Wold, S. (2001). Personal memories of the early PLS development. *Chemometrics and Intelligent Laboratory Systems*, **58**(2), 83–84.

Wold, S., Martens, H., and Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. In *Matrix Pencils*, pages 286–293. Springer, Springer-Verlag Berlin, Heidelberg, Deutschland.

Wold, S., Ruhe, A., Wold, H., and Dunn, III, W. (1984). The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, **5**(3), 735–743.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, **2**(1), 37–52.

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, **58**(2), 109–130.

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *the Annals of Statistics*, **14**(4), 1261–1295.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, **15**(2), 265–286.

UNIVERSITÉ DE STRASBOURG

Jérémy MAGNANENSI

Amélioration et développement de méthodes de sélection du nombre de composantes et de prédicteurs significatifs pour une régression PLS et certaines de ses extensions à l'aide du bootstrap.

École Doctorale des Sciences de la Vie et de la Santé STRASBOURG

Résumé

La régression Partial Least Squares (PLS), de part ses caractéristiques, est devenue une méthodologie statistique de choix pour le traitement de jeux de données issus d'études génomiques.

La fiabilité de la régression PLS et de certaines de ses extensions repose, entre autres, sur une détermination robuste d'un hyperparamètre, le nombre de composantes. Une telle détermination reste un objectif important à ce jour, aucun critère existant ne pouvant être considéré comme globalement satisfaisant. Nous avons ainsi élaboré un nouveau critère de choix pour la sélection du nombre de composantes PLS basé sur la technique du bootstrap et caractérisé notamment par une forte stabilité. Nous avons ensuite pu l'adapter et l'utiliser à des fins de développement et d'amélioration de procédés de sélection de prédicteurs significatifs, ouvrant ainsi la voie à une identification rendue plus fiable et robuste des *probe sets* impliqués dans la caractéristique étudiée d'une pathologie.

Régression PLS – Bootstrap – Composantes – Puces à ADN – Sélection de variable

Résumé en anglais

The Partial Least Squares (PLS) regression, through its properties, has become a versatile statistic methodology for the analysis of genomic datasets.

The reliability of the PLS regression and some of its extensions relies on a robust determination of a tuning parameter, the number of components. Such a determination is still a major aim since no existing criterion could be considered as a global benchmark one in the state-of-art literature. We developed a new bootstrap based stopping criterion in PLS components construction that guarantee a high level of stability. We then adapted and used it to develop and improve variable selection processes, allowing a more reliable and robust determination of significant probe sets related to the studied feature of a pathology.

PLS regression – Bootstrap – Components – Microarray – Variable selection