université de BORDEAUX

Thèse présentée pour obtenir le grade de

DOCTEUR
DE L'UNIVERSITÉ DE BORDEAUX

École doctorale Sociétés, Politique, Santé Publique
Spécialité Santé Publique, option Biostatistique

Par Perrine Soret

Régression pénalisée de type Lasso pour l'analyse de données biologiques de grande dimension

Application à la charge virale du VIH censurée par une limite de quantification et aux données compositionnelles du microbiote

Soutenue le 28/11/2019

Membres du jury

Chiquet Julien	DR INRA, Université Paris-Saclay, France	Rapporteur
Flandre Philippe	CR INSERM, Institut Pierre Louis d'Épidémiologie et de Santé Publique, Paris, France	Rapporteur
Meza Cristian	Professeur, Université de Valparaiso, Chili	Examinateur
Rondeau Virginie	DR INSERM, Université de Bordeaux, France	Examinatrice
Delhaes Laurence	PU-PH, Université de Bordeaux, France	Invitée
Thiébaut Rodolphe	PU-PH, Université de Bordeaux, France	Directeur d'équipe
Avalos Marta	MCU, Université de Bordeaux, France	Directrice de thèse

Avant-Propos

Le premier titre envisagé pour ma thèse a été *Lasso pour l'analyse de données longitudinales de grande dimension*, avec une application à l'analyse de données issues d'un essai pour un vaccin thérapeutique contre le VIH (Dalia-1). Ce sujet de thèse constituait ainsi la suite du stage de Master 2 Biostatistique, Université de Montpellier, que j'ai effectué au sein de l'équipe SISTM sous la direction de Marta Avalos, au cours duquel j'ai effectué un état des lieux de l'adaptation des méthodes d'apprentissage statistique pour des données longitudinales.

Nous nous sommes premièrement attaqués dans le cas des études transversales à un problème retrouvé dans des données VIH : la censure due à la limite de détection et cela en grande dimension. Cette recherche a été effectuée en collaboration avec Linda Wittkop responsable de l'équipe *VIH, Hépatites virales et co-morbidités : épidémiologie clinique et santé publique*.

Cette partie a néanmoins occupé les deux premières années de ma thèse, compromettant ainsi les développements méthodologiques liés à l'adaptation de la méthode Lasso aux données longitudinales.

Une réorientation a alors été envisagée : investir l'expertise développée dans l'application des régressions pénalisées pour des données biologiques à des données présentant une structure particulière, celles du microbiome. Ce travail a été réalisé en collaboration avec l'équipe de Laurence Delheas, notamment avec Louise Eva Vandenborght. Mon exploration des méthodes appliquées à l'analyse des données de microbiote a coïncidé avec un séjour de 3 mois en 2017 dans l'équipe Data61 de Canberra (Australie), où j'ai rejoint ma directrice de thèse en mobilité. Les échanges avec Cheng Soon Ong sur la thématique ont été enrichissants et ont ouvert de nouvelles pistes de recherche, développées en dehors de ma thèse.

Une difficulté éprouvée tout au long de la thèse a été celle qui concerne la restitution écrite de la recherche, que ce soit la rédaction d'article ou du mémoire. A cette difficulté se rajoute l'obstacle de la publication. Pour l'article *Lasso-regularization for left-censored outcome and high-dimensional predictors* publié dans *BMC medical research methodoly*, 12 mois se sont écoulés avant d'obtenir le premier rapport de deux rapporteurs. L'article *Respiratory mycobiome and suggestion of inter-kingdom network during acute pulmonary*

exacerbation in cystic fibrosis, à paraître prochainement dans *Scientific reports* a été, quant à lui, préalablement rejeté par trois journaux. L'utilisation de méthodes provenant de l'apprentissage statistique pour des données de grande dimension (avec un faible nombre de patients) est encore récente et contestée. Ces obstacles m'ont conduit à la demande d'une 4ème année de thèse financée par l'équipe SISTM entre octobre 2017 et juin 2018.

En juin 2018 j'ai été recrutée en tant que "biostatisticienne biomarqueur" dans le département *Modélisation et valorisation des méthodes en recherche et développement* des laboratoires Servier. Techniquement, il s'agit d'une année de césure, jusqu'à ma soutenance en 2019, sans que la démarche administrative n'est été effectuée. Mes missions actuelles incluent la restitution des résultats de nos recherches sous forme d'article scientifique et de communication en congrès.

Les années de thèse ont été également l'occasion de m'investir dans des activités liées à l'encadrement et l'enseignement. J'ai ainsi encadré un stage de M1 et co-encadré un stage de M2. J'ai également participé à l'élaboration du cours en ligne nommé Begin'R pour l'apprentissage du logiciel R sous la direction de Florent Arnal. Par ailleurs, j'ai eu l'honneur d'être élue présidente de l'association des doctorants de l'Ecole doctorale Sociétés, Politique, Santé Publique de l'Université de Bordeaux, SP2. Dans le cadre de mes responsabilités administratives, j'ai participé à l'organisation des journées de l'Ecole doctorale SP2.

Table des matières

1 Introduction

1.1 Les données biomédicales de grande dimension

Les progrès scientifiques, techniques et informatiques réalisés ces dernières années ont permis d'augmenter considérablement les capacités de mesure, de stockage et de traitement de données. En particulier, les champs de la médecine et de la biologie sont devenus de grands « producteurs de données » [Marx, 2013; Thiébaut et al., 2014; Mooney and Pejaver, 2018]. Le développement et la réduction des coûts des techniques de séquençage ont accéléré par exemple la recherche médicale. Il y a encore 10 ans, l'étude des microbes s'effectuait en culture sur un nombre très réduit de sujets et sur une faible variété de bactéries. L'analyse en qPCR (quantitative polymerase chain reaction) où le principe est d'amplifier l'Acide désoxyribonucléique (ADN) ou l'Acide ribonucléique (ARN) d'une cible connue à partir d'une faible quantité, avait déjà permis d'améliorer la connaissance des microbes [MacDougall et al., 2018]. Mais ce n'est qu'avec l'arrivée du NGS (*Next-generation Sequencing*) que la recherche médicale autour des bactéries a pris son envol. En culture, la présence des bactéries est explorée une à une. Si des informations ne sont pas disponibles pour guider l'exploration, il ne sera pas possible de conduire l'étude. En revanche, il est possible d'extraire la composition globale de l'environnement étudié sans aucune connaissance préalable sur les bactéries à partir des données issues de séquençages. De ce fait, pour un environnement donné, il est possible de détecter jusqu'à plusieurs centaines d'espèces. Ces nouvelles techniques ont permis de développer de nouvelles stratégies pour répondre à des questions telles que « Est-ce que des *signatures microbiennes* sont associées à des maladies respiratoires chroniques telles que l'asthme, la broncho pneumopathie chronique obstructive ou encore des maladies génétiques telles que la mucoviscidose ? » [Delhaes et al., 2012].

Un autre exemple des apports technologiques est celui de l'adaptation des traitements des patients atteints du Virus d'Immunodéficience Humaine (VIH). De nos jours, les traitements antirétroviraux permettent d'empêcher la progression de la maladie mais ils ne sont pas capables d'éradiquer totalement l'infection. La réplication de virus présentants des mutations de résistance aux traitements antirétroviraux peut conduire à l'échec thérapeutique. Des tests génotypiques qui vérifient la présence ou non de mutations résistantes

dans le génome du patient sont ainsi effectués préalablement à l'initiation d'un nouveau traitement. Ces tests sont basés sur le séquençage de gènes. Des centaines de mutations connues comme étant probablement associées à des résistances du virus sont explorées. Ces tests permettent d'adapter pour un patient donné les traitements antirétroviraux.

Nous faisons face à ces grands volumes de données à analyser où le nombre de variables (p) est, en général, du même ordre, voire plus grand, que le nombre d'individus (n) rendant la statistique classique inadaptée où l'une des conditions essentielles est $n > p$. On parle alors de données de grande dimension.

La recherche statistique répond à cette génération massive de données en développant de nouvelles méthodes d'analyse, souvent empruntées au domaine de l'apprentissage automatique (*machine learning*, en anglais) ou apprentissage statistique. A l'intersection de la statistique, de l'intelligence artificielle, de l'optimisation ou encore de l'automatique, l'apprentissage statistique joue de nos jours un rôle croissant dans de nombreux domaines d'application où le principal objectif est la prédiction.

Cependant, dans la recherche médicale, on cherche généralement à identifier, expliquer les facteurs associés à la réponse d'intérêt : « Quelles sont les bactéries associés à une exacerbation chez des patients atteints de mucoviscidose ? » « Quelles sont les mutations du VIH associées à une augmentation de la charge virale chez des patients infectés par le VIH. » Si pour le clinicien, un objectif explicatif ou un objectif prédictif est semblable, il en est pas de même pour le statisticien.

Dans le premier cas, l'accent est mis sur la minimisation des biais, afin d'obtenir la représentation la plus précise possible du modèle sous-jacent et de comprendre le rôle exact de chaque variable. En revanche, lorsque l'objectif est prédictif, l'accent est mis sur l'obtention d'une prédiction précise de la réponse d'intérêt pour de futurs patients sans forcément se poser des questions en termes de causalité. L'on cherchera à minimiser la combinaison du biais et de la variance, sacrifiant éventuellement l'interprétation au profit d'une prédiction précise [Breiman et al., 2001; Shmueli et al., 2010; Helmbold and Long, 2012; Falissard, 2018] (Épistémologie de la statistique à l'heure du tout digital, Falissard Bruno,*Journées de l'École Doctorale SP2*, 2018, Université de Bordeaux).

Dans les applications bio-médicales, il est ardu de faire un choix entre deux. En général, on souhaitera obtenir un modèle prédictif qui permet de comprendre ou un modèle explicatif qui permet (« logiquement ») de prédire. Dans un contexte de grande dimension, il est plus aisé de s'attaquer au problème statistique à partir d'une perspective de prédiction. Il est important d'être alors prudents sur les conclusions lors de l'interprétation du rôle des variables. A l'inverse, si le choix se porte sur un modèle explicatif, il faut admettre que le modèle choisit peut ne pas conduire à la meilleure prédiction. Les discussions entre les cliniciens et les statisticiens sont cruciales pour définir les objectifs et mettre en garde

sur les limites des analyses en termes d'interprétation.

Des familles de méthodes présentant un bon compromis entre explication/identification et prédiction sont les régressions pénalisées (Ridge [Hoerl and Kennard, 1970], Lasso [Tibshirani, 1996], par exemple) et la régression des moindres carrés partiels (*Partial Least Square*, PLS [Wold et al., 1983]). En effet, les modèles sous-jacents sont ceux bien connus dans le domaine de la biologie et de la médecine (modèle de régression linéaire, modèle de régression logistique, modèle de Cox, ...). Ce sont les étapes d'estimation et sélection de variables qui diffèrent. De plus, des procédures de rééchantillonnage combinées avec les régressions pénalisées permettent d'obtenir des résultats plus robustes en termes d'identification des prédicteurs réellement associées à la variable réponse [Bach, 2008].

La grande dimension n'est pas le seul défi engendré par les nouvelles techniques NGS. L'augmentation de la quantité d'information est souvent couplée à des structures de données complexes générées à partir de sources différentes. En effet, selon la problématique clinique de départ, le design, les données nécessaires et les techniques utilisées peuvent être différents. Cette thèse traite deux structures en particulier. Dans un premier temps, nous avons étudié le problème de la censure due à des limites de quantification engendrées par le manque de sensibilité des techniques de mesure. Dans un second temps, nous avons exploré la structure hiérarchique et de composition des données microbiote (la communauté microbienne vivant dans un environnement spécifique, chez l'homme ici).

Cette thèse est centrée sur l'exploration, l'adaptation ou encore l'extension de méthodes de pénalisation telles que le Lasso pour prendre en compte la complexité des données décrites ci-dessus.

1.2 Objectifs cliniques et défis méthodologiques

1.2.1 Prise en compte de la censure pour des données de grande dimension : application à la prédiction de la charge virale du VIH à partir des mutations du VIH

La censure à gauche due à un défaut de la sensibilité de l'appareil de mesure est une problématique commune dans de nombreux domaines d'application comme la biologie, la chimie et l'environnement. Un exemple est la quantification de la charge virale, dans le plasma, chez des patients atteints de VIH. La « vraie » valeur de la charge virale n'est alors pas connue, on dit qu'elle est « indétectable ». La sensibilité des techniques de mesure a cependant évolué et les seuils de détection ont diminué de 10,000 copies/mL à 20 copies/mL d'ARN. En statistique, on parle de données censurées à gauche par limite de quantification que l'on peut considérer comme un cas particulier du problème des valeurs

manquantes. Plusieurs approches statistiques ont été proposées pour tenir compte de la censure de ces variables quantitatives : dans des études transversales mais également longitudinales. Les méthodes standards incluent l'imputation multiple, les méthodes de survie en renversant le problème de la censure à droite, la régression quantile ou la régression quantile censurée. Une autre approche standard est celle du modèle Tobit [Tobin, 1958]. A partir d'une variable réponse censurée supposée gaussienne,les paramètres de ce modèle peuvent être estimés par maximum de vraisemblance ou par l'estimateur de Buckley-James [Buckley and James, 1979]. Dans un contexte classique de faible dimension, ces méthodes ont montré de meilleures performances que l'imputation simple (substitution de la valeur censurée par une autre valeur) .

Dans le contexte d'une infection par le VIH, on cherche à analyser l'association entre certaines mutations du VIH et la réponse au traitement antirétroviral. Les souches VIH circulant dans un individu peuvent présenter des mutations associées à une défaillance du traitement antirétroviral (cART). On les appelle les mutations résistantes (Figure 1.1). En conséquence, pour évaluer la résistance aux médicaments, des tests génotypiques sont effectués chez les patients commençant un nouveau traitement antirétroviral (2ème cART dans la figure 1.1) ou chez de nouveaux patients infectés par le VIH, en raison de la transmission de souches résistantes [Hirsch et al., 2008; Wittkop et al., 2011; Hofstra et al., 2016; Wensing et al., 2017].

Les analyses statistiques ont pour objectif de trouver les mutations génotypiques associées à la réponse virologique afin de prédire la résistance au traitement. La régression linéaire et logistique de type Lasso, [Rabinowitz et al., 2006; Beerenwinkel et al., 2013; Cozzi-Lepri et al., 2011], la régression logistique PLS [Wittkop et al., 2008] et les tests univariés ajustés sur la multiplicité [Assoumou et al., 2010] ont été appliqués dans le contexte commun de grande dimension avec $p = 100$ et $n = 1000$ [Rhee et al., 2006]. Cependant, ces études utilisent une réponse binaire ou l'imputation simple, pour s'affranchir de la censure dans le contexte de la grande dimension. Une limite de la dichotomisation d'une variable quantitative est la perte d'information et de puissance du modèle. Nous faisons l'hypothèse que les approches prenant en compte la censure à gauche amènent à de meilleurs résultats que l'imputation simple autant dans le cas de la grande dimension que dans le cas de la faible dimension ($n > p$).

Certains travaux adressent simultanément la problématique de la censure et de la grande dimension, mais sont essentiellement limités aux données de survie (censure à droite). Les régressions pénalisées ont largement été étendues à la censure à droite avec Tibshirani [1997] pour les modèles de Cox, la régression quantile, [Shows et al., 2010; Wang et al., 2010, 2013; Müller and van de Geer, 2015], ou, plus facilement interprétables, les modèles à durée de vie accéléré (AFT pour *Accelerated Failure Time*), connus pour être

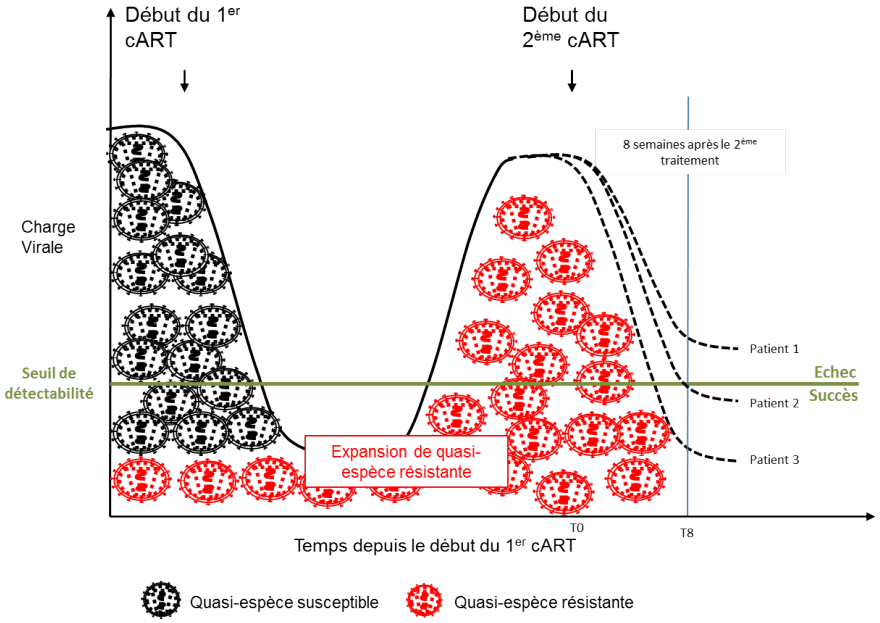

FIGURE 1.1 – Représentation schématique de deux réponses virologiques suite à deux phases de traitement.

plus facilement interprétables [Huang et al., 2006; Datta et al., 2007; Wang et al., 2008; Johnson, 2008, 2009b,a; Ueki, 2009; Wang and Wang, 2010; Chung et al., 2013; Zhao et al., 2014; DiRienzo, 2016]. Par ailleurs, d'autres méthodes d'apprentissage ont été proposées pour les analyses de survie telles que les forêts aléatoires [Ishwaran et al., 2008], SVM [Wang et al., 2016], et les réseaux de neurones [Van der Burgh et al., 2017].

L'algorithme itératif de Buckley-James, imputant les valeurs censurées à droite par une espérance conditionnelle basée sur l'estimateur de Kaplan-Meier, a été adapté à la grande dimension dans le cas d'analyse de survie. Des extensions de la régression PLS [Huang et al., 2004] et de la régression Lasso ont été proposées pour gérer la censure à droite [Johnson, 2008, 2009b; Cai et al., 2009; Wang and Wang, 2010]. Plus récemment, Wiegand et al. [2016] a appliqué l'algorithme à un cas de censure à gauche dans une étude transversale et de faible dimension. Nos travaux proposent une extension Lasso de la méthode de Buckley-James. Notre première approche consiste à renverser la variable réponse pour obtenir une censure à droite afin d'utiliser la version non-paramétrique pénalisée de Buckley-James [Wang and Wang, 2010]. Dans un second temps, nous proposons une version paramétrique pénalisée de l'estimateur de Buckley-James basée sur une hypothèse gaussienne.

Le jeu de données motivant cette étude provient du projet « *Standardization and Clinical Relevance of HIV Drug Resistance Testing Project* » [Cozzi-Lepri, 2008]. Les patients inclus dans cette étude profitent d'un nouveau régime incluant la molécule Abacavir, une molécule de type NRTI (*Nucleoside Reverse Transcriptase Inhibitor*) qui bloque la

transcriptase inverse. La taille de l'échantillon est légèrement plus faible que le nombre de prédicteurs. Environ la moitié des prédicteurs correspondent à la présence/absence de mutations spécifiques au gène de la transcription inverse. Les autres sont spécifiques du gène protéase. La variable réponse est la log-charge virale mesurée à la huitième semaine après le traitement, soumis à une limite de détection fixée à 100 copies/mL pour un taux de censure modéré. Dans le cadre de cette thèse, nous nous sommes limités au dévelppement de méthodes avec un objectif prédictif.

1.2.2 Analyse des données de microbiote : compréhension de l'association avec la sévérité de la mucoviscidose

La réduction du coût des techniques de séquençage haut débit (NGS), en particulier le gène 16S séquençant les composantes bactériennes de la communauté microbienne humaine (microbiote), a donné une nouvelle perspective à la recherche des maladies humaines tel que le montre le nombre de publications depuis le début des années 2000. En conséquence, des associations entre le microbiote et diverses maladies ont été détectées. Par exemple, des liens entre une altération de la flore intestinale et le diabète de type 2 [Qin et al., 2012], les maladies cardiovasculaires [Koeth et al., 2013], les conditions psychologiques grace à l'axe « intestin-cerveau » [Cryan and O'Mahony, 2011], mais également la maladie de Crohn et le syndrome du côlon irritable, ont été observés. Le microbiote intestinal reste le plus étudié jusqu'à présent mais l'exploration d'autres microbiomes est en plein essor.

Jusqu'à présent, les voies aériennes étaient considérées comme stériles. Cependant, les techniques de séquençage ont révélé l'existence d'une flore polymicrobienne (bactérienne, virale ou fongique), bactérienne, virale ou fongique, constituant le microbiote pulmonaire. Celui-ci, moins diversifié que le microbiote intestinal, est maintenant considéré comme un écosystème spécifique où l'abondance et la diversité des micro-organismes sont uniques à chaque individu [Dickson et al., 2013; Marsland and Gollwitzer, 2014; Andréjak and Delhaes, 2015]. La littérature sur ce domaine porte principalement sur la comparaison entre les communautés bactériennes pulmonaires de sujets sains [Charlson et al., 2011] et celles de patients atteints de maladies pulmonaires chroniques telles que l'asthme [Hilty et al., 2010; Goleva et al., 2013; Marri et al., 2013; Simpson et al., 2016; Zhang et al., 2016; Durack et al., 2017; Sverrild et al., 2017], la maladie broncho-pulmonaire obstructive chronique (BPCO) [Weinreich and Korsgaard, 2008; Molyneaux et al., 2013; Zakharkina et al., 2013; Huang et al., 2014], la mucoviscidose [Frayman et al., 2017; Pittman et al., 2017; Leite et al., 2017; Acosta et al., 2017; Heirali et al., 2017; Feigelman et al., 2017; Cox et al., 2017; Boutin and Dalpke, 2017; Nguyen et al., 2016; Beaume et al., 2017; Cribbs

and Beck, 2017; de Koff et al., 2016], la fibrose pulmonaire idiopathique [Wang et al., 2017] ou encore le cancer du poumon [Hosgood et al., 2014; Yu et al., 2016; Lee et al., 2016].

Dans les maladies pulmonaires, en particulier la mucoviscidose (ou fibrose kystique), l'infection joue un rôle critique. Les évaluations actuelles sont nouvellement enrichies par l'analyse de données NGS d'une communauté polymicrobienne présente dans les voies aériennes des patients malades [O'Brien and Fothergill, 2017; Botterel et al., 2017; Nguyen et al., 2016; Quinn et al., 2016b; Whiteson et al., 2014; Willger et al., 2014; Lim et al., 2013; Charlson et al., 2012; Delhaes et al., 2012; Willner et al., 2012]. L'exacerbation pulmonaire aigüe dans la mucoviscidose représente un évènement clinique majeur, impactant significativement le déclin de la fonction pulmonaire et la progression de la maladie, qui requiert des traitements anti-infection adaptés [Bhatt, 2013; Stenbit and Flume, 2011; Bilton et al., 2011; Goss and Burns, 2007]. L'association entre l'exacerbation et la flore bactérienne a récemment été confirmée en utilisant des approches omiques [Nguyen et al., 2016; Quinn et al., 2016a, 2015; Carmody et al., 2013; Tunney et al., 2013; Zemanick et al., 2013; Filkins et al., 2012; Fodor et al., 2012; Zhao et al., 2012]. En ce qui concerne la composition fongique, même si des espèces sont associées à un déclin de la fonction pulmonaire, peu d'études ont étudié l'association avec une exacerbation aigüe [Nguyen et al., 2016; Willger et al., 2014].

Les données de séquençage se présentent sous forme de données de comptage (avec, en général, un nombre important de zéros), interprétées comme des abondances de taxons dans la communauté microbienne. Pour rendre les échantillons comparables, les données sont normalisées et exprimées en abondance relative sur l'ensemble des bactéries observées. Il s'agit des données compositions (CoDA). Les CoDA sont une collection de mesures non négatives sommant à 1. Connaissant la somme, une composante peut être déterminée à partir de la somme des autres. Les parties de la composition sont alors dépendantes mathématiquement et appartiennent à un espace spécifique appelé le simplex. Les CoDA peuvent être transférées dans l'espace euclidien en utilisant des transformations non linéaires pour permettre des inférences valides [Aitchison and Bacon-Shone, 1984].

En outre, les données sur le microbiote sont organisées selon une structure phylogénétique (Figure 1.2) qui peut conduire à une situation de grande dimension lorsqu'on la profondeur de l'arbre est importante. En parallèle, en réponse aux besoins, il y a une émergence de méthodes statistiques spécifiques et d'outils de calcul. En raison de la nouveauté, il est encore tôt pour évaluer l'applicabilité et la précision des méthodes disponibles. Les études de simulation, imitant une distribution de données réelles, sont un outil standard pour comparer la performance de méthodes statistiques compétitives. Cependant, la complexité des données sur le microbiote rend difficile la production de données réalistes.

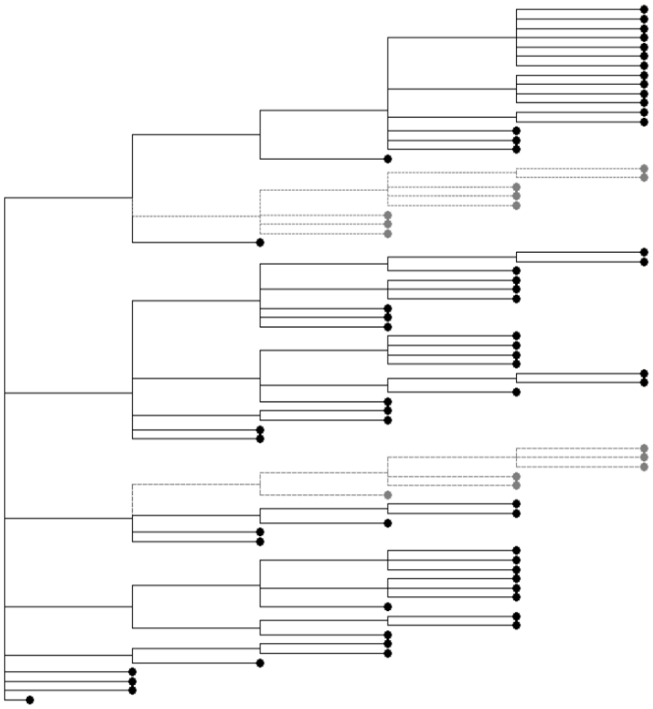

FIGURE 1.2 – Schématisation d'un arbre phylogénétique. Chaque trait représente une branche, une branche se sépare à un noeud et les points représentent les feuilles.

Dans le cadre de cette thèse, nous proposons une revue « non exhaustive » des méthodes d'analyse de données issues du microbiote, allant de l'analyse de diversité aux méthodes statistiques d'apprentissage. Une étude de simulation a également été conduite pour comparer les méthodes de pénalisation dans le but d'identifier les microbes associés à un statut clinique. Dans un deuxième temps, nous avons considéré les compositions bactériennes et fongiques comme des entités uniques dans l'étude de l'association entre le microbiote pulmonaire et la sévérité. Nos résultats ont été interprétés selon une approche de co-occurence/co-exclusion mettant en relation les micro-organismes [Delhaes et al., 2012; Charlson et al., 2012; Conrad et al., 2013; Quinn et al., 2014; Whiteson et al., 2014; Willger et al., 2014; Kramer et al., 2015; Quinn et al., 2016a,b; Nguyen et al., 2016; O'Brien and Fothergill, 2017; Botterel et al., 2017; Krause et al., 2017].

Les données analysées proviennent d'une étude cas-témoins nichée au sein de la cohorte MucoFong : un faible nombre de patients pour lesquels des données de métagénomique ciblée déterminant microbiote et mycobiote du tractus respiratoire sont disponibles. Dans le cadre de cette thèse, nous avons appliqué les méthodes d'analyse statistique des données du microbiote identifiées dans la littérature dans l'objectif de mieux comprendre le rôle des composantes bactériennes et fongiques dans la sévérité de la mucovicidose.

1.3 Structure de la thèse

Cette thèse propose d'analyser des données biomédicales de grande dimension présentant des structures particulières sous l'angle des méthodes de pénalisation de type Lasso.

Le chapitre 2 propose un rappel des méthodes de pénalisation qui constituent la base des méthodes dévéloppées ou utilisées pour l'analyse des données des projets du projet « Standardization and Clinical Relevance of HIV Drug Resistance Testing Project » et MucoFong.

Dans le chapitre 3, nous proposons une adaptation de l'estimateur Buckley-James pour la censure à gauche par limite de détection, puis nous l'étendons à la régression pénalisée pour des données de grande dimension. Une étude de simulation et une application sur les données réelles sont présentées. Ce chapitre fait l'objet d'un article publié dans *BMC Medical Research Methodology*.

Le chapitre 4 expose une revue des méthodes statistiques pour l'analyse de données de microbiote. Une étude de simulation est proposée pour comparer les méthodes adaptées à la grande dimensionidentifiées dans la littérature. L'étude de simulation a pour but de nous aider à faire le choix d'analyse des données MucoFong. Une analyse statistique des données de MucoFong fait l'objet d'un article accepté dans *Scientific Reports*.

Le dernier chapitre résume l'ensemble de la thèse sous forme de conclusion et d' ouvertures pour de nouvelles recherches statistiques.

2 Les régressions pénalisées

Le problème de sélection de variables est un des axes de recherche majeurs en statistique. Depuis une vingtaine d'années, la recherche statistique s'efforce de trouver des critères de pénalisation alternatifs à la classique sélection de variables pas à pas insatisfaisante en grande dimension [Tibshirani, 1996]. Ce chapitre se concentre principalement sur les méthodes de régression pénalisée. L'approche par modèle pénalisé permet d'apporter une solution en présence de multicolinéarité entre les variables (pourvu qu'elle ne soit pas sévère) ou encore lorsque n est inférieur ou comparable à p. Certaines pénalisations sont très avantageuses dans les cas habituellement difficiles à traiter lorsque le nombre de variable est supérieur au nombre d'individus.

2.1 Notations

- n le nombre d'individus
- p le nombre de variables explicatives ou prédicteurs (selon contexte)
- Y_i la réponse clinique de l'individu i
- $\mathbf{Y} = (Y_1, \cdots, Y_n)^\top$ le vecteur réponse
- X_{ij} la j-ème variable explicative de l'individu i
- $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip}), i = 1, \cdots, n$ vecteur ligne des variables aléatoires pour l'individu i
- $\mathbf{X}_j = (X_{1j}, \cdots, X_{nj})^\top, j = 1, \cdots, p$ vecteur colonne de la variable aléatoire j sur l'ensemble des individus
- $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ vecteur colonne des paramètres de régression

Soit l'échantillon $(Y_i, \mathbf{X}_i)$ avec $i = 1, \cdots, n$. Nous considérons les données centrées et réduites, par simplicité dans la suite.

2.2 Modèle

Considérons le modèle de régression linéaire suivant :

$$Y_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \cdots, n \tag{1}$$

avec $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ un p-vecteur contenant les paramètres associés à chacune des covariables et ε_i le terme d'erreur supposé indépendant et identiquement distribué (i.i.d.) suivant une distribution normale de moyenne 0 et de variance σ^2. Soit $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_n)^\top$ la matrice de taille $n \times p$ et $\mathbf{Y} = (Y_1, \cdots, Y_n)^\top$ le $n \times 1$ vecteur.

Les paramètres inconnus du modèle $\boldsymbol{\beta}$ sont couramment estimés par minimisation des moindres carrés soit la somme des carrés des résidus :

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

Si $\mathbf{X}^\top\mathbf{X}$ est inversible alors l'estimateur des moindres carrés $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}$.

Dans cette thèse, nous nous plaçons dans le contexte suivant :

— **Grande dimension** : p est du même ordre voire plus grand que n

— **Parcimonie** : on suppose que certains coefficients de $\boldsymbol{\beta}$ sont égaux à zéro

— **Pénalité convexe** : résolution numérique plus accessible

Dans ce contexte, la méthode des moindres carrés rencontre certains problèmes. La matrice $\mathbf{X}^\top\mathbf{X}$ n'est plus inversible et le nombre de degrés de liberté du système est tel que l'estimateur n'est pas défini de manière unique. Par ailleurs, si certaines variables sont corrélées, le calcul de l'inverse de $\mathbf{X}^\top\mathbf{X}$ peut être numériquement instable engendrant des estimateurs avec des grandes variances donnant lieu à des intervalles de confiances larges.

Les méthodes de sélection pas à pas largement utilisées en dimension faible à modérée, présentent de nombreux problèmes lorsque le nombre de variables candidates dépasse le nombre d'individus. De nombreux auteurs ont écrit sur le sujet dans le but de remédier à ces limites. Parmi l'étendue des méthodes trouvées dans la littérature (sélection automatiques, méthodes bayésiennes, $\cdots$) nous avons choisi de porter notre attention sur les régressions pénalisées.

Leur objectif est de minimiser les moindres carrés en imposant une contrainte sur les coefficients du modèle afin de combiner bonne prédiction et parcimonie du modèle.

2.3 Ridge

La régression Ridge utilise une pénalité en norme L_2, qui permet de réduire la valeur des coefficients de régression (shrinkage), mais pas jusqu'à 0. Elle permet de stabiliser l'es-

timation des coefficients dans le cas de fortes corrélations entre les variables. L'estimateur est obtenu en minimisant le critère suivant :

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\} \qquad \lambda \geq 0$$

avec $\|\boldsymbol{\beta}\|_2^2 = \sum_j \beta_j^2$. Il existe un λ à partir duquel l'inverse de la matrice $\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}$ existe. Il en découle un problème d'optimisation convexe qui possède une solution analytique :

$$\hat{\boldsymbol{\beta}}_{(\lambda)} = \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{Y}$$

Cet estimateur fut introduit par Hoerl and Kennard [1970]. Bien que biaisé, l'estimateur $\hat{\boldsymbol{\beta}}_R$ possède une variance plus faible que l'estimateur des moindres carrés, il s'agit donc de trouver le bon compromis biais-variance. Le fait d'ajouter un terme sur la diagonale de $\mathbf{X}^\top\mathbf{X}$ permet de stabiliser les calculs de l'inverse, ce qui rend cet estimateur plus stable au problème de multi-colinéarité.

Le coin du UseR : Le package `glmnet` développé par Friedman et al. [2010] permet d'effectuer une régression Ridge en fixant l'option `alpha = 0` dans la fonction `glmnet`

2.4 Lasso

Introduit par Tibshirani [1996], le Lasso (Least Absolute Shrinkage and Selection Operator) utilise comme fonction de pénalisation la norme L_1 du vecteur de paramètres $\boldsymbol{\beta}$ qui peut combiner les capacités d'une élimination des variables classique en supprimant automatiquement les petits coefficients, et les capacités d'une norme L_2 pour stabiliser les estimations. L'estimateur Lasso est solution du problème d'optimisation suivant :

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_j |\beta_j| \right\} \qquad \lambda \geq 0 \tag{2}$$

Le paramètre de régularisation λ contrôle l'importance de la pénalité. Plus particulièrement, pour $\lambda = 0$ aucune pénalité n'est appliquée, on retrouve l'estimateur des moindres carrés et pour $\lambda \to \infty$, toutes les variables sont associées à un coefficient estimé nul.
L'utilisation de la norme L_1 permet de réduire les petits coefficients exactement à zéro et donc de simplifier le modèle. Cette norme a l'avantage d'être convexe et de garantir l'unicité de la solution (si $n > p$ et la corrélation entre les variables n'est pas sévère). L'inconvénient de cette pénalité est que sa non différentiabilité en zéro complique son calcul. Il n'existe pas de solution analytique. De nombreux algorithmes ont été développés pour trouver la solution au problème d'optimisation (2) : programme linéaire [Tibshirani,

1996], algorithme du *shooting Lasso* [Fu, 1998], Newton modifiée [Fan and Li, 2001], algorithme *Least Angle Regression* [Efron et al., 2004] et descente de coordonnées circulaire [Friedman et al., 2010], parmi d'autres. Le Lasso peut biaiser les estimations de façon importante en réduisant trop fortement les coefficients. Pour remédier à cette problématique, Fan and Li [2001] proposent la pénalité *Smoothly Clipped Absolute Deviation* qui pénalise plus fortement les coefficients les plus élevés (en valeurs absolues). Toutefois, le problème algorithmique devient plus compliqué.

Le coin du UseR : Le package `glmnet` permet d'effectuer une régression Lasso en fixant l'option `alpha = 1` dans la fonction `glmnet`

2.5 La régression Bridge

Introduit par Frank and Friedman [1993], la régression Bridge propose le problème d'optimisation suivant :

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_j |\beta_j|^\gamma \right\} \qquad \gamma > 0, \lambda \geq 0 \tag{3}$$

Pour un $\lambda \geq 0$. Cette régression cherche alors à minimiser la somme des carrés des résidus pénalisée où pour chaque $\lambda \geq 0$ il existe un $c \geq 0$ tel que le problème (3) puisse s'écrire :

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \right\}, \qquad \text{sous contrainte} \qquad \sum_j |\beta_j|^\gamma < c \qquad \gamma > 0 \tag{4}$$

Le paramètre γ permet de définir différentes régions de contraintes (figure 2.1, en incluant les cas particuliers de la Ridge et le Lasso). Fu [1998] étudie les propriétés de l'estimateur pour $\gamma \geq 1$ et propose une version modifiée de Newton-Raphson pour $\gamma > 1$. Plus le paramètre γ est faible, plus la zone de contrainte est restreinte à une faible zone de possibilités pour l'estimateur. Pour $\gamma < 2$, les zones de contraintes admettent des sommets (Figure 2.1), cette particularité permet la sélection. Si l'optimum se trouve sur un sommet de la zone alors le paramètre est estimé exactement à zéro. Pour $\gamma \geq 1$, les pénalités sont convexes, ce qui permet d'obtenir des propriétés intéressantes, utiles à la résolution du problèmes d'optimisation. L'objectif est alors de trouver le compromis entre une zone de contrainte restreinte et un estimateur le moins biaisé possible. Deux paramètres, γ et λ sont donc à fixer pour déterminer $\hat{\boldsymbol{\beta}}$.

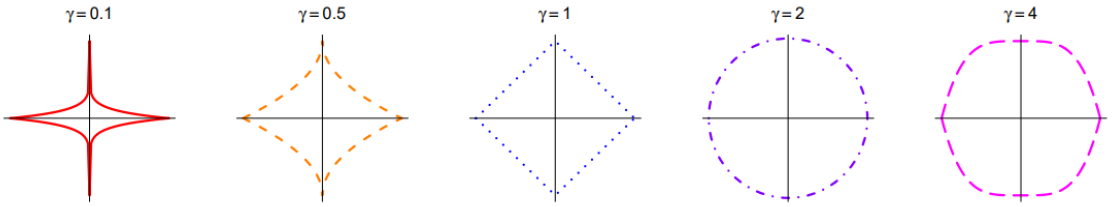

FIGURE 2.1 – Régions de contraintes en fonction de γ pour l'espace de deux paramètres (β_1, β_2). L'axe des abscisses représente les valeurs possible de β_1 et l'axe des ordonnées représente les valeurs possible de β_2

Le coin du UseR : https ://www.rdocumentation.org/packages/lqa/versions/1.0-3/topics/bridge

2.6 ElasticNet

Il peut être intéressant de profiter du partage des poids dans la pénalisation Ridge tout en gardant les capacités parcimonieuses de la régression Lasso. L'ElasticNet proposé par Zou and Hastie [2005] combine la norme L_1 du Lasso et la norme L_2 :

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \sum_j |\beta_j| + \lambda_2 \|\boldsymbol{\beta}\|_2^2 \right\}$$

Lorsque deux variables pertinentes sont corrélées, le lasso aura tendence à sélectionner une seule des deux. L'ajout de la pénalité Ridge favorisera la sélection de l'ensemble. Cependant, la régression ElasticNet peut montrer certaines limites du fait de la nécessité de fixer deux paramètres de pénalisation qui peut mener à de faibles performances en prédiction.

Le coin du UseR : Le package `glmnet` permet d'effectuer une régression de type ElasticNet en donnant une valeur comprise entre 0 et 1 à l'option `alpha` dans la fonction `glmnet`.

2.7 Adaptive-Lasso

Le Lasso a tendance à conserver des variables non associées directement à la réponse mais qui permettent d'améliorer la prédiction. Cependant en terme d'interprétation, il peut être souhaitable de sélectionner uniquement les variables pertinentes (dans un objectif d'explication/identification plutôt que de prédiction). L'Adaptive-Lasso [Zou, 2006]

rajoute un poids sur les variables pertinentes afin d'améliorer la consistance de la procédure :

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_j \omega_j |\beta_j| \right\} \qquad \lambda \geq 0$$

où $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_p)$ est un vecteur de poids qui traduit l'importance, défini tel que $\hat{\boldsymbol{\omega}} = \frac{1}{|\hat{\boldsymbol{\beta}}|^\gamma}$ avec $\gamma > 0$.

Le coin du UseR : Le package `parcor` développé par Kraemer et al. [2009] permet d'effectuer une régression de type Adaptive-Lasso par la fonction `adalasso`. D'autres packages implémentant le Lasso et proposant un argument de poids, tel que `glmnet`, sont également utilisables.

2.8 BoLasso

Dans un objectif d'identification, des méthodes de stabilité de la sélection ont été étudiées, par exemple, la méthode BoLasso [Bach, 2008]. Cette approche perturbe le jeu de données en créant un certain nombre d'échantillons bootstrap. Les variables retenues seront les variables sélectionnées dans l'ensemble d'échantillons bootstrap. Une version moins exigente du Bolasso sélectionne les variables qui ont été retenues avec une grande proportion. Cette proportion est donc un paramètre à régler (tout comme le paramètre de régularisation λ). Les méthodes de bootstrap permettent également de construire des intervalles de confiance des paramètres $\boldsymbol{\beta}$ [Liu et al., 2017]. L'algorithme 1 montre le principe du BoLasso.

Le coin du UseR : Le package `SparseLearner` développé par Guo and Hao [2015] permet d'effectuer du Bolasso par la fonction `Bolasso`.

2.9 GroupLasso

Le Lasso sélectionne des variables de manière individuelle. Cependant, certaines problématiques cliniques peuvent amener à traiter le problème en terme de groupes, comme par exemple les bactéries du microbiome pulmonaire appartenant à la même famille. Dans ce cas, on souhaite sélectionner l'ensemble du groupe de gènes et non qu'un seul gène. Yuan and Lin [2006] ont introduit le GroupLasso qui permet de raisonner en groupes de variables. Cette méthode utilise une pénalité sur le groupe qui permet de soit sélectionner le groupe en entier (i.e. l'ensemble des variables contenues dans le groupe) soit ne pas

Algorithm 1 Algorithme BoLasso

Initialisation :

— Données $(\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^{n\times(p+1)}$

— m : nombre d'échantillons bootstrap

for $k = 1$ **to** m **do**

Génération d'un échantillon boostrap $\left(\mathbf{X}^{(k)}, \mathbf{Y}^{(k)}\right) \in \mathbb{R}^{n\times(p+1)}$

Estimation de $\boldsymbol{\beta}_\lambda^{(k)}$ par une régression Lasso

$$\hat{\boldsymbol{\beta}}_\lambda^{(k)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{n} \|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}\boldsymbol{\beta}^{(k)}\|_2^2 + \lambda \sum_j |\beta_j^{(k)}| \right\}$$

Choix du paramètre de régularisation λ par exemple validation croisée

Calcul du support $J_k = \left\{ j,\ \hat{\boldsymbol{\beta}}_j^{(k)} \neq 0 \right\}$

end for

$J = \bigcap_{k=1}^m J_k$, $\mathbf{X}_J \subset \mathbf{X}$ et $\boldsymbol{\beta}_J$ son J-vecteur de paramètres.

Estimation de $\boldsymbol{\beta}_J$ par une régression non pénalisée

$$\hat{\boldsymbol{\beta}}_J = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{n} \|\mathbf{Y} - \mathbf{X}_J\boldsymbol{\beta}_J\|_2^2 \right\}$$

D'autres possibilités pour le calcul de $\hat{\boldsymbol{\beta}}_J$ sans correction du biais : la médiane ou la moyenne des estimations bootstrap.

le sélectionner (i.e. aucune variable du groupe). Supposons que $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_G)$ est divisé en G groupes de variables, $\mathbf{X}_g$ contient p_g variables et $p = \sum_{g=1}^{G} p_g$, l'estimateur GroupLasso est solution de :

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\mathbf{Y} - \sum_{g=1}^{G} \mathbf{X}_g \boldsymbol{\beta}_g\|_2^2 + \lambda \sum_{g=1}^{G} \|\boldsymbol{\beta}_g\|_2 \right\} \tag{5}$$

avec $\boldsymbol{\beta}_g = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_G)$ les coefficients de régression associés à la matrice $\mathbf{X}_g$. Le problème d'optimisation (5) peut être écrit de façon équivalente :

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\mathbf{Y} - \sum_{g=1}^{G} \mathbf{X}_g \boldsymbol{\beta}_g\|_2^2 + \lambda^* \sum_{g=1}^{G} \sqrt{p_g} \|\boldsymbol{\beta}_g\|_2 \right\} \tag{6}$$

avec $\lambda^* = \lambda / \sqrt{p_g}$ qui permet d'expliciter une pénalisation proportionnelle à (la racine de) la taille du groupe. L'optimum du problème (6) peut être calculé grâce à l'algorithme Group-LARS proposé par Yuan and Lin [2006] et bien d'autres.

Le coin du UseR : Le package `gglasso` développé par Yang and Zou [2013] permet d'effectuer une régression GroupLasso par la fonction `gglasso`.

2.10 Sparse GroupLasso

Le GroupLasso permet de gérer les données en groupe et de les sélectionner par groupe. Cependant, toutes les variables appartenant au groupe ne sont pas forcément associées à la variable clinique étudiée. Simon et al. [2013b] proposent une extension du GroupLasso permettant une sélection de variable de manière individuelle tout en gardant la structure de groupe. Pour cela, une pénalité de norme L_1 de type Lasso est rajoutée au problème (6) :

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\mathbf{Y} - \sum_{g=1}^{G} \mathbf{X}_g \boldsymbol{\beta}_g\|_2^2 + \lambda_1 \sum_{g=1}^{G} \sqrt{p_g} \|\boldsymbol{\beta}_g\|_2 + \lambda_2 \sum_{j} |\beta_j| \right\} \tag{7}$$

Lorsque $\lambda_1 = 0$, on retrouve la pénalité Lasso et lorsque $\lambda_2 = 0$, on retrouve la pénalité GroupLasso.

Le coin du UseR : Le package `SGL` développé par Simon et al. [2013a] permet d'effectuer une régression sparse GroupLasso par la fonction `SGL`.

2.11 Extension à d'autres fonctions de perte

Les pénalisations présentées ci-dessus peuvent être appliquées à d'autres fonctions de perte (« loss function », en anglais, fonction qui mesure la différence entre la réponse Y

et βX fournie par la machine d'apprentissage) telle que la fonction quadratique :

$$\min_{\boldsymbol{\beta}} \left\{ l\left(\mathbf{Y}, \mathbf{X}, \boldsymbol{\beta}\right) + \lambda \sum_{j} |\beta_j| \right\} \tag{8}$$

où l représente n'importe quelle fonction de perte, comme par exemple la log-vraisemblance d'un modèle logistique.

3 Régression pénalisée de type Lasso pour données censurées par une limite de quantification

Valorisation : Ce chapitre a été valorisé par un article dans *BMC Medical research Methodology*

3.1 Introduction

3.1.1 La censure par limite de détection

La censure due à un défaut de la sensibilité de l'appareil de mesure est une problématique commune dans de nombreux domaines d'application comme la biologie, la chimie et l'environnement. La « vraie » valeur n'est pas connue, on dit qu'elle est « indétectable ». En statistique, on parle de données censurées à gauche par limite de quantification. La valeur censurée est une valeur biaisée que l'on doit prendre en compte dans les analyses. L'objectif statistique va être de se rapprocher le plus possible de la « vraie » valeur pour obtenir des résultats valides. Des approches naïves sont appliquées pour s'affranchir du problème de la censure, telles que la substitution de la valeur censurée par une autre valeur, classiquement la limite de détection (LOD, pour limit of detection) ou la moitié de cette valeur (LOD/2) pour s'affranchir du problème de la censure. Cependant, ces méthodes engendrent des biais importants dans l'estimation des paramètres. Plusieurs approches statistiques ont été proposées pour tenir compte de la censure à gauche, dans des études transversales (avec une mesure par sujet) mais également longitudinales (avec plusieurs mesures par sujet). Les méthodes standards incluent l'imputation multiple [Paxton et al., 1997; Helsel, 2005; Lee et al., 2012; Del Greco M. et al., 2016], les méthodes de survie en renversant le problème de la censure à droite [Marschner et al., 1999; Helsel, 2005; Gillespie et al., 2010; Dinse et al., 2014], la régression quantile [Wang et al., 2009; Eilers et al., 2012] ou la régression quantile censurée [Powell, 1984, 1986].

Le modèle Tobit est un modèle statistique utilisé pour décrire une relation entre une

variable dépendante censurée et une variable indépendante [Tobin, 1958]. Il peut être estimé par maximum de vraisemblance [Tobin, 1958; Hughes, 1999; Jacqmin-Gadda et al., 2000; Lynn, 2001; Nie et al., 2010; Fu et al., 2016; Wiegand et al., 2016] ou par l'algorithme itératif de Buckley-James [Buckley and James, 1979; Wiegand et al., 2016]. Ce dernier propose d'imputer les valeurs censurées à droite par une espérance conditionnelle basée sur l'estimateur de Kaplan-Meier. Si $\mathbf{Z}$ une variable censurée à gauche alors $-\mathbf{Z}$ est une variable censurée à droite.

3.1.2 Le contexte du VIH

Le VIH est un retrovirus pouvant causer le SIDA (syndrome d'immunodéficience acquise). L'infection par le VIH atteint le système immunitaire, c'est-à-dire les défenses naturelles du corps contre les infections. ette destruction du système immunitaire contribue à la survenue d'affections opportunistes et au SIDA (syndrome d'immunodéficience acquise). Depuis 1996, la thérapie antirétrovirale combinée (combination antiretroviral therapy, cART), associant au moins trois médicaments antirétroviraux (classe de médicaments utilisés pour le traitement des infections liées aux rétrovirus), est le traitement de référence pour les patients infectés par le VIH. Les molécules disponibles sont divisées en six classes : les inhibiteurs nucléosidiques de la transcriptase inverse, les inhibiteurs non nucléosidiques de la transcriptase inverse, les inhibiteurs de la protéase, les inhibiteurs de fusion, les inhibiteurs de l'intégrase et les inhibiteurs de chémokine co-recepteur 5 (CCR5). De nos jours, la thérapie antirétrovirale n'est pas capable d'éradiquer entièrement le virus. Le but principal de cette thérapie est d'empêcher une progression de l'infection. Pour cela, la thérapie antirétrovirale a comme objectif et de maintenir la charge virale le plus bas possible, c'est à dire au dessous de ce seuil de détection (le plus souvent 20 copies/mL actuellement) aussi longtemps que possible [on Antiretroviral Guidelines for Adults and Adolescents, 2018]. Les échecs thérapeutiques (ré-augmentation de la charge virale au dessus du seuil de détection) peuvent être liés à différentes causes (défaut d'adhérence, concentration plasmatique insuffisante, par exemple). La conséquence directe de ces échecs est une suppression incomplète de la réplication du virus. Cette dernière facilite le développement du virus ainsi que sa résistance aux traitements antirétroviraux : c'est le résultat d'une sélection de mutations génotypiques qui permet au virus de se répliquer en présence du traitement.

Deux types de test sont utilisés pour déterminer la résistance antirétrovirale. En pratique : les tests génotypiques et les tests phénotypiques. Les premiers consistent en un séquençage des gènes du VIH concernés (protéase, par exemple) pour détecter des mutations conduisant à une résistance antirétrovirale. Les tests phénotypiques, quant à eux,

mesurent la réplication du virus en culture cellulaire en présence ou absence d'une molécule antirétrovirale donnée. Ces tests révèlent généralement un nombre conséquent de mutations. Les analyses statistiques visent à identifier les mutations génotypiques, observées lors des tests, associées à la réponse virologique.

3.1.3 Limite de quantification dans l'étude des mutations génotypiques

Les analyses statistiques traditionnelles étudiant l'association entre les mutations génotypiques et la réponse virologique font face à des obstacles : i) le grand nombre de mutations et leurs fortes corrélations entre elles, ii) le faible nombre de patients disponibles et iii) la censure de la réponse virologique. L'analyse de l'effet d'un nombre élevé de mutations mesurées chez un nombre limité de patients peut conduire à des problèmes de sur-ajustement ou, au contraire, de manque de puissance. Pour simplifier l'interprétation, les mutations peuvent être résumées par un score génotypique. Ce score est caractérisé comme la somme des mutations résistantes observées lors du test génotypique pour un patient donné. La composition du score peut être définie par plusieurs stratégies [Brun-Vezinet et al., 2004; Flandre et al., 2005]. Les mutations sont pré-sélectionnées, ce qui peut engendrer une perte d'information. L'analyse en composantes principales (PCA) et la *Partial Least Square* (PLS) ont également été utilisées pour prendre en compte les fortes corrélations entre les mutations [Wittkop et al., 2008]. Cependant, ces deux méthodes n'ont pas montré de performances nettement meilleures en prédiction que le score génotypique. Ces approches reposent sur le modèle logistique. En effet, la charge virale n'est pas considérée comme une variable quantitative mais comme une variable binaire (deux groupes de patients) : la valeur 0 correspondant au succès thérapeutique (la charge virale est inférieure au seuil de détection) et la valeur 1 correspondant à un échec thérapeutique (la charge virale est supérieure au seuil de détection). Dans ce contexte, deux patients ayant des charges virales très proches mais séparées par la limite de détection seront dans des groupes différents. De plus, deux patients avec une charge virale supérieure au seuil de détection peuvent avoir des valeurs très différentes (eg 1000 ou 100000 copies/mL), ce qui peut être très différent sur le plan clinique. La dichotomisation d'une variable quantitative engendre, en général, une perte d'information et de puissance du modèle. L'idée première de ce travail a donc été de conserver le caractère quantitatif de la charge virale.

Une première approche serait d'adapter le modèle Tobit à la grande dimension. La pénalité de type Lasso appliquée à une fonction de vraisemblance est une technique standard dans ce contexte. Cependant, lorsque la fonction de vraisemblance est complexe telle

que le modèle Tobit, l'optimisation devient un réel challenge computationnel.

L'algorithme itératif de Buckley-James, imputant les valeurs censurées à droite par une espérance conditionnelle basée sur l'estimateur de Kaplan-Meier, a déjà été adapté à la grande dimension dans le cas d'analyse de survie. Des extensions de la régression PLS [Huang et al., 2004] et de la régression Lasso ont été proposées pour la censure à droite [Johnson, 2008, 2009b; Cai et al., 2009; Wang and Wang, 2010]. Plus récemment, Wiegand et al. [2016] a appliqué l'algorithme à un cas de censure à gauche dans une étude transversale et de petite dimension. Nos travaux consistent à proposer une version Lasso de la méthode de Buckley-James. Dans un premier temps, en renversant la variable réponse pour se ramener à de la censure à droite et garder la version non-paramétrique. Et dans un second temps, en adaptant l'estimateur de Buckley-James de manière paramétrique en se basant sur une hypothèse sous-jacente de la distribution gaussienne du logarithme en base décimale de la charge virale plasmatique.

3.1.4 Application aux données issues du Forum pour la recherche collaborative contre le VIH

Ce travail a été effectué en collaboration avec Linda Wittkop, responsable de l'équipe « VIH, Hépatites virales et co-morbidités : épidémiologie clinique et santé publique (INSERM U1219) » et associée avec *The Forum Collaborative HIV Research.*

Fondé en 1997, *The Forum Collaborative HIV Research* (HIV Med. 2008 ;7 :27-40). est un partenariat (privé/public) entre l'université de Californie et le campus Berkeley Washington. La mission de ce forum est d'améliorer et de faciliter la recherche contre le VIH. Ceci est accompli en réunissant tous les acteurs pertinents, parties prenantes (gouvernement, industrie, défenseurs des droits des patients, fournisseurs de soins de santé, universités et fondations) pour faire face aux problèmes émergents en matière de VIH. Ce travail s'introduit plus particulièrement dans le groupe de travail sur la pharmacorésistance (*Standardization and Clinical Relevance of HIV Drug Resistance Testing Project*). Le groupe s'intéresse à la comparaison de méthodes quantitatives d'analyse des données de résistance génotypique [Schumi and DeGruttola, 2008; Yang and DeGruttola, 2008; Healy et al., 2008]. Les données contiennent des observations de patients sous traitement ayant commencé le traitement par abacavir et d'autres antirétroviraux ayant un génotype initial [Cozzi-Lepri, 2008]. Le médicament investigué, abacavir, est un inhibiteur nucléosidique de la transcriptase inverse qui bloque la transcriptase inverse du VIH.

Les données à disposition représentent une centaine de patients et répertorient un peu plus de 121 mutations. Ces mutations sont probablement liées à une résistance à un ou plusieurs types de traitements, selon les connaissances au moment de l'étude [Johnson

et al., 2009]. Sur les 121 mutations, 54 correspondent à la présence de mutations spécifiques dans le gène de la transcriptase inverse. Ces dernières seraient probablement liées à la résistance de l'abacavir. Les 67 autres mutations correspondent à la présence des mutations spécifiques au gène de la protéase sûrement associées à la résistance de certains inhibiteurs de la protéase au moment de l'étude [Shafer and Schapiro, 2008; Johnson et al., 2009]. La variable réponse étudiée est la log-charge virale mesurée à la huitième semaine après l'administration du traitement. La limite de détection était de fixer à 100 copies/mL et le taux de censure est modéré (26%).

3.2 Article : Lasso regularization for left-censored Gaussian outcome and high-dimensional predictors

Contexte : Les tests biologiques pour la quantification des marqueurs peuvent être soumis à une limite de détection analytique. C'est le cas de la charge virale du virus de l'immunodéficience humaine (VIH). En dessous de ce seuil, la valeur exacte est inconnue et les valeurs sont par conséquent censurées à gauche. Des méthodes statistiques ont été proposées pour traiter la censure à gauche, mais peu sont adaptées dans le contexte de données de grande dimension.

Méthodes : Nous proposons d'inverser l'algorithme des moindres carrés de Buckley-James pour traiter les données censurées à gauche et d'intégrer une régularisation de type Lasso permettant s'attaquer à des problèmes de grande dimension. Nous introduisons un algorithme impliquant deux types d'imputation : une version non paramétrique avec un estimateur de Kaplan Meier et une méthode paramétrique basée sur une distribution Gaussienne. Pour régler les paramètres de régularisation, la validation croisée est basée sur une fonction de perte appropriée, qui prend en compte de façon différenciée les contributions d'observations censurées et non censurées.

Résultats : Dans un contexte de grande dimension, les approches tenant compte de la censure sont plus performantes que les méthodes d'imputation simple. Notre algorithme, couplé à une validation croisée basée sur une fonction de perte appropriée, a montré l'erreur de prédiction la plus faible sur des données simulées. De plus, lorsque appliqué à de données réelles, l'algorithme a montré les résultats les plus cohérents avec littérature.

Conclusions : L'approche proposée traite simultanément des problèmes de grande dimension et de données censurés à gauche par une limite de détection. Elle a montré son intérêt pour la prédiction de la charge virale du VIH en fonction des mutations du VIH.

Soret *et al. BMC Medical Research Methodology* (2018) 18:159
https://doi.org/10.1186/s12874-018-0609-4

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Lasso regularization for left-censored Gaussian outcome and high-dimensional predictors

Perrine Soret[1,2,3], Marta Avalos[1,2*], Linda Wittkop[1,2,4], Daniel Commenges[1,2] and Rodolphe Thiébaut[1,2,3,4]

Abstract

Background: Biological assays for the quantification of markers may suffer from a lack of sensitivity and thus from an analytical detection limit. This is the case of human immunodeficiency virus (HIV) viral load. Below this threshold the exact value is unknown and values are consequently left-censored. Statistical methods have been proposed to deal with left-censoring but few are adapted in the context of high-dimensional data.

Methods: We propose to reverse the Buckley-James least squares algorithm to handle left-censored data enhanced with a Lasso regularization to accommodate high-dimensional predictors. We present a Lasso-regularized Buckley-James least squares method with both non-parametric imputation using Kaplan-Meier and parametric imputation based on the Gaussian distribution, which is typically assumed for HIV viral load data after logarithmic transformation. Cross-validation for parameter-tuning is based on an appropriate loss function that takes into account the different contributions of censored and uncensored observations. We specify how these techniques can be easily implemented using available R packages. The Lasso-regularized Buckley-James least square method was compared to simple imputation strategies to predict the response to antiretroviral therapy measured by HIV viral load according to the HIV genotypic mutations. We used a dataset composed of several clinical trials and cohorts from the Forum for Collaborative HIV Research (HIV Med. 2008;7:27-40). The proposed methods were also assessed on simulated data mimicking the observed data.

Results: Approaches accounting for left-censoring outperformed simple imputation methods in a high-dimensional setting. The Gaussian Buckley-James method with cross-validation based on the appropriate loss function showed the lowest prediction error on simulated data and, using real data, the most valid results according to the current literature on HIV mutations.

Conclusions: The proposed approach deals with high-dimensional predictors and left-censored outcomes and has shown its interest for predicting HIV viral load according to HIV mutations.

Keywords: Limit of detection, Buckley-James least squares procedure, HIV viral load, Drug resistance, HIV genotypic mutations, Cross-sectional studies

*Correspondence: marta.avalos-fernandez@u-bordeaux.fr
[1]Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, UMR 1219, F-33000 Bordeaux, France
[2]Inria SISTM Team, F-33405 Talence, France
Full list of author information is available at the end of the article

© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Background

Left-censoring due to the lower detection limit of an assay is a common problem in many fields including biology, chemistry, and the environmental sciences. One example is the quantification of the human immunodeficiency virus (HIV) viral load in plasma. The sensitivity of assays has improved and the detection threshold has decreased from 10,000 copies/mL to 20 or fewer copies/mL today. Several statistical methods have been proposed to account for left-censoring of such quantitative variables in cross-sectional (with one measure per subject) and longitudinal (with several measures per subject) studies. Standard methods include multiple imputation [1–4], reverse survival analysis methods [2, 5–7], quantile regression [8, 9] and censored quantile regression [10, 11]. Furthermore, the Tobit model with censored outcome which is supposed to be normally distributed can be estimated by maximum likelihood [12–18] or by the Buckley-James estimator [18, 19]. Indeed, HIV viral load appears to have an underlying Gaussian distribution truncated by the detection limit that justifies the normality hypothesis [13–15, 17, 18]. As expected, approaches accounting for left-censoring outperform simple imputation of a constant [2, 4, 13–16, 18, 20–22].

Another issue may arise when the number of predictors (p) is high compared to the number of statistical units (n), without excluding the possibility that $n < p$. This is known as high dimensionality. In the context of HIV infection, this can be illustrated by analyzing the association between the presence of HIV mutations and the response to antiretroviral therapy which is measured by HIV viral load. HIV strains circulating in a given individual can present mutations associated with antiretroviral treatment failure (detectable HIV viral load), also called HIV drug resistance mutations. Thus, genotypic tests allowing the detection of HIV drug resistance mutations are commonly performed in patients starting a new antiretroviral regimen or even in newly HIV-infected patients because of the transmission of resistant strains [23–26]. Lasso linear [27, 28] and logistic regressions [29], principal component and partial least square logistic regressions [30], and multiple testing correction [31] have been used to deal with more than 100 predictors and fewer than a few hundred of patients, a common situation in this context [32].

These studies use a dichotomized outcome or simple imputation by a constant to circumvent the problem of censoring. One limitation of dichotomizing a continuous outcome is the loss of information and hence power. In addition, success is usually defined as achieving an undetectable HIV viral load. However, the detection limit, although not random, depends on several factors that differ from one study to another. Thus, there is no reason, except convenience, for the detection limit to correspond to the threshold for dichotomization.

We hypothesize that approaches accounting for left-censoring will exhibit better results compared to simple imputation strategies in a high-dimensional setting similar to what has been found in low-dimensional settings.

Some works have simultaneously addressed both censoring and high-dimensional problems using the Lasso [33–43], partial least squares [44], random forests [45], support vector machines [46], and deep learning [47]. These examples were developed for right-censored survival data. A main approach to left-censored data analysis is based on methods typically used with right-censored survival data such as the Buckley-James estimator. Left-censored data are then previously reversed to right-censored data. While from a statistical point of view, the nature of the outcome (time-to-event or quantitative measurement below a limit of detection) is secondary, this can impact the choice of adequate probability distribution functions and other practical issues.

We propose a Lasso-regularized Buckley-James least squares method with both, non-parametric imputation using Kaplan-Meier and parametric imputation based on the Gaussian distribution. The non-parametric Buckley–James estimator, which simply replaces censored residuals by their conditional expectations in an iterative way, has been previously applied to left-censored HIV viral load data in a cross-sectional study [18]. On the other hand, the Lasso extension of the non-parametric Buckley–James method has been proposed for right-censored data [36, 38, 40, 48]. Our contribution consists in using the latter method for left-censored outcomes and high-dimensional predictors. Furthermore, we propose an original parametric version of the Buckley-James method, which is adapted to the typical assumption of a Gaussian distribution of HIV viral load. We demonstrate the value of these approaches by comparing them to Lasso linear regression with simple imputation [28] for predicting the response to antiretroviral therapy by HIV genotypic mutations.

Our primary objective is to predict as accurately as possible responses in future patients who will switch to a similar regimen. Thus, comparisons are based on mean square prediction error. The prediction performances of the different methods were assessed on simulated data that reproduced the observed data. Then, methods were applied to data obtained in a collaborative study from clinical trials and cohorts provided by the Standardization in Clinical Relevance of HIV Drug Resistance Testing Project from the Forum for Collaborative HIV Research [49]. The actual data presented a moderate censoring rate of 26 %, i.e. a realistic magnitude [18, 50]. However, high (around 50 %) or even severe (around 70 %) censoring

rates could be observed in older studies with a high limit of detection (LOD) or particular populations with low treatment failure rate, e.g. HIV controllers [18, 51, 52]. Thus, we also explored the impact of high and severe censoring rates on performance.

We detail how to use publicly available R packages to compute Lasso estimates with left-censored data.

Finally, we discuss possible extensions and applications of our work.

Methods

Methods to analyze left-censored outcome

In this section, we review the simplest models and estimation methods used to deal with left-censoring in cross-sectional studies. For a more extensive and comprehensive review of these methods, see [2, 18]. Thereafter, we consider the Lasso extension of those methods that support simple implementations.

The linear model

First, consider the general linear regression model

$$Y_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \cdots, n \tag{1}$$

where $\mathbf{X}_i$ is a p-vector of fixed predictors, Y_i is the uncensored continuous random outcome variable, $\boldsymbol{\beta} = \left(\beta_1, \cdots, \beta_p\right)^\top$ is a p-vector of unknown regression parameters and ε_i are independent and identically normally distributed random variables with mean 0 and constant variance σ^2. Let $\mathbf{X}$ be the $n \times p$ matrix $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top$ and $\mathbf{Y}$ the $n \times 1$ vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$. The intercept is omitted in the model for simplicity, and all predictor variables are assumed to be standardized (i.e. zero mean and unit variance).

Lasso on complete data

The Lasso (Least Absolute Shrinkage and Selection Operator) [53] is one of the most popular methods in high-dimensional data analyses. It allows for simultaneous estimation and variable selection and has efficient algorithms available. It is considered here as the *Gold Standard* for our simulation studies. The Lasso estimator of parameters in model (1) is:

$$\hat{\boldsymbol{\beta}}(\lambda)_{GoldS} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right\|_2^2 + \lambda \left\|\boldsymbol{\beta}\right\|_1 \tag{2}$$

where $\left\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right\|_2^2 = \sum_{i=1}^{n} (Y_i - \mathbf{X}_i\boldsymbol{\beta})^2$ is the quadratic loss, $\left\|\boldsymbol{\beta}\right\|_1 = \sum_{j=1}^{p} \left|\beta_j\right|$ is the Lasso penalty on the parameter size, and $\lambda > 0$ controls the amount of regularization. When λ is large enough (which depends on data), all coefficients are forced to be exactly zero. Inversely, $\lambda = 0$ corresponds to the unpenalized ordinary least-squares estimate.

This model on complete data (no left-censored measures) is considered as a reference when comparing the other methods applied to incomplete datasets that include left-censored values.

The Tobit model

Because of the detection limit, Y_i can be left-censored. Let LOD_i be the (fixed and known) censoring threshold of subject i. To simplify, we consider $\text{LOD}_i = \text{LOD}$. Z_i is the observed response. The so-called Tobit model [12] can be defined as:

$$Z_i = \begin{cases} Y_i & \text{if} \quad Y_i > \text{LOD} \\ \text{LOD} & \text{if} \quad Y_i \leq \text{LOD} \end{cases} \tag{3}$$

where Y_i is the response variable defined in model (1). We can equivalently write:

$$Z_i = \max\left(Y_i, \text{LOD}\right), \text{ or } Z_i = \delta_i Y_i + (1 - \delta_i)\,\text{LOD}, \tag{4}$$

where $\delta_i = \mathbb{I}_{(Y_i > \text{LOD})}$ is a censoring indicator. The idea behind the Tobit regression model is to deal with the left-censored variable Z as the outcome of a normally distributed latent variable Y.

Simple imputation

Simple imputation is a substitution method that replaces left-censored values with a single value, LOD. LOD/2 is another common choice. Let $\hat{\boldsymbol{\beta}}_{LOD}$ be the ordinary least squares estimate of model:

$$Z_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \cdots, n \tag{5}$$

Simple imputation is widely used for its simplicity. However, replacing any censored observation by a single value may lead to biased parameter estimates.

Beerenwinkel et al. [28] applied Lasso-regularized linear regression with the naïve approach of replacing the unobserved undetectable value with the limit of detection of the assay. Then the Lasso estimator of parameters in model (5) is:

$$\hat{\boldsymbol{\beta}}(\lambda)_{LOD} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\|\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}\right\|_2^2 + \lambda \left\|\boldsymbol{\beta}\right\|_1. \tag{6}$$

Maximum likelihood estimation

In the Tobit model (1)-(3), one can assume that when $Z = \text{LOD}$, the density function of Z is equal to the probability of observing $Y \leq \text{LOD}$ and for $Z > LOD$ the density function of Z is the same as the density of Y. The likelihood function takes the form:

$$\mathcal{L}\left(\boldsymbol{\beta}, \sigma^2\right) = \prod_{i=1}^{n} \mathbb{P}(Y_i \,|\, Y_i > \text{LOD}, \mathbf{X}_i)^{\delta_i}\, \mathbb{P}(Y_i \,|\, Y_i \leq \text{LOD}, \mathbf{X}_i)^{1-\delta_i}$$

When the Gaussian distribution for the outcome is assumed, the log-likelihood function can be written as:

$$\ln\mathcal{L}\left(\boldsymbol{\beta},\sigma^2\right)=\sum_{i=1}^{n}\delta_i \ln f_G\left(Y_i,\mathbf{X}_i,\boldsymbol{\beta},\sigma^2\right) + (1-\delta_i)\ln F_G\left(\text{LOD},\mathbf{X}_i,\boldsymbol{\beta},\sigma^2\right) \quad (7)$$

with $f_G\left(u,\mathbf{X}_i,\boldsymbol{\beta},\sigma^2\right) = \frac{e^{\frac{(u-\mathbf{X}_i\boldsymbol{\beta})^2}{-2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$ the Gaussian probability density function of Y_i with mean $\mathbf{X}_i\boldsymbol{\beta}$ and constant variance σ^2 evaluated at u and $F_G\left(v,\mathbf{X}_i,\boldsymbol{\beta},\sigma^2\right) = \int_{-\infty}^{v} f_G\left(u,\mathbf{X}_i,\boldsymbol{\beta},\sigma^2\right)\mathrm{d}u$ is the corresponding Gaussian cumulative distribution function evaluated at v. Let $\hat{\boldsymbol{\beta}}_{MLE}$ be the maximum likelihood estimation obtained by maximizing (7). Extensions to other distributions have also been explored [54]. Several works have shown the superiority of this method [18, 20–22]. However, when the parametric model is misspecified, the sample size is small or the percent censoring is high, the maximum-likelihood estimation method has been shown to perform poorly [2].

The Lasso penalty applied to some likelihood function has become an established and relatively standard technique. However, when the likelihood function is a more complex function of the model parameter, such as the likelihood function for the Tobit model (7), adding a non-differentiable penalty leads to a computational challenging optimization.

Quantile regression and censored quantile regression

Quantile regression, particularly least absolute deviations (LAD) regression, has been applied to left-censored data [9]:

$$\hat{\boldsymbol{\beta}}_{LAD} = \underset{\boldsymbol{\beta}}{\text{argmin}}\ \frac{1}{n}\sum_{i=1}^{n}|Z_i - \mathbf{X}_i\boldsymbol{\beta}|$$

Median regression is a natural alternative to the usual mean regression in the presence of heteroscedasticity or when the normality assumption is violated. Simple imputation using robust regression may be less sensitive to the influence of censored observations.

Lasso-regularized least absolute deviations regression has been investigated in the literature (e.g. [55]).

Powell [10] proposed the LAD estimate specifically for censored data:

$$\hat{\boldsymbol{\beta}}_{CLAD} = \underset{\boldsymbol{\beta}}{\text{argmin}}\ \frac{1}{n}\sum_{i=1}^{n}|Z_i - \max\{\text{LOD},\mathbf{X}_i\boldsymbol{\beta}\}|$$

Later, the approach was extended to more general quantiles [11].

The Lasso extension of censored quantile regression (basically, censored LAD) has been analyzed for right-censored survival data [41, 42, 56] and specifically for left-censored data [57–62]. Yu et al. [58] and Alhamzawi et al. [61] proposed Bayesian approaches using different hyperparameters priors. These methods rely on computationally intensive algorithms. In practice, applications are limited to the $n >> p$ case. Others [57, 59, 60, 62] derived theoretical properties of the Lasso-regularized censored least absolute deviations regression, but the algorithmic development was not a priority in these works and the practical use was limited to $n >> p$ or not addressed. To our knowledge, there are no publicly available software tools that implement the Lasso extension of Powell's approach and no simple implementation relying on existing packages seems straightforward.

Non-parametric Buckley-James

Left-censored outcome data can be analyzed using methods designed for right-censored survival data by reversing the outcome scale. For instance, Gillespie et al. [6] proposed the reverse Kaplan-Meier and Dinse et al. [7] reversed the Cox method (though in the case of left-censored exposures and uncensored outcome). After the Cox model, the accelerated failure time model is the most frequently used regression model for right-censored survival data. It directly links the expected response to predictors, analogously to the classical linear regression approach. A popular method for fitting the accelerated failure time model is the Buckley-James estimator [19], an extension of the least squares principle. The idea is to impute the censored values by their estimated conditional mean to provide censoring and predictor values:

$$Z_i^* = \delta_i Y_i + (1-\delta_i)\,\mathbb{E}\left(Y_i\,|Y_i \le \text{LOD},\mathbf{X}_i\right) \quad (8)$$

with

$$\mathbb{E}\left(Y_i\,|Y_i \le \text{LOD},\mathbf{X}_i\right) = \int_{-\infty}^{\text{LOD}} \frac{uf\left(u,\mathbf{X}_i,\boldsymbol{\beta}\right)\mathrm{d}u}{F\left(\text{LOD},\mathbf{X}_i,\boldsymbol{\beta}\right)}$$

where $f\left(u,\mathbf{X}_i,\boldsymbol{\beta}\right)$ is the (unknown) probability density function of Y_i with mean $\mathbf{X}_i\boldsymbol{\beta}$ evaluated at u and $F\left(u,\mathbf{X}_i,\boldsymbol{\beta}\right)$ is the corresponding cumulative distribution function. By "flipping" the data (turning it from left-censored to right-censored), the application of algorithms previously developed for right-censoring is direct and has been performed in other contexts [5]. We consider M an arbitrary constant that equals or exceeds the largest observation. Then subtract all uncensored and left-censored outcomes from M. The left-censored at LOD variable $\mathbf{Z}$ is then replaced by $M - \mathbf{Z}$ which is right-censored at $M - \text{LOD}$. Let $(M - Z_i)^*$ be imputed as $\delta_i\,(M - Y_i) + (1-\delta_i)\,\mathbb{E}\left(M - Y_i\,|M - Y_i \ge M - \text{LOD},\mathbf{X}_i\right)$. Then, we can calculate the conditional expectation by

$$\mathbb{E}\left(M - Y_i\,|M - Y_i \ge M - \text{LOD},\mathbf{X}_i\right) = \int_{M-\text{LOD}}^{\infty} \frac{uf\left(u,\mathbf{X}_i,\boldsymbol{\beta}\right)\mathrm{d}u}{1 - F\left(M - \text{LOD},\mathbf{X}_i,\boldsymbol{\beta}\right)} \quad (9)$$

where $F(u, \mathbf{X}_i, \boldsymbol{\beta})$ is now the (unknown) cumulative distribution function of $M - Y_i$ with mean $M - \mathbf{X}_i\boldsymbol{\beta}$ evaluated at u, which can be estimated, for example, by Kaplan-Meier. The Buckley-James estimate, $\hat{\boldsymbol{\beta}}_{NonParBJ}$, can be computed using a semiparametric iterative algorithm that alternates between imputation of censored values according to (9) and least-squares estimation.

The main drawback of this method is that convergence of the algorithm is not guaranteed. Due to the discontinuous nature of the estimating function (formulation (8) makes $\hat{\boldsymbol{\beta}}_{NonParBJ}$ to be a piecewise linear function in $\boldsymbol{\beta}$), the iterative procedure may oscillate between different parameter values. The problem is of practical importance in situations where the effect of predictors is small or in small samples [63] (which could be worse in high-dimensional settings). To circumvent this problem, a one-step algorithm that stops at the first iteration is used in some works [36, 64]. This approach is close to a substitution method in which values below the detection limit are replaced by expected values of the missing measurements, provided they are less than the detection limit [65].

Several authors have proposed combining the iterative Buckley-James imputation and methods handling high-dimensional predictors: Johnson et al. [36, 48] and Cai et al. [38] used the Lasso, Wang et al. [37], used boosting, Wang et al. [40] used ElasticNet, Johnson et al. [66] and Li et al. [67] used the Dantzig selector, and Dirienzo et al. [68] used parsimonious covariate selection. The Buckley-James estimate can be computed using an iterative algorithm that alternates between imputation of censored values according to (9) and the Lasso:

$$\hat{\boldsymbol{\beta}}(\lambda)_{NonParBJ} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\| (M-\mathbf{Z})^* - (M-\mathbf{X}\boldsymbol{\beta}) \right\|_2^2 + \lambda \left\| \boldsymbol{\beta} \right\|_1 \tag{10}$$

Gaussian Buckley-James

Alternatively, the Buckley-James imputation (8), assuming the logarithm of HIV viral load follows a Gaussian distribution, can be calculated with the conditional expectation:

$$\mathbb{E}\left(Y_i \,|\, Y_i \leq \mathrm{LOD}, \mathbf{X}_i\right) = \int_{-\infty}^{\mathrm{LOD}} \frac{u f_G\left(u, \mathbf{X}_i, \boldsymbol{\beta}, \sigma^2\right) \mathrm{d}u}{F_G\left(\mathrm{LOD}, \mathbf{X}_i, \boldsymbol{\beta}, \sigma^2\right)} \tag{11}$$

where f_G and F_G are the Gaussian density and cumulative distribution functions defined in (7). Again, the solution can be computed by iteratively alternating between imputation based on (11) and parameter estimation using ordinary least squares, $\hat{\boldsymbol{\beta}}_{GaussBJ}$, or the Lasso:

$$\hat{\boldsymbol{\beta}}(\lambda)_{GaussBJ} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\| \mathbf{Z}^* - \mathbf{X}\boldsymbol{\beta} \right\|_2^2 + \lambda \left\| \boldsymbol{\beta} \right\|_1 \tag{12}$$

Graphical illustration in a low-dimensional setting

To illustrate the difference between estimation methods we generated data from the simple linear model ($p = 1$): $Y_i = X_i\beta + \varepsilon_i$ with $i = 1, \cdots, n, \mathbf{X} \sim N(0,1), \boldsymbol{\varepsilon} \sim N(0, \sigma^2)$. β was set to 10 and σ^2 was chosen such that the signal-to-noise ratio was 4:3. A limit of detection was then fixed to obtain the desired censoring rate: moderate, 20 %, high, 50 % or severe, 70 %.

In Fig. 1, predicted regression lines are obtained using different methods: the true model that generated the data, the gold standard (ordinary least squares with uncensored data), maximum likelihood estimation (MLE), which is identical to the Gaussian Buckley-James estimation (BJ) when $p = 1$, non-parametric Buckley-James (BJ), least absolute deviations (LAD) and censored LAD regressions, simple imputation by the limit of detection (LOD) and by LOD/2.

Notice that simple imputation by LOD and LOD/2 are the most distant regression lines from the true and gold standard lines, in an opposite way: simple imputation by LOD tends to overestimate the response values while simple imputation by LOD/2 tends to underestimate them.

Maximum likelihood estimation shows one of the best behaviors, but the computational complexity dramatically increases with p (results not showed). Mean and median regressions with simple imputation by LOD are quite close and are the closest when the estimation situation is easy (high n, low censoring rate). The censored LAD shows better results than censored mean regression (MLE and Buckley-James) for small sample size while the inverse is observed when $n = 100$. Gaussian Buckley-James and MLE are identical, but their differences increase when p increases (results not showed). In this i.i.d. generated from a Gaussian distribution example, the Gaussian Buckley-James estimate shows better behavior than non-parametric Buckley-James, the difference being higher when n is small.

Tuning parameter selection

K-fold cross-validation is routinely applied to select the optimal regularization parameter when the main goal of the study is prediction. Data $\mathbf{D}$ is randomly chunked into K disjoint blocks of approximately equal size. To avoid a potentially unbalanced partition, we consider stratified K-fold cross-validation, i.e. each fold contains roughly the same proportion of censoring as in the whole sample. $\mathbf{D}_{\setminus k}$ is the learning data, used to estimate coefficients. $\mathbf{D}_k$ is the test data, not used in the estimation process and then used to evaluate the loss function L. This K-fold cross-validation can be written as:

$$\mathrm{CV}(\lambda) = \frac{1}{K} \sum_{k=1}^{K} L\left(\hat{\boldsymbol{\beta}}(\lambda)_{\mathbf{D}_{\setminus k}}, \mathbf{D}_k\right) \tag{13}$$

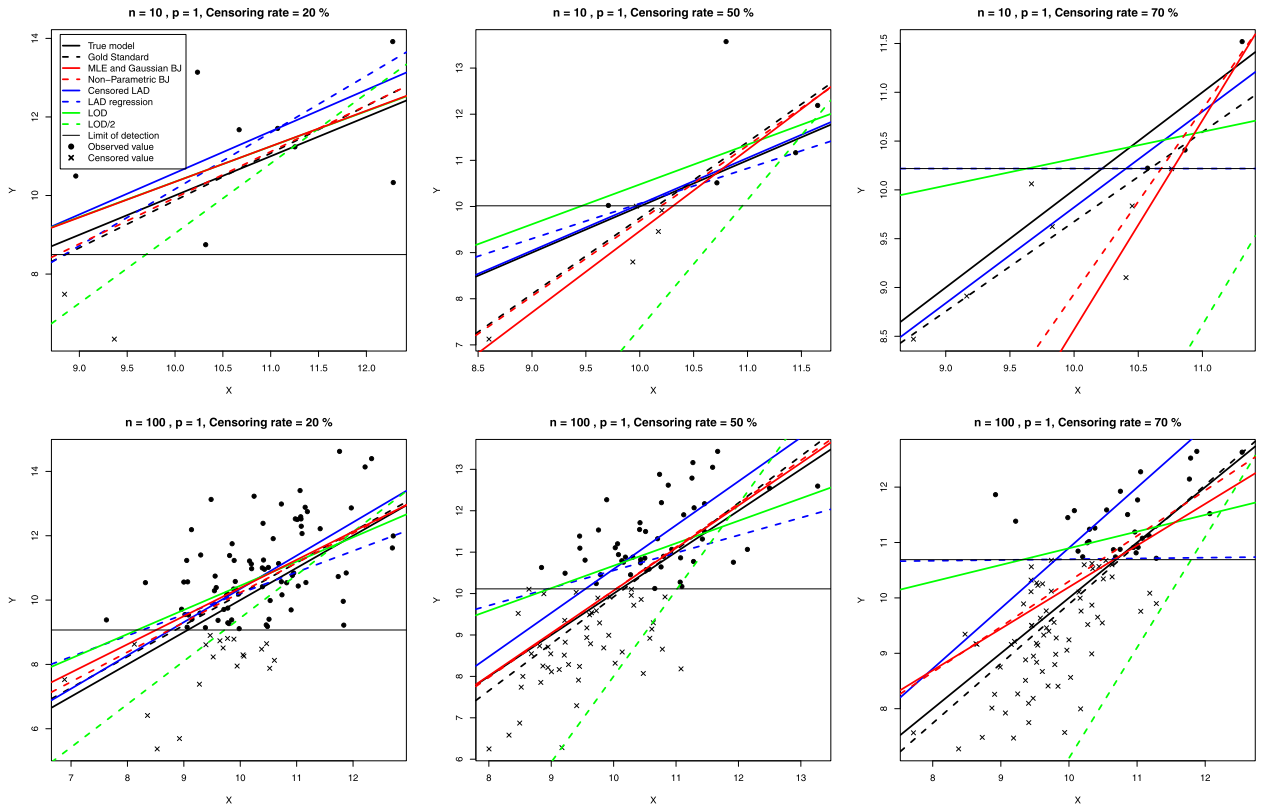

Fig. 1 Observed and predicted simulated censored data. Regression lines are obtained using: the true model that generated the data, the gold standard, maximum likelihood estimation (MLE), non-parametric Buckley-James (BJ), least absolute deviations (LAD) and censored LAD regressions, simple imputation by the limit of detection (LOD) and by LOD/2

CV is evaluated on a grid of λ–values. The highest value, $\lambda_{\max}$, corresponds to the smallest value of λ for which all coefficients are zero. The lowest value, $\lambda_{\min}$, corresponds to the unpenalized solution (when feasible). We choose the λ value that minimizes the CV function.

Squared error loss is one of the most widely used loss functions:

$$L(\hat{\boldsymbol{\beta}}(\lambda)_{\mathbf{D}_{\setminus k}}, \mathbf{D}_k) = \frac{1}{n_k} \sum_{i \in \mathbf{D}_k} \left(Y_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}(\lambda)_{\mathbf{D}_{\setminus k}} \right)^2, \quad (14)$$

where n_k is the sample size of $\mathbf{D}_k$. However, $\mathbf{Y}$ is a latent variable not fully observed due to the detection limit. This loss function could be used only for the gold standard in (2), with simulated data. Again, the simplest imputation strategy consists in replacing $\mathbf{Y}$ with $\mathbf{Z}$, in $\mathbf{D}_k$. Alternatively, Buckley-James strategies could replace censored Y_i values in the test data $\mathbf{D}_k$ by their conditional expectation estimated using the learning data [48].

On the other hand, a loss function differentiating the contribution of uncensored and censored data would be useful. Assuming the Gaussian distribution of the HIV viral load (7), the following loss function could be derived:

$$L_G\left(\hat{\boldsymbol{\beta}}(\lambda)_{\mathbf{D}_{\setminus k}}, \mathbf{D}_k\right) = \frac{1}{n_k^{\text{unc}}} \sum_{\substack{i \in \mathbf{D}_k \\ i \text{ uncensored}}} \left(Y_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}(\lambda)_{\mathbf{D}_{\setminus k}} \right)^2 + \frac{2\hat{\sigma}^2_{\mathbf{D}_{\setminus k}}}{n_k^{\text{unc}}} \sum_{\substack{i \in \mathbf{D}_k \\ i \text{ censored}}} -\ln F_G\left(\text{LOD}, \mathbf{X}_i, \hat{\boldsymbol{\beta}}(\lambda)_{\mathbf{D}_{\setminus k}}, \hat{\sigma}^2_{\mathbf{D}_{\setminus k}}\right) \quad (15)$$

where n_k^{unc} is the number of uncensored observations in $\mathbf{D}_k$. The loss function L_G in (15) is proportional and equivalent to the negative Gaussian log-likelihood loss function, but allows for comparison with the squared loss in (14).

Implementation issues

All statistical analyses, comparisons and implementations were performed using the computing environment R (R Development Core Team, 2017) [69]. We used the function `cv.glmnet` from package `glmnet` [70] to choose the optimal λ value of Lasso linear regression on complete data (*GoldS*) and Lasso linear regression with simple substitution of left-censored values by the detection limit (*LOD*). We implemented the Lasso non-parametric Buckley-James (*NonParBJ*) using the `bujar`

package [71]. We modified the function to support stratified K-fold cross-validation and conserve the same proportion of censoring in all the folds. Lasso Gaussian Buckley-James (*Gaussian BJ*) was implemented in a new function `cvGaussBJ`. Algorithm 1 specifies how to solve the problem. The stopping criterion is based on the difference between current and previous regression coefficient estimates, variance estimates, and imputed data. Because of the tendency to oscillate between different parameter values of the iterative procedure, the algorithm is also stopped if the number of oscillations is high [40]. Alternatively, we also considered the one-step algorithm, which stops at the first iteration [36, 64]. Cross-validation based on both imputation and loss function accounting for censored and uncensored contributions is considered. The Lasso estimation step depends on package `glmnet`.

All these implementations and an artificial example are available at: https://github.com/psBiostat/left-censored-Lasso.

Prediction of HIV viral load from HIV genotypic mutations: real and simulated data

HIV is highly replicative and thus presents high mutation and recombination rates which could lead to the development of HIV drug resistance and consequently reduce the efficacy of antiretroviral treatment. To optimize the control of the evolution of HIV drug resistance, HIV viral load is routinely monitored to identify treatment failure, and HIV genotypic tests are commonly performed before a switch to a new treatment regimen in patients already treated or at the initiation of the first treatment in naive HIV-infected patients [72].

Algorithm 1 Lasso-regularized Gaussian Buckley-James

Initialization:

$$\hat{\boldsymbol{\beta}}^{(0)} \leftarrow \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} (Z_i - \mathbf{X}_i\boldsymbol{\beta})^2 + \lambda\|\boldsymbol{\beta}\|_1$$

$$\hat{\sigma}^{2(0)} \leftarrow \frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}^{(0)}\right)^2$$

while the stopping criterion is not satisfied **do**

Imputation step

$$Z_i^{*(k)} \leftarrow \delta_i Y_i + (1-\delta_i)\int_{-\infty}^{\mathrm{LOD}} \frac{u f_G(u, \mathbf{X}_i, \hat{\boldsymbol{\beta}}^{(k-1)}, \hat{\sigma}^{2(k-1)})\,\mathrm{d}u}{F_G(\mathrm{LOD}, \mathbf{X}_i, \hat{\boldsymbol{\beta}}^{(k-1)}, \hat{\sigma}^{2(k-1)})}$$

Lasso estimation step

$$\hat{\boldsymbol{\beta}}^{(k)} \leftarrow \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} (Z_i^{*(k)} - \mathbf{X}_i\boldsymbol{\beta}^{(k)})^2 + \lambda\|\boldsymbol{\beta}^{(k)}\|_1$$

$$\hat{\sigma}^{2(k)} \leftarrow \frac{1}{n}\sum_{i=1}^{n}\left(Z_i^{*(k)} - \mathbf{X}_i\hat{\boldsymbol{\beta}}^{(k)}\right)^2$$

end while

Our objective is to compare methods that handle left-censoring by conditional imputation with methods that handle left-censoring by imputing a single constant value (that is, the Lasso-regularized linear regression with simple imputation by LOD and LOD/2) to predict of HIV viral load by HIV genotypic mutations. The methods accounting for left-censoring by imputing the estimated conditional mean given censoring and predictor values are the Lasso-regularized Buckley-James least square algorithms (with/without Gaussian assumption, with complete convergence/1-step, using cross-validation based on imputation/loss function accounting for censored and uncensored contributions).

Real data

The database used in this study was provided by the Standardization and Clinical Relevance of HIV Drug Resistance Testing Project for the Forum for Collaborative HIV Research [49]. Patients included in this study were all treatment-experienced and switched to an abacavir-containing regimen. The investigated drug, abacavir, is a nucleoside reverse transcriptase inhibitor (NRTI) that blocks HIV reverse transcriptase.

The sample size $n = 99$ was slightly smaller than the number of predictors $p = 121$. 54 of the 121 predictors correspond to the presence or absence of specific mutations in the reverse transcriptase gene (RTG), which were reported to be probably associated with resistance to abacavir, multi-NRTI, NRTI (other than abacavir) or non-nucleoside reverse transcriptase inhibitors (NNRTI) at the time of the study [73, 74]. The number of mutations reported to be probably associated with resistance to abacavir or multi-NRTI is low (14%). The other 67 predictors correspond to the presence or absence of specific mutations in the protease gene (PG) reported to be probably associated with resistance to one or several protease inhibitors (PI) at the time of the study [73, 74]. The number of molecules, including abacavir, ranged from 1 to 6 (with the median number of molecules being 3 and interquartile range 2). In particular, a PI was prescribed in 59% of the patients and 43% received an NNRTI. The response variable is the log-HIV viral load measured at t_8 (8 weeks after treatment initiation at t_0). LOD was fixed at 100 copies/mL and the censoring rate was moderate (26%).

Generation of simulated data

HIV viral load appears to have an underlying Gaussian distribution when log-transformed. Therefore, our outcome, $Y_i^{(8)}$ is generated from a Gaussian distribution. We simulated 200 data sets of size $n = 100$ and $p = 100$ predictors from the model:

$$Y_i^{(8)} = \beta_0 + \beta_1^{(0)} Y_i^{(0)} + \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i \qquad i = 1, \cdots, n$$

where

- $Y_i^{(0)}$, the HIV viral load at t_0 is generated by a normal distribution with mean 12 ($\log_{10}$ copies/mL) and variance 1.
- $\beta_1^{(0)}$ represents the change of the slope between the HIV viral load on the day of treatment, t_0, and 8 weeks later, at t_8, when no mutations are present and for 1 log_{10}/mL higher concentration of viral load at t_0. We fix β_0 and $\beta_1^{(0)}$ to obtain the desired censoring rates: 20% (moderate), 50% (high), and 70% (severe).
- $\mathbf{X}_{(n\times p)}$, representing the presence or absence of HIV mutations, is generated by a multinomial distribution with mean 0.15 (the fixed prevalence for all the 100 mutations) and covariance matrix $\mathbf{\Sigma}$ where $\Sigma_{ij} = 0.4^{|i-j|}$ (the closer the mutations, the more positively they are correlated).
- $\boldsymbol{\beta} = \left(\beta_1, \cdots, \beta_p\right)^\top$. Among $p = 100$ candidate mutations only 10% are relevant with effects $\beta_j = 1$, if $j = 1, \ldots, 10$ and 0 if $j > 10$. A 1-unit increase in HIV viral load is expected per occurrence of these relevant mutations for a given baseline HIV viral load.
- $\boldsymbol{\varepsilon}$ is generated from a normal distribution with mean 0 and variance σ^2 chosen such that the signal-to-noise ratio is fixed at $3:1$.

Our primary goal was to compare competing methods in terms of prediction accuracy. Consequently, we simulated training and test datasets. The former were used to estimate, the latter were used to evaluate the prediction performance. We ensured that training and test datasets contained roughly the same proportions of censoring. We computed the mean squared error on test data as:

$$\text{MSE} = \frac{1}{n^{\text{test}}} \sum_{i=1}^{n^{\text{test}}} \left(Y_i^{\text{test}} - \mathbf{X}_i^{\text{test}} \hat{\boldsymbol{\beta}}(\hat{\lambda}) \right)^2$$

with $\hat{\boldsymbol{\beta}}(\hat{\lambda})$ estimated on training data by stratified 5-fold cross-validation using a given regression method (Gold standard, Lasso-regularized Non-parametric and Gaussian Buckley-James -with complete convergence and 1-step- and Lasso-regularized linear regression with simple imputation -by LOD and LOD/2-) and the corresponding loss function in the cross-validation criterion. The gold standard uses (14), the others replace the censoring values according to their imputation strategy, and the Gaussian Buckley-James also uses the loss function in (15).

Results

Simulation results

Figure 2 shows the mean prediction error results. "OLS on True Model" corresponds to ordinary least squares for the linear model with uncensored data and only relevant predictors (the so-called oracle estimator). Gold standard (*GoldS*) corresponds to the Lasso estimation for the linear model with uncensored data. These results allow for reference prediction errors when the true model is known (the first one) or due to censoring data (both).

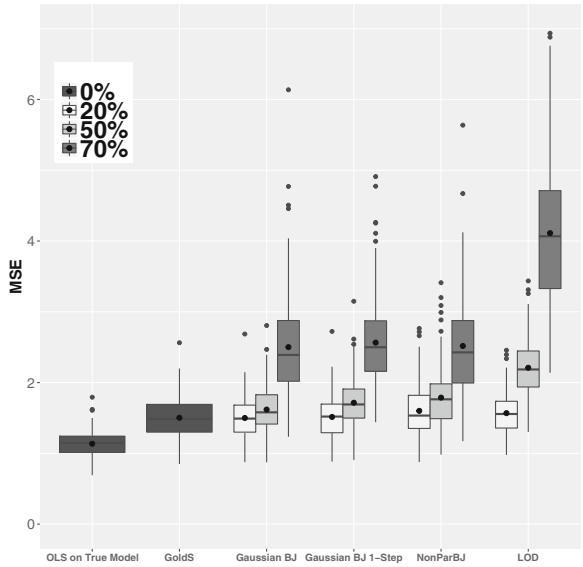

Fig. 2 Mean Square Error (MSE) calculated on test data of size $n_{test} = 100$ and 200 replications using gold standard method (*GoldS*), Gaussian Buckley-James (*Gaussian BJ*) and 1-Step version (*Gaussian BJ 1-Step*), non-Parametric Buckley-James (*NonParBJ*) and simple imputation by LOD (*LOD*). Data were generated to present different censoring rates: 0% (uncensored), 20% (moderate), 50% (high), and 70% (severe), which are represented by different gray degrees

The imputation by LOD/2 led to poorer results than the imputation by LOD. Thus, for the simple imputation, only LOD imputation (*LOD*) results are shown. For the Gaussian Buckley-James algorithms (*Gaussian BJ*), the error is calculated by using both cross-validation with imputation and cross-validation with the loss function indicated in (15), but only the best results are shown.

The Gaussian Buckley-James method presented an oscillating behavior in 9.5% of the generated samples when the convergence rate was 20%. This percentage rose to 82.5% and 95.0% when the convergence rates were 50% and 70%, respectively. For the Gaussian Buckley-James using the 1-step algorithm (*Gaussian BJ 1-Step*), results using the two cross-validation approaches were almost identical. Nevertheless, for the Gaussian Buckley-James with complete convergence, a notable improvement was obtained when applying (15).

The higher the rate of censoring, the less information is available to train the models and, unsurprisingly, the higher is the prediction error. For a moderate rate of censoring (20%), all methods show a good performance close to that of the gold standard *GoldS*. When the rate of censoring is 50%, *Gaussian BJ* shows the lowest prediction error, followed by *Gaussian BJ 1-step*, *NonParBJ* and finally simple imputation, which shows more errors.

The same patterns but more pronounced were observed with a severe rate of censoring (70%). Taking the knowledge about the distribution into account appears to have only a slight impact. Simple imputation yields the poorest results. In addition, it showed high variability with some extreme errors.

Application to real data

The Lasso-regularized Buckley-James least square algorithm that showed the best behavior in the simulation study (with complete convergence and cross-validation based on the loss function L_G in (15)) was applied to real data, as well as the Lasso-regularized non-parametric Buckley-James method and simple imputation (by LOD and LOD/2).

Regularization parameters were estimated by stratified cross-validation in order to ensure that each fold had the same proportion of censoring as in the corresponding data set (26%). In addition, because some mutations were relatively infrequent, we used 20-fold cross-validation. Indeed, the higher the number of folds, the lower the probability of randomly obtaining test sets with no subject exposed to infrequent HIV drug resistance mutations.

Figure 3 shows two examples of the observed HIV viral load at t_0 and the observed and estimated HIV viral load at t_8. As in the low-dimensional case shown in Fig. 1, simple imputation by LOD predicted the highest values of HIV viral load. Inversely, simple imputation by LOD/2 estimates the lowest values of HIV viral load. The difference between the two estimates at 8 weeks is $>0.5\ log_{10}$ copies/mL, which is clinically relevant. Lasso-regularized Buckley-James least square algorithms (with/without Gaussian assumption), often gave a prediction in between. This tendency was increased when the censoring rate was high or when the sensitivity of the assay was low.

When applying the Gaussian Buckley-James method to real data, no oscillating behavior was observed.

Table 1 indicates the number of HIV genotypic mutations selected from the list of mutations that may contribute to a reduced virologic response known at the time of the study [73, 74], according to each method applied. The Lasso-regularized Gaussian Buckley-James selected several HIV genotypic mutations suspected of being associated with abacavir or multi-NRTI resistance. Furthermore, it selected a high number of HIV genotypic mutations probably associated with PI resistance and a few probably associated with NNRTI resistance. This selection of a large number of candidate predictors seems to be relevant because all patients received an abacavir-containing regimen and a high percentage of patients received regimens including a PI- and/or NNRTI. The Lasso-regularized non-parametric Buckley-James selected fewer mutations, and especially fewer mutations in PG, probably due to PI resistance. Simple imputation of the LOD or LOD/2 selected few mutations. In particular, only 1 of the 5 mutations in RTG probably associated with abacavir resistance was retained.

Discussion

Simple imputation of the detection limit or of half of this limit is an ad hoc approach to address left-censored outcome data. However, in standard (low-dimensional) settings, it leads to biased estimates of parameters and standard errors. In our high-dimensional simulation study, simple imputation using Lasso-regularized least-squares showed poor performance. As in low-dimensional

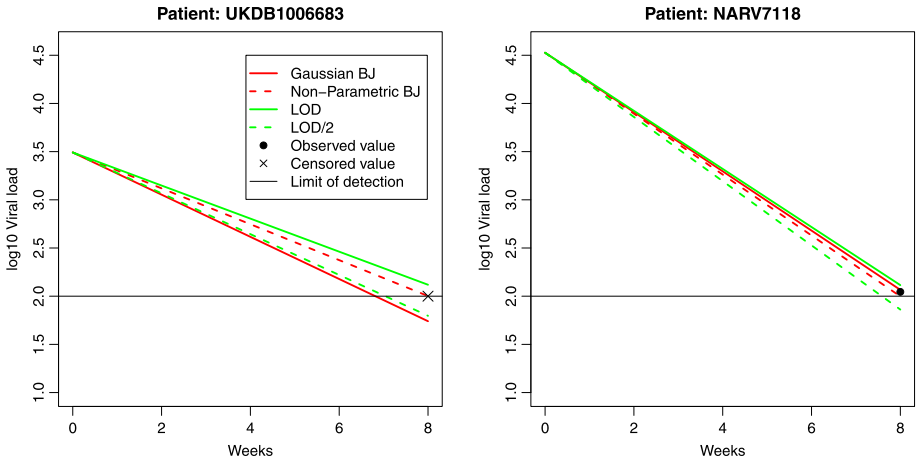

Fig. 3 Two examples of individual predicted values using Lasso regularization techniques (Gaussian Buckley-James, non-parametric Buckley-James, simple imputation by LOD and LOD/2) and observed values. One example is censored (at left) and the other is uncensored (at right)

Table 1 Distribution of 121 HIV genotypic mutations included in real data study according to knowledge at study time and reported in [73, 74] and number of HIV genotypic mutations selected by Lasso regularized methods

Number of HIV genotypic mutations present in real data study	Gaussian BJ	NonPar BJ	LOD	LOD/2
5 in RTG probably associated with abacavir resistance	3 (60%)	3 (60%)	1 (20%)	1 (20%)
13 in RTG probably associated with multi-NRTI resistance	7 (54%)	4 (31%)	4 (31%)	4 (31%)
6 in RTG probably associated with NRTI resistance (other than abacavir)	3 (50%)	2 (33%)	1 (17%)	1 (17%)
30 in RTG probably associated with NNRTI resistance	12 (40%)	11 (37%)	8 (27%)	7 (23%)
67 in PG probably associated with PI resistance	40 (60%)	22 (33%)	15 (22%)	14 (21%)
121 Total	65 (54%)	42 (35%)	29 (24%)	27 (22%)

settings, approaches accounting for left-censoring outperformed simple imputation.

In this work, we propose a Lasso-regularized Gaussian Buckley-James algorithm, according to the usual Gaussian assumption of log-transformed HIV viral load. Because of the well-known convergence problems of the iterative Buckley-James procedure, we implemented two algorithms, the first algorithm running until convergence and the second one being stopped after one step [36, 64]. This one-step algorithm showed similar results in the simulation study. Other solutions have been proposed to deal with convergence problems in low-dimensional settings [39, 64] and could be investigated in future research.

As in other works [48], we implemented a cross-validation criterion for the tuning parameter based on imputing values to Y_i in the test set from conditional expectations estimated using the learning set. We also proposed a cross-validation criterion based on a loss function that accounts for the different contribution of censored and uncensored values. Almost identical results were obtained when applying the two cross-validation criteria to the one-step algorithm. However, when running the algorithm until convergence, better results were obtained with the cross-validation criterion based on a loss function that accounts for censored and uncensored contributions.

On the other hand, we reversed the Lasso-regularized non-parametric Buckley-James method previously applied to right-censored survival data [36, 38, 40, 48] in order to apply to left-censoring due to detection limits. Foreseeably, in our homoscedastic Gaussian outcome data scenario, the Gaussian Buckley-James showed better behavior than the non-parametric algorithm. However, accounting for the knowledge about the distribution seems to have had a slight influence. When the Gaussian assumption is violated, non-parametric imputation using Kaplan Meier is perhaps the best option.

We provide a publicly available R code to compute the methods introduced in this work (https://github.com/psBiostat/left-censored-Lasso). It would be interesting to compare the Lasso-regularized Buckley-James least squares method to Lasso-regularized censored LAD method. The Lasso extension of censored LAD has been proposed in different works [41, 42, 56–62]. However, to our knowledge, there is no publicly available implementation, and no simple implementation relying on existing packages seem straightforward. Moreover, several works have shown the superiority of maximum likelihood estimation in low-dimensional settings when the Gaussian assumption is valid [18, 20–22]. Nevertheless, optimization strategies for complex likelihood functions (such as that in Eq. 7) including penalties that are not smooth are not obvious.

To illustrate the application of the methods on real data, we consider a data set from the Standardization and Clinical Relevance of HIV Drug Resistance Testing Project for the Forum for Collaborative HIV Research. The data set used to illustrate the initial data set is characterized by a sample size-to-predictors ratio of around 1. There is no gold standard to measure and compare predictive performance of the different methods when using censored outcome data. All patients were being treated with abacavir, an NRTI, so we expected our methods to select a high number of HIV genotypic mutations known to contribute to abacavir and NRTI resistance. Furthermore, a high number of patients were on PI- and/or NNRTI-containing regimens, and a selection of several HIV genotypic mutations reported to be probably associated with resistance to any of these molecules was also expected [73, 74]. In that sense, the Gaussian and non-parametric Buckley-James methods showed more coherent results with the literature compared to simple imputation.

Otherwise, the data presented a moderate censoring rate of 26%, which is a realistic magnitude [18, 50] in

studies measuring HIV viral load. However, high or even severe censoring rates were found in older studies with a high limit of detection (LOD) or particular populations with a low treatment failure rate [18, 51, 52]. Furthermore, left-censoring due to the lower detection limit of an assay is a problem in many fields such as biology, immunology, chemistry, and the environmental sciences in which high censoring rates may be frequent. Our simulation study shows that the difference in performance between Lasso-regularized Buckley-James methods and Lasso-regularized simple imputation methods increased with the censoring rate.

In our simulations and real application, the detection threshold was the same for all subjects. The detection threshold may vary among subjects, for example, in multicentric studies. Our R code also supports multiple lower limits of quantification. However, the findings should be interpreted with caution: differences in technological equipment could be a confounding factor that might help explain the differences in patient response to HIV treatment (in addition to HIV mutations). Adjusting or stratifying for the hospital would then be necessary.

In this study we focused on the prediction performance of Lasso-regularized methods. In clinical applications, even when prediction accuracy is the main objective, researchers aim to identify which predictors are more strongly associated with outcome. Our proposal could be easily extended or adapted to support other Lasso-type penalties. When the primary goal is to infer the set of truly relevant variables, the adaptive Lasso and the bootstrap-enhanced Lasso could thus be considered.

Acknowledgements
We acknowledge members of the Standardization and Clinical Relevance of HIV Drug Resistance Testing Project for the Forum for Collaborative HIV Research. We thank "Sidaction, Ensemble contre le Sida", France, for their continuous support. We would like to thank Binbin Xu, postdoctoral researcher at SISTM research team from Inserm BPH U1219 & Inria BSO, for his help on testing the R code.

Funding
This work was partially supported by the "Investissements d'Avenir" program managed by the ANR under reference ANR-10-LABX-77

Availability of data and materials
R code and artificial example are available at https://github.com/psBiostat/left-censored-Lasso.

Authors' contributions
PS developed the algorithms and corresponding R code, carried out the statistical analysis and helped to draft the manuscript. MA developed the algorithms, revised the R code and drafted the manuscript. RT designed and supervised the applied research. RT and LW interpreted the results of the analysis. DC revised the methodology. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, UMR 1219, F-33000 Bordeaux, France. [2]Inria SISTM Team, F-33405 Talence, France. [3]Vaccine Research Institute (VRI), F-94000 Créteil, France. [4]CHU Bordeaux, Department of Public Health, F-33000 Bordeaux, France.

Received: 4 October 2017 Accepted: 2 November 2018
Published online: 04 December 2018

References

1. Paxton W, Coombs R, McElrath M, Keefer M, Hughes J, Sinangil F, Chernoff D, Demeter L, B BW, Corey L. Longitudinal analysis of quantitative virologic measures in human immunodeficiency virus-infected subjects with > or = 400 CD4 lymphocytes: implications for applying measurements to individual patients. National Institute of Allergy and Infectious Diseases AIDS Vaccine Evaluation Group. J Infect Dis. 1997;175(2):247–54.
2. Helsel DR. More than obvious: Better methods for interpreting nondetect data. Environ Sci Technol. 2005;39(20):419–23.
3. Lee M, Kong L, Weissfeld L. Multiple imputation for left-censored biomarker data based on Gibbs sampling method. Stat Med. 2012;31:1838–48.
4. Del Greco M F, Pattaro C, Minelli C, Thompson JR. Bayesian analysis of censored response data in family-based genetic association studies. Biom J. 2016;58(5):1039–53.
5. Marschner I, Betensky R, DeGruttola V, Hammer S, Kuritzkes D. Clinical trials using HIV-1 RNA-based primary endpoints: Statistical analysis and potential biase. J Acquir Immune Defic Syndr Hum Retrovirol. 1999;20(3):220–7.
6. Gillespie BW, Chen Q, Reichert H, Franzblau A, Hedgeman E, Lepkowski J, Adriaens P, Demond A, Luksemburg W, Garabrant DH. Estimating population distributions when some data are below a limit of detection by using a reverse Kaplan-Meier estimator. Epidemiology. 2010;21:64–70.
7. Dinse G, Jusko A, Ho L, Annam K, Graubard B, Hertz-Picciotto I, Miller F, Gillespie B, Weinberg C. Accomodating measurements below a limit of detection: A novel application of Cox regression. Am J Epidemiol. 2014;179(8):1018–24.
8. Wang HJ, Zhu Z, Zhou J. Quantile regression in partially linear varying coefficient models. Ann Stat. 2009;37(6B):3841–66.
9. Eilers PH, Röder E, Savelkoul HF, van Wijk RG. Quantile regression for the statistical analysis of immunological data with many non-detects. BMC Immunol. 2012;13:13–37.
10. Powell JL. Least absolute deviations estimation for the censored regression model. J Econ. 1984;25:303–25.
11. Powell JL. Censored regression quantiles. J Econom. 1986;32:143–55.
12. Tobin J. Estimation of relationships for limited dependent variables. Econometrica. 1958;26:24–36.
13. Hughes JP. Mixed effects models with censored data with application to HIV RNA levels. Biometrics. 1999;55:625–9.
14. Jacqmin-Gadda H, Thiébaut R, Chêne G, Commenges D. Analysis of left-censored longitudinal data with application to viral load in HIV infection. Biostatistics. 2000;1(4):355–68.
15. Lynn HS. Maximum likelihood inference for left-censored HIV RNA data. Stat Med. 2001;20:33–45.
16. Nie L, Chu H, Liu C, Cole SR, Vexler A, Schisterman EF. Linear regression with an independent variable subject to a detection limit. Epidemiology. 2010;21:17–24.
17. Fu P, Hughes J, Zeng G, Hanook S, Orem J, Mwanda O, Remick S. A comparative investigation of methods for longitudinal data with limits of detection through a case study. Stat Methods Med Res. 2016;25(1):153–66.
18. Wiegand RE, Rose CE, Karon JM. Comparison of models for analyzing two-group, cross-sectional data with a gaussian outcome subject to a detection limit. Stat Methods Med Res. 2016;25(6):2733–49.

19. Buckley J, James I. Linear regression with censored data. Biometrika. 1979;66:429–36.
20. Hewett P, Ganser GH. A comparison of several methods for analyzing censored data. Ann Occup Hyg. 2007;51:611–32.
21. Uh H-W, Hartgers FC, Yazdanbakhsh M, Houwing-Duistermaat JJ. Evaluation of regression methods when immunological measurements are constrained by detection limits. BMC Immunol. 2008;9(1):59.
22. Kafatos G, Andrews N, McConway KJ, Farrington P. Regression models for censored serological data. J Med Microbiol. 2013;62(Pt 1):93–100.
23. Hirsch MS, Günthard HF, Schapiro JM, Vézinet FB, Clotet B, Hammer SM, Johnson VA, Kuritzkes DR, Mellors JW, Pillay D, et al. Antiretroviral drug resistance testing in adult HIV-1 infection: 2008 recommendations of an International AIDS Society-USA panel. Clin Infect Dis. 2008;47(2):266–85.
24. Wittkop L, Günthard H, de Wolf F, Dunn D, Cozzi-Lepri A, de Luca A, Kücherer C, Obel N, von Wyl V, Masquelier B, Stephan C, Torti C, Antinori A, Garcia F, Judd A, Porter K, Thiébaut R, Castro H, van Sighem A, Colin C, Kjaer J, Lundgren J, Paredes R, Pozniak A, Clotet B, philipps A, Pillay D, Chêne G, study group E-C. Effects of transmitted drug resistance on virological and immunological response to initial combination antiretroviral therapy for HIV (euro-coord-chain joint project): a european multicohort study. Lancet Infect Dis. 2011;11(5):363–71.
25. Hofstra LM, Sauvageot N, Albert J, Alexiev I, Garcia F, Struck D, Van de Vijver DA, Åsjö B, Beshkov D, Coughlan S, et al. Transmission of HIV drug resistance and the predicted effect on current first-line regimens in europe. Clin Infect Dis. 2016;62(5):655–63.
26. Wensing AM, Calvez V, Günthard HF, Johnson VA, Paredes R, Pillay D, Shafer RW, Richman DD. 2017 update of the drug resistance mutations in HIV-1. Top Antivir Med. 2017;24(4):132.
27. Rabinowitz M, Myers L, Banjevic M, Chan A, Sweetkind-Singer J, Haberer J, McCann K, Wolkowicz R. Accurate prediction of HIV-1 drug response from the reverse transcriptase and protease amino acid sequences using sparse models created by convex optimization. Bioinformatics. 2006;22(5):541–9.
28. Beerenwinkel N, Montazeri H, Schuhmacher H, Knupfer P, von Wyl V, Furrer H, Battegay M, Hirschel B, Cavassini M, Vernazza P, Bernasconi E, Yerly S, Böni J, Klimkait T, Cellerai C, Günthard HF, Study TSHC. The individualized genetic barrier predicts treatment response in a large cohort of HIV-1 infected patients. PLoS Comput Biol. 2013;9(8):1–11.
29. Cozzi-Lepri A, Prosperi MCF, Kjær J, Dunn D, Paredes R, Sabin CA, Lundgren JD, Phillips AN, Pillay D, for the EuroSIDA, the United Kingdom CHIC/United Kingdom HDRD Studies. Can linear regression modeling help clinicians in the interpretation of genotypic resistance data? an application to derive a lopinavir-score. PLoS ONE. 2011;6(11):1–9.
30. Wittkop L, Commenges D, Pellegrin I, Breilh D, Neau D, Lacoste D, Pellegrin J-L, Chêne G, Dabis F, Thiébaut R. Alternative methods to analyse the impact of HIV mutations on virological response to antiviral therapy. BMC Med Res Methodol. 2008;8(1):68.
31. Assoumou L, Houssaïna A, Corstagliola D, Flandre P, Standardization and clinical relevance of HIV drug resistance testing project from the forum for collaborative HIV research. Relative contributions of baseline patient characteristics and the choice of statistical methods to the variability of genotypic resistance scores: the example of didanosine. J Antimicrop Chemother. 2010;65(4):752–60.
32. Rhee S, Taylor J, Wadhera G, Ben-Hur A, Brutlag D, Shafer R. Genotypic predictos of human immunodeficiency cirus type 1 drug resistance. Proc Natl Acad Sci USA. 2006;103(46):17355–60.
33. Tibshirani R. The Lasso method for variable selection in the Cox model. Stat Med. 1997;16:385–95.
34. Huang J, Ma S, Xie H. Regularized estimation in the accelerated failure time model with high-dimensional covariates. Biometrics. 2006;62: 813–20.
35. Datta S, Le-Rademacher J, Datta S. Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and Lasso. Biometrics. 2007;63:259–71.
36. Johnson BA. Variable selection in semiparametric linear regression with censored data. J R Stat Soc Ser B Stat Methodol. 2008;70:351–70.
37. Wang S, Nan B, Zhu J, Beer DG. Doubly penalized Buckley-James method for survival data with high-dimensional covariates. Biometrics. 2008;64(1):132–40.
38. Cai T, Huang J, Tian L. Regularized estimation for the accelerated failure time model. Biometrics. 2009;65:394–404.
39. Ueki M. A note on automatic variable selection using smooth-threshold estimating equations. Biometrika. 2009;96(4):1005–11.
40. Wang Z, Wang CY. Buckley-james boosting for survival analysis with high-dimensional biomarker data. Stat Appl Genet Mol Biol. 2010;9(1):24.
41. Shows JH, Lu W, Zhang HH. Sparse estimation and inference for censored median regression. J Stat Plan Infer. 2010;140:1903–17.
42. Wang HJ, Zhou J, Li Y. Variable selection for censored quantile regression. Stat Sin. 2013;23(1):145–67.
43. Chung M, Long Q, Johnson BA. A tutorial on rank-based coefficient estimation for censored data in small-and large-scale problems. Stat Comput. 2013;23(5):601–14.
44. Huang X, Pan W, Park S, Han X, Miller LW, Hall J. Modeling the relationship between LVAD support time and gene expression changes in the human heart by penalized partial least squares. Bioinformatics. 2004;20(6):888–94.
45. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat. 2008;2:841–60.
46. Wang Y, Chen T, Zeng D. Support vector hazards machine: A counting process framework for learning risk scores for censored outcomes. J Mach Learn Res. 2016;17(167):1–37.
47. Van der Burgh HK, Schmidt R, Westeneng H-J, de Reus MA, van den Berg LH, van den Heuvel MP. Deep learning predictions of survival based on MRI in amyotrophic lateral sclerosis. NeuroImage Clin. 2017;13:361–9.
48. Johnson BA. On lasso for censored data. Electron J Stat. 2009;3: 485–506.
49. Cozzi-Lepri A. Initiatives for developing and comparing genotype interpretation systems: external validation of existing rule-based interpretation systems for abacavir against virological response. HIV Med. 2008;9(1):27–40.
50. Marks G, Gardner LI, Craw J, Giordano TP, Mugavero MJ, Keruly JC, Wilson TE, Metsch LR, Drainoni M-L, Malitz F. The spectrum of engagement in HIV care: do more than 19% of HIV-infected persons in the US have undetectable viral load?. Clin Infect Dis. 2011;53(11):1168–9.
51. Dao CN, Patel P, Overton ET, Rhame F, Pals SL, Johnson C, Bush T, Brooks JT, Study to Understand the Natural History of HIV and AIDS in the Era of Effective Therapy (SUN) Investigators. Low vitamin D among HIV-infected adults: prevalence of and risk factors for low vitamin D levels in a cohort of HIV-infected adults and comparison to prevalence among adults in the US general population. Clin Infect Dis. 2011;52(3):396–405.
52. Leon A, Perez I, Ruiz-Mateos E, Benito JM, Leal M, Lopez-Galindez C, Rallon N, Alcami J, Lopez-Aldeguer J, Viciana P, et al. Rate and predictors of progression in elite and viremic HIV-1 controllers. AIDS. 2016;30(8): 1209–20.
53. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B Methodol. 1996;58(1):267–88.
54. Sigrist F, Stahel WA. Using the censored gamma distribution for modeling fractional response variables with an application to loss given default. ASTIN Bull J Int Actuar Assoc. 2011;41(02):673–710.
55. Belloni A, Chernozhukov V. L1-penalized quantile regression in high-dimensional sparse models. Ann Stat. 2011;39(1):82–130.
56. Xue X, Xie X, Strickler HD. A censored quantile regression approach for the analysis of time to event data. Stat Methods Med Res. 2018;27(3): 955–65.
57. Zhanfeng W, Yaohua W, Lincheng Z. A lasso-type approach to variable selection and estimation for censored regression model. Chin J Appl Probab Stat. 2010;26(1):66–80.
58. Yue YR, Hong HG. Bayesian tobit quantile regression model for medical expenditure panel survey data. Stat Model. 2012;12(4):323–46.
59. Liu X, Wang Z, Wu Y. Group variable selection and estimation in the tobit censored response model. Comput Stat Data Anal. 2013;60:80–9.
60. Zhou X, Liu G. LAD-lasso variable selection for doubly censored median regression models. Commun Stat Theory Methods. 2013;45(12):3658–67.
61. Alhamzawi R. Bayesian elastic net tobit quantile regression. Commun Stat Simul Comput. 2016;45(7):2409–27.
62. Müller P, van de Geer S. Censored linear model in high dimensions. TEST. 2015;25(1):75–92.
63. Peter Wu C-S, Zubovic Y. A large-scale monte carlo study of the Buckley-James estimator with censored data. J Stat Comput Simul. 1995;51(2-4):97–119.
64. Wang Y-G, Zhao Y, Fu L. The Buckley–James estimator and induced smoothing. Aust N Z J Stat. 2016;58(2):211–25.

65. Gleit A. Estimation for small normal data sets with detection limits. Environ Sci Technol. 1985;19(12):1201–6.
66. Johnson BA, Long Q, Chung M. On path restoration for censored outcomes. Biometrics. 2011;67:1379–88.
67. Zhao SD, Lee D, Li Y. The Dantzig selector for censored linear regression models. Stat Sin. 2014;24(1):251–68.
68. DiRienzo AG. Parsimonious covariate selection with censored outcomes. Biometrics. 2016;72:452–62.
69. R Core Team. R: A language and environment for statistical computing. Vienna: R foundation for statistical computing; 2017. ISBN 3-900051-07-0, http://www.R-project.org.
70. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22.
71. Wang Z, Wang MZ, Suggests T. bujar: Buckley-James regression for survival data with high-dimensional covariates. 2015. R package version 0.2-1. https://CRAN.R-project.org/package=bujar.
72. Iyidogan P, Anderson KS. Current perspectives on HIV-1 antiretroviral drug resistance. Viruses. 2014;6(10):4095–139.
73. Shafer RW, Schapiro JM. HIV-1 drug resistance mutations: an updated framework for the second decade of HAART. AIDS Rev. 2008;10(2):67.
74. Johnson VA, Brun-Vézinet F, Clotet B, Gunthard H, Kuritzkes DR, Pillay D, Schapiro JM, Richman DD. Update of the drug resistance mutations in HIV-1: December 2009. Top HIV Med. 2009;17(5):138–45.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

3.3 Conclusion

La présence d'une réponse censurée est un problème récurrent dans les études biologiques. Jusqu'à présent celle-ci était imputée par une valeur fixe et traitée par un modèle de régression classique ou dichotomisée et traitée dans un modèle de régression logistique. Dans ce chapitre, nous avons montré l'intérêt de tenir compte de l'incertitude autour cette réponse censurée pour des données de grande dimension. L'algorithme Buckley-James combiné à une pénalité Lasso apporte une solution au problème et montre un avantage dans la prédiction de la réponse face aux techniques naïves.

Dans ce chapitre, nous avons démontré l'apport de la prise en compte de la censure pour des problèmes de prédiction. Nous souhaitions, en effet, dans un premier temps évaluer les méthodes et les algorithmes dans la situation la plus simple.

Toutefois, si la problématique est l'identification des variables pertinentes, l'adaptive Lasso ou le Bolasso sont plus appropriés. L'algorithme de Buckley-James a l'avantage de pouvoir être adapté à d'autres pénalisations. Notons cependant, que le méthode Bolasso implique des boucles supplémentaires dans l'algorithme, afin d'obtenir des proportions de sélection de variables. De plus, la paramétrisation de l'adaptive Lasso n'est pas une question évidente. Nous pourrions utiliser le maximum de vraisemblance pour le modèle censuré de manière univarié (une variable par modèle) et utiliser les estimations du maximum de vraissemblance comme pondération de l'adaptive Lasso. La question sur la définition de la variance de manière itérative est encore ouverte.

La question traitée dans ce travail est peut-être moins d'actualité aujourd'hui dans le cadre du VIH (avec le seuil de l'ordre de 20 et de nombreuses valeurs en dessous du seuil), mais l'idée peut être généralisée à d'autres problèmes en biologie, écologie,... avec ou sans l'hypothèse de normalité (la version non paramétrique ayant été implémentée et étudiée existant également) [Wei et al., 2018; Keizer et al., 2015].

4 Analyse de données de microbiote : état de l'art

Dans ce chapitre, nous introduisons la spécificité des données du microbiote humain et proposons un état de l'art des méthodes statistiques disponibles pour les analyser. Les techniques d'extraction de données et les informations disponibles sont présentées dans la Section 4.2. Leur pré-traitement est expliqué dans la Section 4.3. La Section 4.4 est consacrée à la revue de méthodes statistiques partant de l'analyse de diversité à l'extension des méthodes de pénalisation.

4.1 Introduction

Dès notre naissance, nous vivons en symbiose avec des centaines de milliards de micro-organismes. Ces microbes vivent à la surface et dans les profondeurs de nombreuses parties de notre corps. La bouche, la gorge, le nez, la peau, le vagin et tout le tube digestif représentent les différents **microbiotes**. Le mot microbiote désigne les espèces qui prédominent ou sont durablement adaptées à la surface et à l'intérieur d'un organisme. La diversité des espèces (nombre d'espèces différentes) est très variable entre les différents organismes, mais également au sein d'un même organisme [Huttenhower et al., 2012]. On parle de **microbiome** lorsque l'on fait référence aux génomes du microbiote.

La connaissance du microbiome a longtemps été limitée. Seules quelques bactéries (connues des chercheurs) pouvaient être cultivées in-vitro. Le développement des techniques de séquençage haut débit a donné un nouvel élan à la recherche. La nature des intéractions entre le microbiote et notre métabolisme est maintenant plus accessible à la recherche médicale. Le nombre de publications scientifiques sur le sujet a augmenté de manière exponentielle lors des dix dernières années (Figure 4.1).

Le microbiote intestinal a été le premier a être exploré à partir des nouvelles techniques de séquençage. On sait désormais que l'altération qualitative et fonctionnelle de la flore intestinale, connue sous le nom de dysbiose, est une piste sérieuse pour comprendre certaines maladies. Des liens significatifs entre la flore intestinale et différentes pathologies ont été mis en évidence. Nous pouvons citer le syndrome du côlon irritable ou la

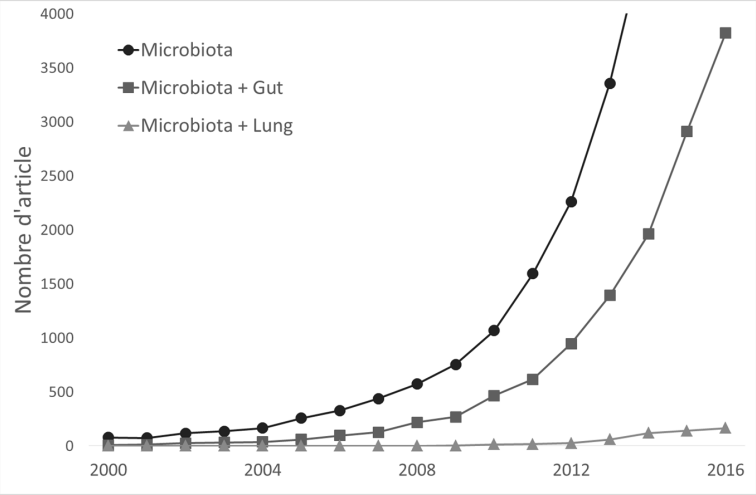

FIGURE 4.1 – Nombre d'articles indexés dans Medline où les mots clés dans le titre sont « Microbiota » (noir), « Gut Microbiota » (gris) et « Lung Microbiota » (gris clair). Consulté le 20 Février 2018.

maladie de Crohn, le diabète de type 2 [Qin et al., 2012] et les maladies cardiovasculaires [Koeth et al., 2013]. Le microbiote intestinal impacterait également les conditions psychologiques [Cryan and O'Mahony, 2011]. L'analyse de données de séquençage du microbiote a également montré des associations avec l'eczéma [Kong et al., 2012], le risque de naissance prématurée [Romero et al., 2014] ou encore des pathologies respiratoires chroniques [Charlson et al., 2011; Dickson et al., 2013; Morris et al., 2013; Marsland and Gollwitzer, 2014; Andréjak and Delhaes, 2015]. Il n'est toutefois pas possible dans la majorité de ces études de discerner si l'altération du microbiote est à l'origine de la maladie ou à l'inverse.

Les données microbiome ont une structure particulière. En effet, un microbe a une histoire évolutive et des informations sur ses liens de parenté sont disponibles. On parle de structure phylogénétique. De plus, les données issues du séquençage sont des données d'abondance, qui doivent être préalablement normalisées avant d'être explorées. Cette normalisation engendre un changement d'espace de définition qui soulève de nouveaux défis statistiques. Tenir compte de l'ensemble de ces complexités est un réel enjeu pour la recherche statistique.

4.2 Extraction des données et informations disponibles

La quantification de la composition du microbiote peut passer par le séquençage d'un seul gène spécifique [Tringe and Rubin, 2005], appelé séquençage de l'amplicon. En par-

ticulier le gène 16S est omniprésent dans le royaume des bactéries ou encore le gène ITS omniprésent dans le royaume des champignons. Le gène 16S est composé d'environ 1500 nucléotides, de sept régions conservées et de neuf régions hypervariables (Figure 4.2). L'ensemble permet donc théoriquement d'utiliser ce gène pour identifier et détecter toutes les espèces bactériennes.

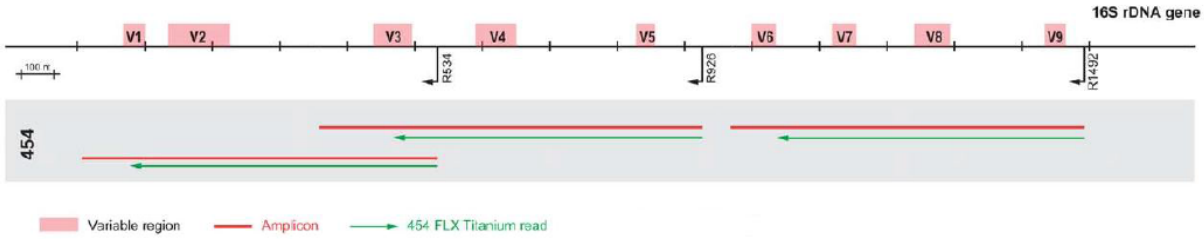

FIGURE 4.2 – Représentation schématique du gène 16S. Les zone rouges définies par la lettre V représentent les régions variables du gène ciblées par le séquençage.

Dans un premier temps, l'idée est d'isoler les brins d'ADN de toutes les bactéries correspondant aux régions variables du gène. L'étape suivante consiste à nettoyer les séquences brutes. Les séquences trop courtes ou contenant des erreurs dues à la technique de séquençage sont supprimées. Elles sont ensuite réparties et comptabilisées en Unité Taxonomique Opérationnelle (OTU) pour un certain niveau de similarité [Schloss et al., 2009; Caporaso et al., 2010a], le plus souvent à 97%. Les séquences sont comptées pour chacun des individus que l'on appelle données d'abondance. Le Tableau 4.1 montre un exemple de tableau d'abondance avec en ligne les individus et en colonne les OTU. Chaque cellule du tableau contient le nombre de séquences pour un OTU et pour un individu. Chaque OTU est représenté par une séquence ADN. Il peut se voir attribuer une lignée taxonomique (Figure 4.3) en la comparant à une base de données d'Acide ribonucléique ribosomique (ARNr) 16S bactérienne connue. La lignée taxonomique est composée de plusieurs rangs taxonomiques représentant chacun le niveau hiérarchique de la classification scientifique du monde vivant allant du règne à l'espèce. Les OTU peuvent donc être utilisés pour classer les espèces taxonomiques (Tableau 4.2), bien qu'elles ne représentent pas précisément les espèces bactériennes (par manque de sensibilité des techniques dans l'identification). Il est possible d'agréger les OTU du même Genre et analyser les abondances au niveau du Genre, ce qui est plus robuste à l'erreur de séquençage et peut réduire considérablement le nombre de variables dans les analyses (si un intérêt particulier n'est pas porté au niveau de l'espèce). De même, en utilisant des bases de données taxonomiques connues, les OTU peuvent être affectés à des Familles, des Ordres, des Classes ou des Phylums.

Une alternative au séquençage de l'amplicon est de séquencer tout l'ADN génomique microbien, une technique dans laquelle l'ADN est extrait de l'environnement et cisaillé

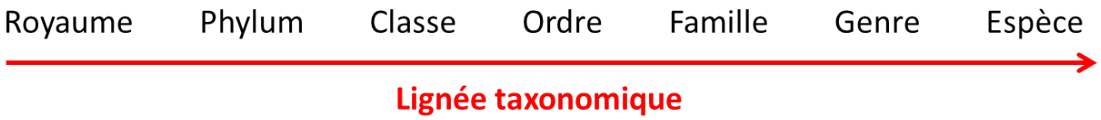

FIGURE 4.3 – Lignée taxonomique

ID	OTU 1	OTU 2	OTU 3	OTU 4	OTU 5	OTU 6	OTU 7	OTU 8	OTU 9	....	OTU q	Total
Sample 1	0	0	12	0	447	0	0	0	17	....	13	7538
Sample 2	0	0	1	0	232	0	0	0	2	....	8	7332
Sample 3	0	0	3	0	118	0	0	0	0	....	0	10099
Sample 4	0	0	8	0	50	0	0	0	0	....	1	12258
Sample 5	3	0	6	1	23	0	0	0	10	....	0	5106
Sample 6	0	0	35	1	480	0	0	0	54	....	0	7092
Sample 7	0	7	72	0	43	0	0	0	1	....	1	9108
Sample 8	0	1	69	0	177	0	0	0	0	....	10	5308
Sample 9	0	0	0	1	5	0	0	0	0	....	0	6054
Sample 10	0	0	9	0	7	0	0	0	2	....	0	10675
Sample 11	2	0	121	2	98	0	0	0	0	....	0	7457
Sample 12	7	0	2	1	4	0	0	0	0	....	0	14751
Sample 13	0	0	42	0	181	0	0	0	15	....	0	8741
Sample 14	0	0	173	0	343	0	0	375	5	....	2	8994
....	....	....	....	....	....	....	....	....	....	....	....	....
Sample n	2	0	63	0	164	5	0	0	0	....	0	6187
Total	15	11	1059	67	4427	8	3	477	114	....	37	261061

Tableau 4.1 – Exemple de table d'abondance pour un échantillon de taille n et une précision de q OTU. Les valeurs présentées dans le tableau sont fictives.

au hasard en plus petits fragments. Cette technique est appelée « *Shotgun metagenomic sequencing* ». Cette méthode ne sera pas décrite dans cette thèse, plus de détails sont disponibles dans la revue de Li [2015].

4.3 Pré-traitement

La grande proportion de zéros dans les données d'abondance est courante [McDonald et al., 2012; Paulson et al., 2013]. Ce nombre excessif de zéros est dû au fait que les nombres d'OTU sont dépendants du sujet, c'est-à-dire que la composition est unique à chaque sujet. Afin d'éviter des données trop clairsemées, les OTU présents dans moins de 10% des individus sont soit résumés au rang taxonomique supérieur, soit supprimés de l'analyses. Le plus souvent, la deuxième solution est utilisée. La dimension du problème et la proportion de zéros sont assurément réduites. Cependant, la perte d'information engendrée peut amener à des conclusions erronées [McMurdie and Holmes, 2014].

Le nombre total de séquences est souvent très hétérogène entre les individus à cause de la technique de séquençage mais également la variabilité biologique. La comparaison d'abondance brute chez deux individus peut ainsi amener à de fausses interprétations. La normalisation des données permet alors de relativiser les abondances brutes pour les rendre comparables entre échantillons. Se ramener à une abondance relative est l'approche

OTU	Kingdom	Phylium	Class	Order	Family	Genus
OTU 1	Bacteria	NA	NA	NA	NA	NA
OTU 2	Bacteria	Actinobacteria	Actinobacteria	NA	NA	NA
OTU 3	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Actinomyces
OTU 4	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Corynebacteriaceae	Corynebacterium
OTU 5	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Micrococcaceae	Rothia
OTU 6	Bacteria	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Alloscardovia
OTU 7	Bacteria	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium
OTU 8	Bacteria	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Scardovia
OTU 9	Bacteria	Actinobacteria	Coriobacteriia	NA	NA	NA
...	...	...	...	...	...	...
OTU q	Bacteria	Tenericutes	Mollicutes	Mycoplasmatales	Mycoplasmataceae	Mycoplasma

Tableau 4.2 – Exemple de lignée taxonomique pour chacun des OTU correspondant à la table d'abondance 4.1. Pour certaine bactérie, il n'est pas possible d'apporter de précisions sur un niveau inférieur au royaume, par exemple. Un OTU représentant le royaume permet alors de cataloguer l'ensemble de bactéries qui n'ont pas d'autre précision sur le niveau phylum ni aux niveaux inférieurs.

la plus utilisée dans l'étude des données de microbiote. Chaque abondance brute est divisée par le nombre total de séquences de l'individu pour obtenir une table d'abondances relatives dont la somme de chaque ligne est égale à 1. Prenons un exemple simple : on considère deux individus A et B dont leur nombre total de séquences est de 3489 et 2589 respectivement. Parmi ces séquences, 790 et 634 proviennent de l'OTU j chez l'individu A et chez l'individu B, respectivement. L'analyse directe de ces chiffres nous indique que cet OTU est plus présent chez le premier patient que chez le deuxième. Cependant, si on relativise par rapport au nombre total de séquences de chacun des patients, on remarque que l'OTU est présent en plus grande proportion chez l'individu B ($634/3489 = 25\%$) que chez l'individu A ($790/2589 = 23\%$). Selon la mesure utilisée, les conclusions sont différentes. D'autres normalisations ont été proposées dans la littérature. Une description détaillée et une comparaison de ces différentes normalisations sont décrites dans les revues de Thorsen et al. [2016] et Weiss et al. [2017].

La normalisation des données permet de comparer les échantillons mais certaines espèces peuvent toutefois dominer la composition et influencer les résultats de l'analyse. L'OTU avec la plus grande abondance relative est considéré comme un OTU dominant lorsque celui-ci a une abondance relative égale ou supérieure à deux fois l'abondance du deuxième OTU le plus abondant [Carmody et al., 2013]. La transformation logarithmique $\log(x)$, appliquée après la gestion des abondances nulles, est régulièrement utilisée et permet de réduire l'effet des OTU dominants.

4.4 État de l'art des méthodes d'analyse de données issues du microbiote

L'analyse de données de microbiote se décompose en plusieurs étapes :

A. L'**analyse de diversité** (diversité α et β) : Combien y a-t-il d'espèces différentes dans l'échantillon ? Quelles sont les espèces dominantes de notre échantillon ? Existe-t-il des structures phylogénétiques différentes ?

B. L'**analyse de corrélation** : Existe-t-il des co-occurences significatives entre espèces ?

C. L'**analyse différentielle d'abondance (tests statistiques, analyse de covariance, identification et prédiction)** : La composition du microbiote est-elle différente selon des groupes d'individus ? Quels OTU sont différentiellement abondants entre des groupes d'individus ? Quelles sont les covariables (variables explicatives) significativement associées à la composition du microbiote ? Quelles sont les OTU associés à un statut clinique ? A partir d'individus ayant une composition similaire d'un nouvel individu, est-il possible de prédire sa réponse clinique associée ?

4.4.1 Notations

— n le nombre d'individus
— q le nombre d'OTU considéré dans le n-échantillon
— p le nombre de covariables (autre que les OTU) considérées dans le n-échantillon
— Y_i la réponse clinique de l'individu i
— $\mathbf{Y} = (Y_1, \cdots, Y_n)^\top$ le vecteur réponse
— $Z_{ij} \in \mathbb{N}^+$ l'abondance brute (comptage) de l'OTU j chez le patient i
— $\tilde{Z}_{ij} = \frac{Z_{ij}}{\sum_{j=1}^{q} Z_{ij}} \in [0, 1]$ l'abondance relative de l'OTU j chez le patient i
— $\mathbf{Z} = (\mathbf{Z}_1, \cdots, \mathbf{Z}_n)^\top$ et $\tilde{\mathbf{Z}} = \left(\tilde{\mathbf{Z}}_1, \cdots, \tilde{\mathbf{Z}}_n\right)^\top$ les matrices d'abondance respectivement brute et relative de taille $n \times q$.
— $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_n)^\top$ la matrice de covariables de taille $n \times p$

4.4.2 Analyse de diversité

La **diversité** biologique fait référence à la composition du microbiote étudié. Les études sur sujets sains [Huttenhower et al., 2012] ont montré qu'un microbiote équilibré et diversifié est associé au bon fonctionnement de l'environnement dans lequel il se trouve (les voies aériennes, par exemple). A l'inverse, un microbiote non diversifié et déséquilibré est

associé à un dysfonctionnement de l'environnement [Huttenhower et al., 2012]. On parle de microbiote déséquilibré lorsque celui-ci est dominé anormalement par une espèce ou à l'inverse lorsqu'une espèce est anormalement absente (toutefois, la notion de déséquilibre ou dysbiose ne fait pas l'unanimité Hooks and O'Malley [2017]). A ce jour, il est nécessaire de garder une nuance dans la notion de microbiote équilibré ou déséquilibré par le manque d'information confirmatoire à ce sujet. En effet, la composition d'un microbiote est très variable au cours du temps et il est donc difficile de définir une composition « standard », même chez des sujets sains.

Évaluer la diversité d'un microbiote est complexe. Cependant, il existe des indicateurs simples tels que le nombre d'espèces différentes présentes ou encore le nombre d'individus ayant des compositions différentes. Nous présentons dans cette partie les méthodes proposées dans la littérature pour explorer la diversité d'un microbiote.

4.4.2.1 Diversité α

La diversité α permet de mesurer :

— la « **richesse spécifique** » représentant le nombre d'espèces différentes S dans le microbiote étudié pour un individu donné et

— l'« **équitabilité** » représentant la régularité de la distribution d'une espèce au sein du microbiote pour un individu donné.

Deux échantillons ayant le même nombre d'espèces différentes peuvent avoir une équitabilité différente.

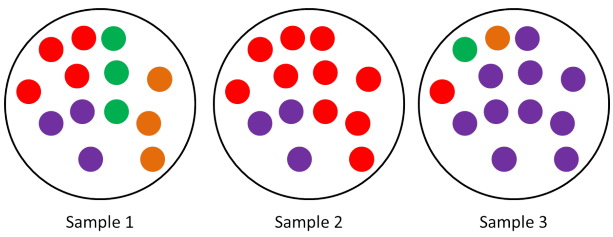

FIGURE 4.4 – Illustration d'un exemple de diversité dans trois échantillons

La Figure 4.4 représente un exemple de la composition du microbiote chez trois individus. Chaque couleur désigne une espèce différente. Les individus 1 et 3 ont le même nombre d'espèces différentes mais avec une distribution différente. L'individu 1 contient quatre espèces différentes dont la distribution des abondances de chacune des espèces est semblable. L'individu 3 contient également quatre espèces différentes mais la distribution de chacune des espèces n'est pas régulière. En effet, l'espèce violette est plus abondante par rapport aux autres. Les individus 2 et 3 n'ont pas le même nombre d'espèces mais

ont en revanche une distribution similaire.

Soit S le nombre d'espèces existantes au sein d'une population. Dans un échantillon de n individus, on note s^n le nombre d'espèces différentes observées, n_j le nombre d'individus ayant l'espèce j et p_j la probabilité qu'un individu ait l'espèce j.

La richesse spécifique peut être estimée par des indicateurs non-paramétriques qui visent à extraire le maximum d'information de la distribution de n_j pour estimer le nombre d'espèces non observées. Plus particulièrement, l'indicateur de **Chao1** estime le nombre d'espèces non observées à partir du nombre d'espèces observées une fois et du nombre d'espèces observées exactement deux fois. On note s_k^n le nombre d'espèces observées k fois dans un échantillon de n individus. La probabilité que l'espèce j soit observée k fois est celle d'une distribution binomiale. L'espérance du nombre d'espèces observées k fois, $\mathbb{E}(s_k^n)$, s'écrit :

$$\mathbb{E}(s_k^n) = \binom{n}{k} \sum_{j=1}^{s^n} p_j^k \left(1 - p_j\right)^{n-k}$$

L'espérance du nombre d'espèces non observées $\mathbb{E}(s_0^n)$ est donc égale à $\sum_j (1-p_j)^n$ et on peut démontrer que :

$$\mathbb{E}(s_0^n) \geq \frac{n-1}{n} \frac{\left[\mathbb{E}(s_1^n)\right]^2}{2\mathbb{E}(s_2^n)}$$

En remplaçant les espérances par les valeurs observées, l'indice de Chao1 s'écrit :

$$\hat{S}_{Chao1} = s^n + \frac{(n-1)(s_1^n)^2}{2ns_2^n}$$

Le coin du UseR : Le package `fossil` développé par Vavrek [2011] pour l'étude des espèces écologiques permet d'obtenir l'estimation de l'indice de Chao1 par la fonction `fossil()`.

D'autres indicateurs de richesse spécifique existent, tel que l'estimateur **Chao2**, similaire à Chao1 mais ne considère que la présence de l'espèce. L'estimateur *Abundance-based coverage estimator* (**ACE**) prend en compte le coefficient de variation de la distribution des fréquences $(\hat{p}_j)$. Plus les probabilités sont hétérogènes, plus le nombre d'espèces non observées sera grand. Plus de détails sont disponibles dans le cours de Marcon [2015].

Ces indices ne prennent pas en compte la distribution des espèces au sein de l'échantillon. Comme nous l'avons vu dans l'exemple précédent, deux individus peuvent avoir le même nombre d'espèces différentes mais des distributions distinctes. Les indices de **Simpson** et **Shannon** permettent de mesurer la richesse et l'équitabilité.

L'indice de Simpson se calcule de la façon suivante :

$$E = 1 - \sum_{j=1}^{S} p_j^2$$

Il peut être interprété comme la probabilité que deux séquences tirées au hasard soient d'espèces différentes. Il est compris dans l'intervalle $[0, 1]$. Sa valeur diminue avec la régularité de la distribution : $E = 0$ signifie qu'une seule espèce a une probabilité de 1 et $E = 1$ est atteint pour un nombre infini d'espèces, de probabilités nulles.

L'indice de Shannon, aussi appelé indice de Shannon-Weaver ou Shannon-Winer, ou simplement « entropie » est dérivé de la théorie de l'information qui consiste à supposer qu'une espèce de faible abondance peut apporter autant, voir plus, d'information qu'une espèce fortement abondante. L'indice de Shannon se calcule de la façon suivante :

$$H = -\sum_{j=1}^{S} p_j \log p_j$$

En théorie, cet indice n'a pas de limite supérieure. Il est difficilement interprétable seul.

Le coin du UseR : Le package `vegan` développé par Oksanen et al. [2017] permet de calculer les estimateurs de Shannon et Simpson par la fonction `diversity()` (argument `index = shannon"` ou `index = "simpson"`)

4.4.2.2 Diversité β

Deux individus ayant une diversité α similaire peuvent avoir des **structures phylogénétiques** différentes. Nous pouvons traduire cette notion par le fait que les espèces présentes chez les deux individus sont différentes et/ou éloignées dans la phylogénie. La Figure 4.5 représente l'abondance de trois individus A, B et C sur un arbre phylogénétique. Les individus A et B ont trois espèces différentes avec une distribution d'abondance similaire. En revanche, l'individu C n'a qu'une seule espèce. Les individus A et B ont donc une diversité α semblable par rapport à l'individu C. Cependant, en terme de structure phylogénétique, l'individu B est plus proche de C que de A car les espèces présentes dans B et dans C ont plus de branches en commun.

La notion de diversité β introduite par Whittaker [1960] permet de comparer ces différentes structures entre des groupes d'individus en calculant des distances. Dans ce cas, ces distances sont associées à des **dissimilarités**. On considère deux types :

— la **distance non phylogénétique** ne prenant en compte que la co-présence ou non d'individus au niveau des feuilles de l'arbre phylogénétique. Deux individus ayant deux espèces proches dans l'arbre phylogénétique vont avoir une même distance phylogénétique que deux individus ayant deux espèces éloignées dans l'arbre.

— la **distance phylogénétique** prenant en compte la co-présence de deux individus au niveau des feuilles mais également à tous les rangs taxonomiques de l'arbre. Dans ce cas, les individus ayant des espèces proches dans l'arbre vont avoir une distance inférieure à deux individus ayant des espèces éloignées.

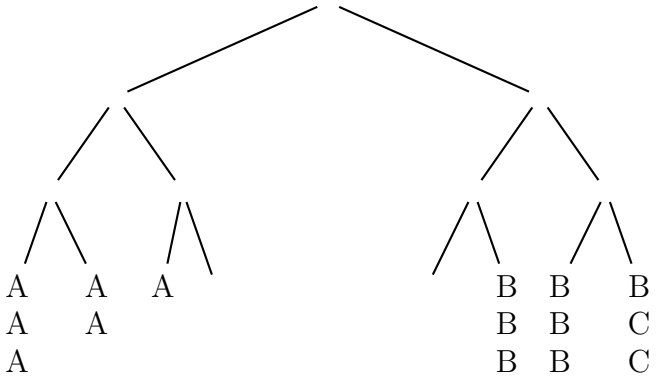

FIGURE 4.5 – **Exemple de répartition des abondances chez trois individus sur un arbre phylogénétique.** Les feuilles de l'arbre représentent une espèce et les lettres A, B et C montrent l'abondance pour chacune des espèces. Le nombre de lettres à chaque feuille montre l'abondance brute de l'espèce chez cet individu.

Les distances non phylogénétiques les plus courantes dans l'étude du microbiote sont la distance $\mathbf{D}^{(BC)}$ de **Bray-Curtis** [Bray and Curtis, 1957] et la distance $\mathbf{D}^{(J)}$ de **Jaccard** [Chao et al., 2005] définies comme :

$$d_{ii'}^{(BC)} = 1 - \frac{2\sum_{j=1}^{q}\min(Z_{ij}, Z_{i'j})}{\sum_{j=1}^{q}(Z_{ij} + Z_{i'j})} \qquad d_{ii'}^{(J)} = 1 - \frac{2\sum_{j=1}^{q}\min(Z_{ij}, Z_{i'j})}{\sum_{j=1}^{q}\max(Z_{ij}, Z_{i'j})}$$

avec $d_{ii'}$ la distance entre l'individu i et i'. Si on reprend notre exemple associé à l'arbre de la Figure 4.5, les distances de Bray-Curtis entre A et B et entre B et C sont égales à :

$$\begin{aligned} d_{AB}^{(BC)} &= 1 - \frac{2\left(\min(3,0) + \min(2,0) + \min(1,0) + \min(0,3) + \min(0,3) + \min(0,1)\right)}{(3+0) + (2+0) + (1+0) + (0+3) + (0+3) + (0+1)} \\ &= 1 - \frac{2 \times 0}{13} = 1 \\ d_{BC}^{(BC)} &= 1 - \frac{2\left(\min(3,0) + \min(3,0) + \min(1,2)\right)}{(3+0) + (3+0) + (1+2)} = 1 - \frac{2 \times 1}{9} = 0.78 \end{aligned}$$

Ces distances sont qualifiées de « pondérées » lorsqu'elles sont calculées à partir des abondances relatives et de « non pondérées » quand les abondances brutes sont utilisées.

Le coin du UseR : Le package `vegan` développé par Oksanen et al. [2017] permet de calculer les matrices de distances de Bray-Curtis et de Jaccard par la fonction `vegdist()` (argument `method = "bray"` ou `method = "jaccard"`).

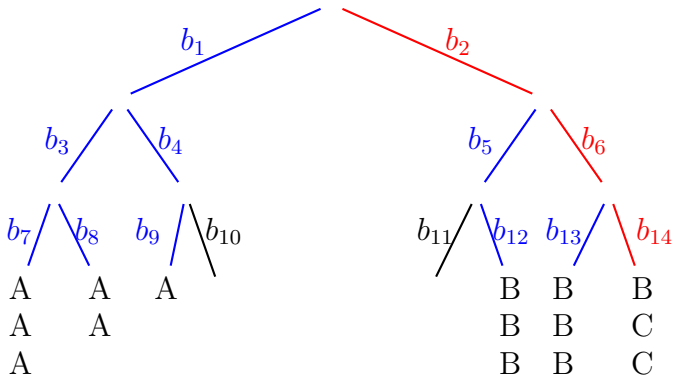

FIGURE 4.6 – **Exemple de répartition des abondances sur un arbre phylogénétique.** Les feuilles de l'arbre représentent chacune une espèce et les lettres A, B et C montrent l'abondance pour chacune des espèces pour trois individus : A, B et C. Le nombre de lettres à chaque feuille montre l'abondance brute de l'espèce chez cet individu. b_m avec $m = 1, \cdots, M$ représente la longueur de la branche m. Les taxons rouges représentent les taxons communs entre deux individus et les taxons bleus, les taxons présents chez un seul individu.

Lozupone and Knight [2005] ont proposé la distance **UniFrac** pour prendre en compte la phylogénie des données. Elle satisfait toutes les propriétés mathématiques pour une distance (positive et inégalité triangulaire). La distance UniFrac exploite les différents degrés de similarité entres les séquences, contrairement aux précédentes. Elle mesure la distance entre des ensembles de taxons dans un arbre phylogénétique comme la fraction de longueur de branche de l'arbre (Figure 4.6). Trois versions de distance Unifrac sont proposées [Lozupone and Knight, 2005; Lozupone et al., 2007; Fukuyama et al., 2012; Chen et al., 2012a], « non pondérée » ($\mathbf{D}^{(U)}$), « pondérée » ($\mathbf{D}^{(WU)}$) et « généralisée » ($\mathbf{D}^{(\xi)}$)

$$d_{ii'}^{(U)} = \frac{\sum_{m=1}^{M} b_m |\mathbb{1}_{(Z_{im}>0)} - \mathbb{1}_{(Z_{i'm}>0)}|}{\sum_{m=1}^{M} b_m} \qquad d_{ii'}^{(WU)} = \frac{\sum_{m=1}^{M} b_m |\tilde{Z}_{im} - \tilde{Z}_{i'm}|}{\sum_{m=1}^{M} b_m (\tilde{Z}_{im} + \tilde{Z}_{i'm})}$$

$$d_{ii'}^{(\xi)} = \frac{\sum_{m=1}^{M} b_m (\tilde{Z}_{im} + \tilde{Z}_{i'm})^{\xi} |\frac{\tilde{Z}_{im} - \tilde{Z}_{i'm}}{\tilde{Z}_{im} + \tilde{Z}_{i'm}}|}{\sum_{m=1}^{M} b_m (\tilde{Z}_{im} + \tilde{Z}_{i'm})^{\xi}}$$

avec M le nombre de branches dans l'arbre et b_m la longueur de la branche m. La longueur des branches indique un changement génétique, c'est-à-dire que plus la branche est longue, plus il y a eu de changement génétique (ou de divergence).

Le coin des UseR : Le package `Gunifrac` développé par Chen et al. [2012b] permet de calculer les distances UniFrac par la fonction `GUnifrac()`.

L'analyse en coordonnées principales (PCoA), un cas particulier de *Multidimensional scaling* (MDS), permet d'explorer et visualiser des similarités ou dissimilarités d'un jeu de données [Paliy and Shankar, 2016; Ramette, 2007]. C'est une généralisation de l'analyse en composantes principales (PCA) quand la distance n'est pas euclidienne on peut également la voir comme une PCA à noyau [Mariette and Villa-Vialaneix, 2017]. . Dans le même principe qu'une PCA, les résultats sont des vecteurs de coordonnées principales. Chacune de ces composantes explique un pourcentage de variance. Le but est de projeter les données sur un nombre réduit de composantes avec un pourcentage de variance expliquée maximum. La PCoA dépend d'une matrice de distance $\mathbf{D}$ de dissimilarité et son interprétation est simple : plus les individus sont proches dans la projection, plus leurs structures phylogénétiques sont similaires. A partir de groupes d'individus définis selon un phénotype, la PCoA permet de visualiser si les structures phylogénétiques sont discriminantes pour ces groupes.

Le coin des UseR : Le package `ape` développé par Paradis et al. [2004] permet d'effectuer une PCoA sur une matrice de distance $\mathbf{D}$ (Bray-Curtis, Jaccard, UniFrac) par la fonction `pcoa()`. La fonction `plot` est ensuite utilisée pour visualiser les résultats.

Anderson et al. [2011] ont proposé une revue des analyses de la diversité β sous forme de guide. Seule la notion de variation de la composition entre les différents groupes d'individus y est abordée.

4.4.3 Analyse de corrélation

4.4.3.1 Données de composition

La normalisation des abondances brutes en abondances relatives caractérise les nouvelles données de contraintes mathématiques : elles sont positives et somment à 1. On les appelle les **données de composition** ou *compositional data* (CoDa) [Aitchison, 1986]. D'un point de vue mathématique, cette normalisation est appelée *fermeture* et engendre un nouvel espace de définition. Soit $\mathbf{Z}_i = (Z_{i1}, \cdots, Z_{iq})^\top \in \mathbb{R}_+^q$ le vecteur des abondances brutes pour les OTU du patient i. La fermeture est définie tel que :

$$\left(\frac{Z_{i1}}{\sum_{j=1}^q Z_{ij}}, \cdots, \frac{Z_{iq}}{\sum_{j=1}^q Z_{ij}}\right) = \left(\tilde{Z}_{i1}, \cdots, \tilde{Z}_{iq}\right)^\top = \tilde{\mathbf{Z}}_i \qquad i = 1, \cdots, n$$

$\tilde{\mathbf{Z}}_i$ représente le vecteur d'abondance relative et appartient désormais à l'espace $\mathcal{S}^q$ connu sous le nom de **Simplex** et défini tel que :

$$\mathcal{S}^q = \left\{ \tilde{\mathbf{Z}}_i = \left(\tilde{Z}_{i1}, \cdots, \tilde{Z}_{iq}\right)^\top : \ \tilde{Z}_{ij} > 0, \ j = 1, \cdots, q; \ \sum_{j=1}^q \tilde{Z}_{ij} = 1 \right\}$$

Pour $q = 3$, les données sont souvent visualisées dans un triangle équilatéral (**diagramme ternaire**) dont les sommets représentent les 3 composantes étudiées, par exemple, trois espèces de bactéries. La composition globale est définie par le barycentre des trois sommets affectés de coefficients. Ces derniers sont définis comme les abondances relatives de chacune des composantes. Plus le barycentre est proche d'un sommet, plus la composante associée est abondante relativement aux autres. La Figure 4.7 montre un exemple de représentation de diagramme ternaire.

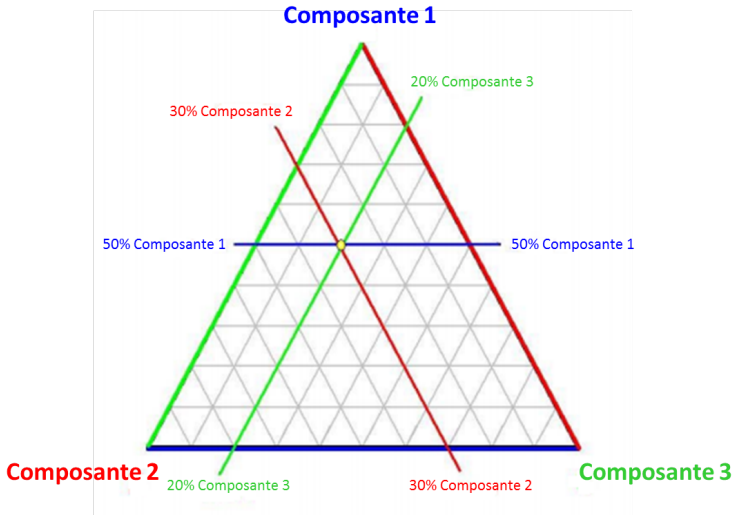

FIGURE 4.7 – Exemple de diagramme ternaire pour trois composantes

Le coin du UseR : Le package `ade4` développé par Dray et al. [2007] pour des données écologiques permet de construire le diagramme ternaire par la fonction `triangle.plot`

La corrélation standard ne peut pas être calculée directement lorsqu'il s'agit de CoDa car les composantes sont mathématiquement dépendantes (dont la somme vaut 1). Les fonctions de transformations applicables doivent répondre à trois conditions définies par Aitchison [Aitchison, 1986] :

- Une fonction $f(.)$ de transformation est **invariante à l'échelle** si pour toutes valeurs de $\lambda \in \mathbb{R}^+$ et pour toutes compositions $\tilde{\mathbf{Z}}_i \in S^q$, la fonction satisfait $f(\lambda\tilde{\mathbf{Z}}_i) = f(\tilde{\mathbf{Z}}_i)$. Cette condition est vraie si $f(.)$ est seulement une fonction log-ratio.
- Une fonction de transformation est **invariante par permutation** lorsque le résultat reste inchangé après changement de place des éléments dans la composition.
- Une fonction de transformation doit respecter **la cohérence de la sous-composition**, i.e. la distance entre deux compositions doit être plus grande que des distances entre toutes leurs sous-compositions respectives.

Le trois principales transformations de type log-ratio proposées dans la littérature sont les suivantes :

— La transformation **ALR** (*Additive log-ratio*) [Aitchison, 1986]. Elle renvoie dans un sous espace $\mathbb{R}^{q-1}$

$$\mathcal{S}^q \to \mathbb{R}^{q-1}, \quad \text{alr}(\tilde{\mathbf{Z}}_i) = \left(\log \frac{\tilde{Z}_{i1}}{\tilde{Z}_{iq}}, \cdots, \log \frac{\tilde{Z}_{iq-1}}{\tilde{Z}_{iq}}\right)$$

Une des composantes doit être préalablement choisie comme référence. Elle servira de dénominateur. L'OTU le plus abondant est souvent utilisé dans les analyses de données de microbiote.

— La transformation **CLR** (*Centered log-ratio*) [Aitchison, 1986]. Elle est régulièrement utilisée pour les techniques basées sur une métrique. Elle renvoie également à un sous espace $\mathbb{R}^q$ qui facilite l'interprétation car le nombre d'OTU reste inchangé après transformation. En revanche cette transformation peut amener à de distributions déformées et généré une matrice singulière.

$$\mathcal{S}^q \to \mathbb{R}^q, \quad \text{clr}(\tilde{\mathbf{Z}}_i) = \left(\log \frac{\tilde{Z}_{i1}}{g\left(\tilde{Z}_i\right)}, \cdots, \log \frac{\tilde{Z}_{iq-1}}{g\left(\tilde{Z}_i\right)}\right)$$

with $g\left(\tilde{\mathbf{Z}}_i\right) = \prod_{j=1}^{q} \tilde{Z}_{ij}^{1/q}$.

— La transformation **ILR** (*Isometric log-ratio*) [Egozcue et al., 2003]. Soit $\{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_{q-1}\}$ une base orthonormale de S^q et on considère la matrice $\mathbf{\Psi}$ de taille $q-1 \times q$ où les lignes sont $\text{clr}(\mathbf{e}_i)$. Une base orthonormale satisfait $\langle \mathbf{e}_i, \mathbf{e}_j \rangle_a = \delta_{ij}$ (δ_{ij} est le delta de Kronecker qui est nul quand $i \neq j$ et 1 quand $i = j$) avec $\langle , \rangle_a$ le produit scalaire dans la géométrie d'Aitchison [Aitchison, 1986] tel que $\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2q} \sum_i^q \sum_j^q \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}$ où $\mathbf{x}, \mathbf{y} \in S^q$. Une composition $\tilde{\mathbf{Z}}_i \in S^q$ peut alors être exprimée de la façon suivante :

$$\tilde{\mathbf{Z}}_i = \bigoplus_{j=1}^{q-1} \tilde{\mathbf{Z}}_{ij}^* \odot \mathbf{e}_{ij}, \qquad i = 1, \cdots, n$$

avec $\tilde{\mathbf{Z}}_i^* = \left(\tilde{\mathbf{Z}}_{i1}^*, \cdots, \tilde{\mathbf{Z}}_{iq-1}^*\right)$ le vecteur de coordonnées de $\tilde{\mathbf{Z}}_i$, $\bigoplus$ et $\odot$ sont les opérateurs de somme et de produit, respectivement, dans la géométrie d'Aitchison.

$$S^q \to \mathbb{R}^{q-1}, \quad \tilde{\mathbf{Z}}_i^* = \text{ilr}(\tilde{\mathbf{Z}}_i) = \left(\langle \tilde{\mathbf{Z}}_i, e_{i1} \rangle_a, \cdots, \langle \tilde{\mathbf{Z}}_i, e_{iq-1} \rangle_a\right)$$

Le coin du UseR : Le package `compositions` développé par van den Boogaart et al. [2010] permet d'effectuer les transformations ALR, CLR et ILR par les fonctions `alr()`, `clr()` et `ilr()` respectivement.

4.4.3.2 Gestion des zéros

Les transformations log-ratio requièrent des éléments non nuls. Comme décrit par Martın-Fernandez et al. [2011], il existe différents types de zéros dans l'analyse de données de compositions : les **arrondis**, les **absolus** (ou structurels) et les **zéros comptés**.

Les arrondis sont dus à un nombre de chiffres significatifs trop petit. Ce ne sont donc pas de « vrais » zéros. Pour les gérer, deux techniques existent. Martín-Fernández et al. [2003] proposent une méthode non-paramétrique appelée **remplacement multiplicatif**, considérée comme une stratégie d'imputation. Elle consiste à remplacer les zéros arrondis par une petite valeur et modifier les valeurs non nulles afin de respecter les hypothèses des données CoDa. Cependant, cette méthode a tendance à générer des corrélations artificielles entres les variables lorsqu'il y a des zéros sur la même ligne [Marks et al., 2011]. Pour éviter ce problème, Palarea-Albaladejo et al. [2007] ont proposé une méthode paramétrique basée sur l'algorithme *Expectation Maximisation* (EM), qui est souvent utilisé dans les problèmes de maximisation de la vraisemblance en présence de données manquantes. Le but est de combiner l'algorithme EM avec la transformation ALR qui génère des estimations adaptées pour les valeurs censurées par limite de quantification. Cette méthode, indépendante du choix du dénominateur dans la transformation ARL remplace les zéros d'arrondis par une imputation, conditionnellement aux valeurs observées. Plus de détails sont données dans Martín-Fernández and Thió-Henestrosa [2006]; Palarea-Albaladejo et al. [2007]; Martın-Fernandez et al. [2011].

Les zéros comptés représentent le nombre de cas où l'évènement n'a pas eu lieu. Cela correspond à des OTU qui ont une probabilité faible mais non nulle d'évènement. En moyenne, il faudra un nombre conséquent de sujet pour l'observer. Dans l'analyse de microbiote, les techniques de séquençage manquent de sensibilité dans la détection des espèces engendrant des zéros comptés. Pour les gérer, Pierotti et al. [2009] ont appliqué un traitement basé sur l'approche **Bayésienne multiplicative**. Soit s_i le nombre total de séquences de l'individu i et $\boldsymbol{\theta}_i$ son paramètre de probabilité associé dérivant d'une distribution multinomiale. L'idée est de supposer une distribution *a priori* sur $\boldsymbol{\theta}_i$ provenant d'une distribution de Dirichlet et de calculer une distribution *a posteriori* grâce au théorème de Bayes. Plusieurs types d'*a priori* peuvent être proposés et leur choix est discuté dans Martın-Fernandez et al. [2011].

Les zéros absolus sont considérés comme les « vrais » zéros. Cela correspond à des OTU qui ont une probabilité nulle d'évènement. L'absence d'une espèce dans le microbiome peut être associée à une certaine pathologie. Il est clair que remplacer ces zéros par une petite valeur n'est pas approprié car ils sont informatifs. Des modèles ont été proposés pour incorporer ces zéros absolus dans le processus de modélisation [Aitchison et al.,

2003; Bacon Shone et al., 2008]. Il est également possible d'interpréter ces zéros dans une certaine composante en tant qu'indicateur de deux sous-groupes d'intérêt différents : les observations avec une valeur à zéro dans cette composante par rapport à l'observation prenant une valeur positive à la place. En réalité, c'est un problème difficile qui reste ouvert. De plus, le choix de considérer des zéros absolus plutôt que des zéros comptés peut être discutable. Il doit être basé sur des hypothèses réelles qu'il est impossible de vérifier en pratique.

4.4.3.3 Analyse de corrélation dans le cas de la grande dimension

L'application d'une corrélation classique aux abondances brutes et aux abondances relatives (ou CoDa) peut conduire à des résultats différents, voire contradictoires. L'étude des corrélations sur les CoDa est un réel enjeu statistique. Les estimations de corrélation sont faussées par la dépendance présente entre les abondances relatives, engendrant des tendances de corrélation négative. Ainsi, les corrélations reflètent la nature compositionnelle des données plutôt que le processus biologique sous-jacent. De plus, le nombre d'OTU concernés étant souvent large ($q = n$ ou $q > n$), des méthodes d'analyse de corrélation adaptées à ce type de données ont été proposées.

Aitchison [1986] suggéra d'utiliser la variance du log-ratio des paires d'OTU (mesurés en termes d'abondances relatives).

$$t_{jk} \equiv Var\left(\log \frac{\tilde{\mathbf{Z}}_j}{\tilde{\mathbf{Z}}_k}\right) \equiv Var\left(-\log \frac{\tilde{Z}_k}{\tilde{Z}_j}\right) \equiv Var\left(\log \frac{\tilde{Z}_k}{\tilde{Z}_j}\right) \equiv t_{kj}$$

où t_{jk} représente la variation entre l'OTU j et l'OTU k. Lorsque les OTU sont parfaitement corrélés, $t_{jk} = 0$. Moins les OTU sont corrélés plus t_{ij} est grand. Cependant, on ne sait pas si un $t_{jk} = 0.1$ représente une corrélation faible, modérée ou forte. Il est donc difficile à interpréter dans l'absolu.

La méthode **CCREPE** (*Composionality Corrected Renormalization and Permutation*) [Faust et al., 2012] propose de construire B jeux de données permutés (au niveau des sujets) à partir du jeu de données réelles, de les normaliser (transformation ALR) et de calculer une matrice de corrélation pour chacun des jeux de données. Cette étape permet d'estimer la distribution des corrélations, sous l'hypothèse nulle d'absence de corrélations. B' échantillons bootstrap sont ensuite générés à partir du jeu de données réelles préalablement normalisé. Une matrice de corrélations est ensuite calculée pour chacun des échantillons bootstrap. Cette étape permet d'obtenir une distribution faisant référence à l'intervalles de confiance des corrélations observées. Les deux distributions sont ensuite comparées par un Z-test avec variance commune.

SparCC (*Sparse Correlation for CoDa*) [Friedman et al., 2012], **CCLasso** (*Correlation inference for CoDa through Lasso*) [Fang et al., 2015] et **REBACCA** (*Regularized*

Estimation of the BAsis Covariance based on Compositional data) [Ban et al., 2015] estiment t_{jk} et l'utilisent pour calculer la matrice de corrélation au lieu des corrélations de Pearson. Faisant l'hypothèse que la matrice de corrélation entre les OTU est parcimonieuse (i.e. que l'on suppose qu'il existe des corrélations nulles entre certains OTU), SparCC repère les corrélations aberrantes par rapport à un seuil fixé alors que CCLasso et REBACCA appliquent une pénalisation L_1 sur les corrélations. Ces méthodes comparent des liens directs entre des paires d'OTU. Cependant, les liens entre les OTU semblent plus complexes et des liens non directs peuvent exister entre OTU. **SPIEC-EASI** [Kurtz et al., 2015] (Sparse Inverse Covariance Estimation for Ecological Association Inference) utilise des méthodes de réseaux graphiques à la place des matrices de corrélations. Cette méthode met l'accent sur l'indépendance conditionnelle permettant de garder les corrélations indirectes avec un degré de séparation approprié.

PLNnetwork [Chiquet et al., 2019] se base sur un modèle Poisson log-normal multivarié. Ce modèle considère une variable latente de distribution gaussienne pour chacun des OTU étudiés. Regarder la structure de dépendances entre les OTU revient à étudier la structure de dépendance des variables latentes associées. Sur l'espace latent Gaussien, les méthodes standard basées sur des distances ou des corrélations sont valides. Pour cela, les auteurs utilisent les modèles graphiques développés par Lauritzen [1996] en l'adaptant à la distribution conjointe gaussienne des variables latentes et introduit une matrice de covariance que l'on suppose parcimonieuse.

Tsilimigras and Fodor [2016] ont proposé une revue et une comparaison de certaines de ces méthodes.

L'ensemble des méthodes citées ci-dessus et leurs implémentations sont rassemblées dans le Tableau 4.3.

Article	Description	Implémentation
Faust et al. [2012]	**CCREPE** : teste les corrélations par une méthode de permutations.	`ccrepe`
Friedman et al. [2012]	**SparCC** : Estime la matrice de variation $\mathbf{T} = (t_{jk})_{j,k}$ et repère les corrélations aberrantes suivant un seuil fixé.	`https://bitbucket.org/yonatanf/sparcc`
Fang et al. [2015]	**CCLasso** : Applique une pénalisation L_1 sur la matrice de corrélation.	`https://github.com/huayingfang/CCLasso`
Ban et al. [2015]	**REBACCA** : Applique une pénalisation L_1 sur la matrice de corrélation.	`http://faculty.wcas.northwestern.edu/hji403/REBACCA`
Kurtz et al. [2015]	**SPIEC-EASI** : Applique un modèle de réseaux graphiques pour prendre en compte l'ensemble des corrélations directes et indirectes entre les OTU.	`https://github.com/zdk123/SpiecEasi`
Chiquet et al. [2019]	**PLNnetwork** : Utilise un réseau de covariance parcimonieuse dans des graphiques.	`https://github.com/jchiquet/PLNmodels`

Tableau 4.3 – Revue des méthodes d'analyses de corrélation pour des données CoDa.

4.4.4 Analyse différentielle d'abondance

Ce type d'analyse permet de comprendre les liens entre des variables cliniques (ou non cliniques) et les OTU. La composition du microbiote est-elle différente entre des groupes de patients ? Les OTUs ont-il une distribution similaire entre des groupes de patients ?

4.4.4.1 Différence globale d'abondance

La PCoA, définie dans la sous section 4.4.2.2, permet d'explorer visuellement les différences de structure phylogénétique entre des groupes d'individus. Des tests sont proposés afin d'évaluer cette différence. L'ensemble des tests présentés dans la suite dépendent de la distance de dissimilarité **D** choisie.

Au début des années 90, Clarke [1993] a proposé une analyse des similarités (**ANOSIM**) entre des groupes d'individus. Inspiré de l'ANOVA, ANOSIM est un test non-paramétrique opérant sur les rangs des dissimilarités (où le rang de 1 correspond à la plus petite dissimilarité). La statistique de test R s'écrit :

$$-1 \leq R = \frac{r_B - r_w}{\frac{n(n-1)}{2}/2} \leq 1$$

avec r_B la moyenne des rangs des dissimilarités d'individus provenant de groupes différents et r_w la moyenne des rangs des dissimilarités d'individus provenant du même groupe. La matrice de dissimilarité est une matrice symétrique d'ordre n dont le sous espace vectoriel est de dimension $n(n+1)/2$. $R = 1$ signifie que toutes les paires d'individus provenant du même groupe sont plus similaires que les paires d'individus provenant de groupes différents. $R = 0$ est une valeur particulière où les moyennes sont égales. $R < 0$ est numériquement possible mais improbable. Pour déterminer la p-valeur allouée à R, on utilise la méthode de permutation. On construit B jeux de données, permutés sur l'assignation des individus aux groupes à partir du jeu de données réelles. La p-valeur est déterminée par :

$$p = \frac{\#\left\{|R^b| \geq |R|, b = 1, \cdots, B\right\}}{B}$$

où R^b est la statistique de test pour le jeu de données permuté b.

Le test **PERMANOVA** [Anderson et al., 2011], inspiré également de l'ANOVA, opère sur les dissimilarités entre les groupes. Soit $SS_T = \frac{1}{n}\sum_{i=1}^{n-1}\sum_{i'=i+1}^{n} d_{ii'}^2$ et $SS_W = \frac{1}{N}\sum_{i=1}^{n-1}\sum_{i'=i+1}^{n} d_{ii'}^2 \delta ii'$ avec d une matrice de distance de dissimilarité et $\delta ii'$ qui prend la valeur 1 quand les individus i et i' sont dans le même groupe et 0 quand ils sont dans des groupes différents. La statistique de test F s'écrit :

$$F = \frac{\left(SS_T - SS_W\right)/(G-1)}{SS_W/(n-G)}$$

où G est le nombre de groupes considérés. La p-valeur est également calculée à partir de la méthode de permutation décrite ci-dessus.

Le test **SPU** (*powered score*) [Pan et al., 2014] se base sur un modèle de régression linéaire :

$$Y_i = \beta_0 + \mathbf{X}_i^\top \boldsymbol{\beta} + \mathbf{Z}_i^\top \boldsymbol{\alpha} + \varepsilon_i \qquad i = 1, \cdots, n$$

avec $\mathbf{Y}$ le vecteur réponse définissant l'appartenance aux groupes, $\mathbf{X}$ la matrice des variables explicatives et $\mathbf{Z}$ la matrice d'abondance brute. Tester que les OTU ne sont pas associés à la variable réponse revient à tester $H_0 : \boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_q)^\top = \mathbf{0}$. Le vecteur du score est défini comme :

$$\begin{aligned}\mathbf{U} &= (U_1, \cdots, U_q) \\ &= \left(\sum_{i=1}^{n}(Y_i - \hat{Y}_i)Z_{i1}, \cdots, \sum_{i=1}^{n}(Y_i - \hat{Y}_i)Z_{iq}\right)\end{aligned}$$

où $\hat{Y}_i$ est la prédiction sous H_0. La statistique de test s'écrit alors :

$$T = \sum_{j=1}^{q} U_j^\gamma \underset{H_0}{\sim} \chi^2_{(q)}$$

avec $\gamma > 1$. Plus γ est grand, plus le test met d'importance sur les composantes les plus abondantes de U.

Le test **MiRKAT** (*Microbiome Regression-Based Kernel Association Test*) [Chen and Li, 2013a] est basé sur un modèle de régression semi-paramétrique tel que :

$$Y_i = \beta_0 + \mathbf{X}_i^\top \boldsymbol{\beta} + h(\tilde{\mathbf{Z}}_i) + \varepsilon_i \qquad i = 1, \cdots, n$$

où $h(.)$ est une fonction inconnue de l'espace de Hilbert représentant le lien entre la structure du microbiote et la variable réponse. Les coefficient de régression $\boldsymbol{\beta}$ et la fonction $h()$ sont estimés par minimisation de la fonction de log-vraisemblance négative pénalisée :

$$\underset{\boldsymbol{\beta}, h()}{\operatorname{argmin}} l(\mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}, h()) - \frac{1}{2}\lambda \alpha^\top \mathbf{K} \alpha$$

avec $h() = \sum_{i=1}^{n} \alpha_i K(., \mathbf{Z}_i)$ pour plusieurs α. $\mathbf{K}$ représente la matrice à noyau telle que $K_{ij} = K(\mathbf{Z}_i, \mathbf{Z}_j)$. λ est le paramètre de réglage contrôlant de manière simultanée l'ajustement du modèle et la complexité de la fonction $h()$. Le modèle de régression gaussien est généralisé aux modèles semi-paramétriques généralisés (incluant en autre la logistique). Tester que les OTU n'ont aucun effet sur la variable réponse revient à tester $H_0 : h(\tilde{\mathbf{Z}}) = 0$. La fonction $h(.)$ est estimée en remplaçant la formulation du noyau par un modèle linéaire mixte :

$$Y_i = \beta_0 + \mathbf{X}_i^\top \boldsymbol{\beta} + h_i + \varepsilon_i \qquad i = 1, \cdots, n$$

avec $\boldsymbol{\beta}$ le vecteur de coefficients des effets fixes, $\mathbf{h} = (h_i, \cdots, h_n) \sim \mathcal{N}(0, \tau\mathbf{K})$ les effets aléatoires spécifiques aux individus, $\mathbf{K}$ est la matrice de noyaux sur les individus et $\tau = \frac{1}{\lambda}$. Dans ce modèle, tester que les OTU n'ont aucun effet sur la variable réponse revient à tester la nullité de la variance de l'effet aléatoire $H_0 : \tau = 0$. Liu et al. [2008] ont proposé le test du score défini par la statistique de test :

$$Q = \frac{1}{2\hat{\sigma}^2}(\mathbf{Y} - \hat{\mathbf{Y}}_0)^\top \mathbf{K}(\mathbf{Y} - \hat{\mathbf{Y}}_0) \sim \text{Mélange de } \chi^2$$

où $\hat{\sigma}^2$ est l'estimation de la variance résiduelle sous H_0 et $\hat{\mathbf{Y}}_0$ est la prédiction sous H_0. $\mathbf{K}$ est construit à partir de la distance de dissimilarité par $\mathbf{K} = -\frac{1}{2}\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n}\right)\mathbf{D}\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n}\right)$. Chen and Li [2013a] ont proposé d'utiliser la distance Unifrac généralisée. Un test exact a également été proposé par [Zhan et al., 2017]. Le test **OMiAT** (*Optimal microbiome-based association test*) [Koh et al., 2017] prend le minimum entre les p-valeurs des tests SPU et MiRKAT.

Pour finir, LaRossa et al. [2012] ont proposé le test **HMP** (*Hypothesis Testing and Power Calculations for Comparing Metagenomic Samples from Human Microbiome Project*) basé sur la distribution multinomiale de Dirichlet (DM). DM est paramétrée par $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_q)$ représentant la fréquence moyenne des OTU. Si l'on souhaite comparer deux groupes d'individus A et B, l'hypothèse nulle est définie par $H_0 : \boldsymbol{\pi}_A = \boldsymbol{\pi}_B$ et la statistique de test associée s'écrit comme :

$$X = \left(\hat{\boldsymbol{\pi}}_A - \hat{\boldsymbol{\pi}}_B\right)^\top (\mathbf{S})^{-1} \left(\hat{\boldsymbol{\pi}}_A - \hat{\boldsymbol{\pi}}_B\right) \sim \chi^2$$

qui est une généralisation du test de Wald et $\mathbf{S}$ est une matrice diagonale représentant la dispersion des données.

L'ensemble de ces tests et leurs implémentations sont répertoriés dans le Tableau 4.4.

Article	Description	Packages R	Fonctions R
Clarke [1993]	**ANOSIM** : Test non-paramétrique basé sur les rangs des dissimilarités.	`vegan`	`anosim()`
Anderson et al. [2011]	**PERMANOVA** : Test non-paramétrique basé sur les distances entres les centres des groupes.	`vegan`	`adonis()`
Pan et al. [2014]	**SPU** : test d'association basé sur un modèle de régression linéaire.		
Chen and Li [2013a]	**MiRKAT** : Test d'association basé sur un modèle de régression semi-paramétrique incluant une composante non-paramétrique $h()$ qui relate des données de microbiote. $h()$ est défini comme un effet aléatoire dans un modèle mixte et suivant une distribution normale de moyenne 0 et de variance τK. K est défini comme un noyau à partir d'une matrice de distance de dissimilarité.	`MiRKAT`	`MiRKAT`
Koh et al. [2017]	**OMiAT** : Méthode robuste prenant le minimum des p-valeurs des tests SPU et MiRKAT	`OMiAT`	`OMiAT`
LaRossa et al. [2012]	**HMP** : Test paramétrique basé sur une distribution DM pour prendre en compte la grande dispersion des données d'abondance.	`HMP`	

Tableau 4.4 – Revue des tests de différences d'abondance globale utilisés dans l'analyse de données de microbiote

4.4.4.2 OTU différentiellement abondants

Les tests globaux, présentés à la sous section 4.4.4.1, permettent de tester si la structure phylogénétique du microbiote est différente selon des groupes d'individus. La deuxième étape consiste à répondre à la question suivante : Quels sont les OTU différentiellement abondants entre des groupes d'individus ? De nombreux tests ont été proposés et des revues ont été publiées [Thorsen et al., 2016; Weiss et al., 2017; Xia and Sun, 2017]. Elles rassemblent l'ensemble des tests de différences d'abondance univariés pour données de microbiote.

Les tests les plus simples sont le ***t*-test** de comparaison de moyennes (test de Welch) ou sa généralisation à plus de 2 groupes, l'ANOVA, sous condition de normalité des données et le test des rangs de **Wilcoxon**. **Lefse** [Segata et al., 2011] et **STAMP** [Parks et al., 2014] constituent des démarches d'analyse d'OTU différentiellement abondants à base de tests de Wilcoxon. Ces tests peuvent être utilisés sur les données d'abondance relative ou sur les données log-transformées. **Metastats** [White et al., 2009] estime la distribution nulle de manière non-paramétrique utilisant une méthode de permutation à partir de t-test.

Une autre approche consiste à considérer la distribution *a priori* des données d'abondance brute. Le plus souvent, pour modéliser des données de comptage, on utilise la distribution de Poisson. En effet, la loi de Poisson décrit le comportement du nombre d'évènements se produisant dans un volume fixé. Elle est paramétrée par λ qui est à la fois le paramètre de moyenne et de variance. Cependant, pour les données de séquençage, il est incohérent de supposer que la moyenne est égale à la variance dû à la sur-dispersion des données. La distribution Binomiale Négative (NB) est une alternative à la distribution de Poisson. La loi NB décrit le comportement du nombre d'échecs avant l'obtention de k succès de probabilité p dans une série d'épreuves de Bernoulli indépendantes et identiquement distribuées. Dans le cas de données de comptage, Z_{ij} (suivant une NB) se modélise par un modèle linéaire généralisé avec une fonction de lien logarithmique :

$$
\begin{aligned}
&Z_{ij} \sim NB(\text{moyenne} = \mu_{ij}, \text{dispersion} = \alpha_j) \qquad i = 1, \cdots, n \qquad j = 1, \cdots, q \\
&\mu_{ij} = s_i \gamma_{ij} \\
&\log \gamma_{ij} = \mathbf{X}_i^\top \boldsymbol{\beta}
\end{aligned}
$$

où s_i est le facteur de taille spécifique de l'individu i représentant le nombre de séquences. Il peut être estimé par une méthode de normalisation telle que *median-of-ratios* décrit dans Anders et al. [2012]. γ_{ij} est la vraie proportion de l'OTU j chez l'individu i. La dispersion est fonction de μ_{ij} telle que $\alpha_j = \mu_{ij} + \frac{\mu_{ij}^2}{l}$ où l est un paramètre de dispersion. Le modèle NB généralisé [Bolker et al., 2009] a longtemps été utilisé en écologie pour

modéliser l'abondance des espèces.

DESeq2 [Love et al., 2014] permet d'estimer s_i comme la médiane des ratios des données observées. En d'autres termes, les données brutes sont divisées par la moyenne géométrique de l'échantillon tel que :

$$\hat{s}_j = \underset{i}{\text{median}} \frac{Z_{ij}}{\left(\prod_{j=1}^{q} Z_{ij}\right)^{1/q}} \qquad i = 1, \cdots, n$$

edegR [Robinson et al., 2010] et **baySeq** [Hardcastle and Kelly, 2013] se basent sur le même modèle et proposent une procédure bayésienne pour l'estimation des paramètres.

Comme nous l'avons dit précédemment, les données de microbiote comportent un nombre excessif de 0. Pour cela, des modèles *Zero inflated* (ZI), peuvent être employés. Le modèle ZI est défini comme un mélange d'un point de masse à 0, $I_{\{0\}}(\mathbf{Z})$ et une distribution dont la densité est notée $f(\boldsymbol{\theta})$ où $\boldsymbol{\theta}$ est le vecteur de paramètres de la densité. La fonction de probabilité s'écrit :

$$f_{ZI}(Z_{ij}; s_i, \beta, \theta_j) = \pi_i(s_i).I_{\{0\}}(Z_{ij}) + (1 - \pi_i(s_i)) .f(Z_{ij}; \theta_j)$$

où π_i est le paramètre de mélange du modèle. Le nombre de zéros dépendant du nombre total de séquences, le paramètre $\pi_i(s_i)$ est paramétré à partir d'une régression logistique :

$$\log \frac{\pi_i}{1 - \pi_i} = \beta_0 + \beta_1 . \log(s_i)$$

Lors de l'analyse du microbiote, différents modèles ZI ont été introduits : les modèles Poisson (ZIP) et NB (ZINB) ont été comparés par Xue et al. [2016] alors que **metagenomeSeq** [Paulson et al., 2013] s'appuie sur un modèle gaussien (ZIG). L'ensemble des paramètres sont estimés par maximum de vraissemblance à l'aide d'un algorithme EM.

Le test **ALDEx2** (*ANOVA-like differential expression*) [Fernandes et al., 2013] a été développé spécifiquement pour le caractère compositionnel des données de microbiote. La première étape de la méthode consiste à convertir la table d'abondances brutes en une distribution de probabilités *a posteriori*. Cette conversion passe par un échantillonnage de Monte Carlo à partir de la distribution de Dirichlet pour chaque individu. La loi de Dirichlet étant une loi conjuguée, si on suppose que le modèle d'échantillonnage des individus suit une loi de Dirichlet, il en sera de même pour la distribution *a posteriori*. Un prior non informatif de 1/2 est utilisé pour modéliser la fréquence des abondances nulles (a priori on a autant de chances d'observer la caractéristique que de ne pas l'observer). Chaque abondance d'OTU est alors représentée par un vecteur de probabilités *a*

posteriori de taille M où M est le nombre d'instances de Dirichlet Monte Carlo échantillonnées. A la seconde étape, chaque instance de Dirichlet Monte Carlo est transformée par la transformation CLR. À l'étape suivante, des tests de comparaisons sont effectués sur chaque instance du vecteur entre les deux groupes. Chacune des M réalisations entre les conditions sont soumises à la fois à un t-test et à un test de Wilcoxon donnant deux vecteurs de p-valeurs. Chacune des M instances de p-valeur est corrigée pour des tests d'hypothèses multiples en utilisant l'approche du taux de fausse découverte (FDR) de Benjamini et Hochberg. La distribution *a posteriori* des p-valeurs obtenues et la distribution *a posteriori* des statistiques FDR peuvent être alors calculées pour les deux tests statistiques.

ANCOM (*Analyse of composition of microbiome*) [Mandal et al., 2015] teste l'hypothèse $H_{0_{jk}} : \mathbb{E}_A\left(\log\left(\frac{\tilde{Z}_j}{\tilde{Z}_k}\right)\right) = \mathbb{E}_B\left(\log\left(\frac{\tilde{Z}_j}{\tilde{Z}_k}\right)\right)$ pour $j > k$ La statistique de test utilisée est celle de l'ANOVA (si les hypothèses sont vérifiées) ou l'alternative non paramétrique (Wilcoxon lorsqu'on compare 2 groupes et Kruskal-Wallis lorsqu'on compare plus de 2 groupes) permet de calculer la p-valeur correspondante. Des corrections de la multiplicité de tests est également appliquée.

L'ensemble de ces tests présentés ci-dessus, ainsi que leurs implémentations sont répertoriés dans le Tableau 4.5.

Article	Description	Normalisation	Package R/implementation	Fonction R
	***t*-test** : Test paramétrique de comparaison de moyennes avec possibilité d'inégalité des variances.	Relative	`stats`	`t.test()`
	Log *t*-test : Test paramétrique de comparaison de moyenne avec possibilité d'inégalité des variances sur données log-transformées.	Relative	`stats`	`t.test()`
Segata et al. [2011]	**Lefse** : Test non-paramétrique de comparaison de distribution.	Relative	`http://huttenhower.sph.harvard.edu/galaxy`	
Parks et al. [2014]	**STAMP** : Test non-paramétrique de comparaison de distribution.	Relative	`http://kiwi.cs.dal.ca/Software/STAMP`	
White et al. [2009]	**Metastats** : Ne faisant pas d'hypothèse de distribution, ils utilisent un test de permutation pour test l'égalité des moyennes.	Relative	`http://metastats.cbcb.umd.edu/`	
Bolker et al. [2009]	**Negative binomial generalized linear model (GLM)** : Méthode longtemps utilisée en écologie dans la modélisation des données d'abondance en ajoutant un paramètre pour prendre en compte la sur-dispersion.	-	`MASS`	`glm.nb()`
Love et al. [2014]	**DESeq2** : Estimation bayésienne basée sur un modèle NB.	-	`DESeq2`*	`DESeq2()`
Robinson and Oshlack [2010]	**edgeR** : Même modèle que DESeq2. La différence est dans l'estimation de la variance qui l'estime moins grande.	TMM	`edgeR`*	`exactTest()`
Hardcastle and Kelly [2013]	**baySeq** : Estimation bayésienne basée sur un modèle Beta-binomial.	-	`baySeq`*	`baySeq()`
Paulson et al. [2013]	**metagenomeSeq ZIG** : Estimation par maximum de vraisemblance à l'aide d'un algortithme EM basé sur un modèle Zero Inflated Gamma (ZIG).	CSS	`metagenomeSeq`*	`fitZig()`
Fernandes et al. [2014]	**ALDEx2** : Basé sur une méthode de Monte Carlo échantillonnant une distribution de Dirichlet et moyennant q valeurs sur tous les échantillons. Un test de comparaison de moyenne est ensuite utilisé pour tester la significativité des OTU.	-	`ALDEx2`*	`aldex()`
Mandal et al. [2015]	**ANCOM** : Ne fait aucune hypothèse de distribution.	-	-	-

Tableau 4.5 – Revue des tests univariés de différences d'abondance dans l'analyse de données de microbiote

4.4.4.3 Analyse de covariance

Le microbiote semblant être étroitement lié à notre métabolisme, il est pertinent d'étudier son association avec un grand nombre de facteurs de manière simultanée. L'étude des facteurs génétiques ou environnementaux a pour objectif de mieux comprendre l'étiologie de nombreuses maladies et de développer des mesures thérapeutiques pour adapter la composition du microbiote. Ces facteurs peuvent être de tout type et en très grand nombre : des variables démographiques telles que l'âge et le sexe, des variables liées au traitement ou encore la composition du microbiote par exemple la présence ou non de certaines bactéries observées en culture. Des approches sont proposées pour identifier un sous ensemble de covariables statistiquement associées au microbiome.

Rappelons que $\mathbf{Z}$ représente la matrice des données de comptages de taille $n \times q$ et $\mathbf{X}$ représente la matrice de covariables de taille $n \times p$. Chen and Li [2013b] ont modélisé les données de comptage selon un modèle de DM

$$f_{DM}(\mathbf{Z}_j;\boldsymbol{\gamma}) = \int f_M(\mathbf{Z}_j;\boldsymbol{\pi}) f_D(\boldsymbol{\pi};\boldsymbol{\gamma}) d\boldsymbol{\pi} \quad j = 1, \cdots, q \tag{9}$$

où f_M est la fonction de probabilité d'un modèle de régression logistique multinomial, $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_q)$ sont les probabilités d'occurrence de chaque OTU avec $\sum_{j=1}^{q} \pi_j = 1$, f_D est la fonction de probabilité d'un modèle de Dirichlet et $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_q)$ sont des paramètres positifs. Les covariables sont introduites à travers le modèle log-linéaire suivant :

$$\log \gamma_j(\mathbf{X}_i) = \alpha_j + \sum_{k=1}^{p} \beta_{jk} X_{ik} \tag{10}$$

où α_j peut être interprété comme l'abondance de base pour l'OTU j, β_{jk} est le coefficient de régression de la covariable k par rapport au j-ième OTU et $\boldsymbol{\beta} = (\beta_{jk})_{j,k}$ représente la matrice de l'ensemble des coefficients de taille $q \times (p+1)$. Soit $l(\boldsymbol{\beta}, \mathbf{Z}, \mathbf{X})$ la log-vraisemblance négative correspondant au modèle (9) incorporant (10). Chen and Li [2013b] ont proposé de pénaliser la log-vraisemblance par une pénalité sparse GroupLasso et de minimiser la fonction suivante :

$$\min_{\boldsymbol{\beta}} \left\{ l(\boldsymbol{\beta}, \mathbf{Z}, \mathbf{X}) + \lambda_1 \sum_{k=1}^{p} \sqrt{\sum_{j=1}^{q} \beta_{jk}} + \lambda_2 \sum_{k=1}^{p} \sum_{j=1}^{q} |\beta_{jk}| \right\}$$

avec λ_1 qui contrôle les coefficients associés à la k-ième covariable et λ_2 contrôle la sélection sur les OTU.

Basé sur le même modèle, Wadsworth et al. [2017] ont proposé une approche de sélection en supposant une loi *a priori* sur $\boldsymbol{\beta}$ tel que :

$$\beta_{jk} \sim \delta_{jk} \mathcal{N}(0, r_j^2) + (1 - \delta_{jk}) \delta_0(\beta_{jk})$$

où $\boldsymbol{\delta}_j = (\delta_{j1}, \cdots, \delta_{jp})$ tel que $\delta_{jk} = 1$ si la k-ième covariable a une influence sur l'abondance de l'OTU j et $\delta_{jk} = 0$ sinon, δ_0 est le Delta-Dirac à 0. Un algorithme *Markov Chain Monte Carlo* (MCMC) est ensuite utilisé pour estimer les paramètres.

Toujours dans un cadre bayésien Xia et al. [2013] ont, quant à eux, modélisé l'abondance relative $\tilde{\mathbf{Z}}$ selon un modèle de régression linéaire logistique pondérée sur les données ALR-transformées. Soit $l(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \tilde{\mathbf{Z}}^*, \mathbf{X})$ la log-vraisemblance du modèle logistique, Xia et al. [2013] ont ajouté une pénalité Group Lasso et minimisé la fonction suivante :

$$\min_{\boldsymbol{\beta};,\boldsymbol{\Sigma}} \left\{ l(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \tilde{\mathbf{Z}}^*, \mathbf{X}) + \lambda \sum_{k=1}^{p} \sqrt{\sum_{j=1}^{q-1} \beta_{jk}} \right\}$$

où $\tilde{\mathbf{Z}}^*$ sont les données ALR-transformées de dimension $q-1$. Un algorithme stochastique est utilisé pour estimer les paramètres.

Chen et al. [2013] ont proposé la méthode **ssCCA** (*Structure-constrained Sparse Canonical Correlation Analysis*) qui intègre la structure compositionnelle des données du microbiote et l'information phylogénétique. Le principe de la méthode CCA est de trouver deux directions de projection $\mathbf{u}_1 \in \mathbb{R}^q$ et $\mathbf{v}_1 \in \mathbb{R}^p$ entre les données d'abondance relative $\tilde{\mathbf{Z}}$ et les covariables $\mathbf{X}$ tel que :

$$\max_{\mathbf{u},\mathbf{v}} Corr(\mathbf{u}^\top \tilde{\mathbf{Z}}, \mathbf{v}^\top \mathbf{X}) \tag{11}$$

L'optimisation (11) est équivalente à :

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^\top \tilde{\mathbf{Z}}^\top \mathbf{X}\mathbf{v}$$
$$\text{sous les contraintes } \mathbf{u}^\top \tilde{\mathbf{Z}}^\top \tilde{\mathbf{Z}}\mathbf{u} \leq 1,\ \mathbf{v}^\top \mathbf{X}^\top \mathbf{X}\mathbf{v} \leq 1,\ \|\mathbf{u}\| \leq c_1,\ \|\mathbf{v}\| \leq c_2,\ \mathbf{u}^\top \mathbf{L}\mathbf{u} \leq c_3$$

où $\mathbf{L} = \mathbf{D} - \mathbf{A}$ avec $\mathbf{D}$ une matrice de distance (sous section 4.4.2.2) et $\mathbf{A}$ définie telle que $a_{ij} = 1/d_{ij}^{*2}$ où d_{ij} sont les éléments de $\mathbf{D}$. c_1, c_2 et c_3 sont des paramètres de réglages.

L'ensemble de ces méthodes présentées ci-dessus et leurs implémentations sont répertoriés dans le Tableau 4.6.

Article	Description	Implémentation
Chen and Li [2013b]	Approche par maximum de vraisemblance basée sur un modèle de régression DM	http://statgene.med.upenn.edu/DirMulti.R.txt
Xia et al. [2013]	Approche par maximum de vraisemblance basé une modèle de régression multinomial normal logistique	PenLNM
Chen et al. [2013]	**ssCCA** : version parcimonieuse de l'analyse de corrélation canonique en intégrant la structure compositionelle et l'information phylogénétique	
Wadsworth et al. [2017]	Approche bayésienne basée sur un modèle de régression DM utilisant le *spike-and-slab* comme *a priori*	https://github.com/duncanwadsworth/dmbvs

Tableau 4.6 – Revue des méthodes de sélection de variables impactant le microbiote

4.4.4.4 Méthodes de prédiction à partir des OTU

Dans cette partie, on s'intéresse aux méthodes qui visent à identifier les OTU prédictifs d'une réponse clinique. Le nombre d'OTU (q) étant souvent plus grand que le nombre d'individus étudiés (n), nous nous plaçons dans un contexte de grande dimension. Nous faisons l'hypothèse que seul un sous ensemble d'OTU est associé à la variable d'intérêt. Certains chercheurs ont étendu des méthodes de sélection de variables à la structure compositionelle des données de microbiote.

Comme défini dans la section 4.4.3.1, nous avons $\tilde{\mathbf{Z}}_i \in \mathcal{S}^q$ avec $i = 1, \quad , n$ le nombre d'individus. Ignorer l'appartenance de $\tilde{\mathbf{Z}}$ à une structure simplex engendrerait des problèmes d'identifiabilité lors d'une régression d'une variable réponse $\mathbf{Y}$ sur $\tilde{\mathbf{Z}}$. Une idée naïve est d'exclure une composante de référence de $\tilde{\mathbf{Z}}$. Shankar et al. [2015] ont proposé d'exclure l'OTU avec la plus grande abondance et d'appliquer des méthodes de sélections de variables à partir d'un modèle de régression de $\mathbf{Y}$ sur $\tilde{\mathbf{Z}}^{\backslash \text{ref}}$. $\tilde{\mathbf{Z}}^{\backslash \text{ref}}$ représente la matrice $\tilde{\mathbf{Z}}$ moins l'OTU ayant la plus grande abondance relative considéré comme l'OTU de référence. Le modèle de régression s'écrit comme :

$$\mathbf{Y} = \tilde{\mathbf{Z}}^{\backslash \text{ref}} \boldsymbol{\beta}^{\backslash \text{ref}} + \boldsymbol{\varepsilon} \tag{12}$$

où $\boldsymbol{\beta}^{\backslash \text{ref}}$ est le vecteur de coefficient de régression associé à $\tilde{\mathbf{Z}}^{\backslash \text{ref}}$ et de taille $1 \times (q-1)$. Shankar et al. [2015] a opposé plusieurs méthodes de sélections de variables : des méthodes fréquentistes et des méthodes bayésiennes. L'approche fréquentiste est la

régression pénalisée de type ElasticNet associée au modèle (12) minimisant le critère suivant :

$$\min_{\beta} \left\{ \frac{1}{2n} \|\mathbf{Y} - \tilde{\mathbf{Z}}^{\setminus \text{ref}} \boldsymbol{\beta}^{\setminus \text{ref}}\|_2^2 + \alpha \|\beta_j^{\setminus \text{ref}}\|_2 + (1-\alpha)\|\beta_j^{\setminus \text{ref}}\|_2^2 \right\}$$

Une version visant la stabilité de sélection [Meinshausen and Bühlmann, 2010] a également été étudiée. L'approche bayésienne consiste à utiliser un *a priori spike-and-slab* (soit un a priori consistant en un mélange de probabilités dont une ayant une masse à zéro) pour sélectionner les OTU. Un algorithme MCMC est ensuite utilisé pour estimer les paramètres.

Exclure un OTU de l'analyse amène cependant à des interprétations et des inférences difficiles. Aitchison and Bacon-Shone [1984]; Aitchison [1986] a proposé le populaire modèle linéaire log-constrat pour prendre en compte la compositionnalité des données. L'idée principal est d'effectuer une transformation log-ratio de type alr sur les données de compositions (cf. section 4.4.3.1) telle que :

$$\tilde{\mathbf{Z}}^{\setminus q*} = \text{alr}(\tilde{\mathbf{Z}}_i) = \left(\log \frac{\tilde{Z}_{i1}}{\tilde{Z}_{iq}}, \cdots, \log \frac{\tilde{Z}_{iq-1}}{\tilde{Z}_{iq}} \right) \in \mathbb{R}^{q-1}$$

où q est l'OTU de référence. Le modèle linéaire log-contrast s'écrit comme :

$$\mathbf{Y} = \tilde{\mathbf{Z}}^{\setminus q*} \boldsymbol{\beta}^{\setminus q*} + \boldsymbol{\varepsilon} \tag{13}$$

avec $\boldsymbol{\beta}^{\setminus q*}$ le vecteur de coefficient de régression associé à $\tilde{\mathbf{Z}}^{\setminus q*}$ de taille $1 \times (q-1)$. Bien que le modèle (13) dépend d'un OTU de référence, il a cependant une forme symétrique. Soit $\tilde{\mathbf{Z}}^* = \log(\tilde{\mathbf{Z}}) \in \mathbb{R}^{(n \times q)}$, le modèle (13) est équivalent à :

$$\mathbf{Y} = \tilde{\mathbf{Z}}^* \boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \text{sous} \qquad \sum_{j=1}^{q} \beta_j = 0 \tag{14}$$

avec $\boldsymbol{\beta}$ le vecteur de paramètres de régression associé à $\tilde{\mathbf{Z}}^*$ de taille $1 \times q$.

Astuce :

$$\begin{aligned}
\mathbf{Y} &= \beta_1 \tilde{\mathbf{Z}}_1^* + \cdots + \beta_{q-1} \tilde{\mathbf{Z}}_{q-1}^* + \boldsymbol{\varepsilon} \\
&= \beta_1 \left(\log \tilde{\mathbf{Z}}_1 - \log \tilde{\mathbf{Z}}_q \right) + \cdots + \beta_q \left(\log \tilde{\mathbf{Z}}_{q-1} - \log \tilde{\mathbf{Z}}_q \right) + \boldsymbol{\varepsilon} \\
&= \sum_{j=1}^{q-1} \beta_j \log \tilde{\mathbf{Z}}_j - \sum_{j=1}^{q-1} \beta_j \log \tilde{\mathbf{Z}}_q + \boldsymbol{\varepsilon} \\
&= \sum_{j=1}^{q} \beta_j \log \tilde{\mathbf{Z}}_j + \varepsilon
\end{aligned}$$

avec $\beta_q = \sum_{j=1}^{q-1} \beta_j$

Lin et al. [2014] a proposé d'étendre le modèle (14) au cadre de la grande dimension en appliquant une pénalité Lasso :

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\mathbf{Y} - \tilde{\mathbf{Z}}^* \boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{q} |\beta_j| \right\} \qquad \text{sous} \qquad \sum_{j=1}^{q} \beta_j = 0$$

Chen et al. [2013] et Shankar et al. [2015] ont pris en compte seulement l'aspect compositionnel des données mais ne font pas intervenir la structure phylogénétique sous jacente. Garcia et al. [2014] ont été les premiers à introduire une structure hiérarchique dans la sélection d'OTU. Ils ont étendu le sparse Group Lasso et ont proposé la méthode **SGSL** (*sparse Group-subgroup Lasso*). La Figure 4.8 décrit le schéma du modèle. On suppose que les OTU appartiennent à des sous-groupes qui eux mêmes appartiennent à des groupes. Le modèle SGSL prend en compte deux niveaux hiérarchiques qui sont donc représentés par deux rangs taxonomiques imbriqués. Par exemple, le groupe correspond au rang Phylum, le sous-groupe au rang Famille et les OTU étudiés sont au rangs Genre. D'un point de vue du modèle, on considère qu'il existe L groupes disjoints et que dans chaque groupe, il y a M_l sous-groupes disjoints avec $l = 1, \cdots L$. Soit q_l le nombre d'OTU dans le groupe l et $q_{l,m}$ dans le sous-groupes m avec $m = 1, \cdots, M_l$ (Figure 4.8). On peut écrire que $\tilde{\mathbf{Z}}^{(l,m)} \subset \tilde{\mathbf{Z}}^{(l)} \subset \tilde{\mathbf{Z}}$ avec $\tilde{\mathbf{Z}}^{(l,m)}$ et $\tilde{\mathbf{Z}}^{(l)}$ les matrices d'abondance relative de taille respective $n \times p_{l,m}$ et $n \times p_l$ correspondant au rang taxonomique choisi. On note $\boldsymbol{\beta}^{(l)}$ et $\boldsymbol{\beta}^{(l,m)}$ les vecteurs de coefficients associés au groupe l et au sous-groupe m du groupe l, respectivement. Garcia et al. [2014] ont donc proposé de minimiser le critère suivant :

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\mathbf{Y} - \sum_{l=1}^{L} \tilde{\mathbf{Z}}^{(l)} \boldsymbol{\beta}^{(l)}\|_2^2 \quad + \quad \alpha_1 \lambda \sum_{l=1}^{L} \sqrt{\sum_{m=1}^{M_k} \beta^{(l,m)}} + \alpha_2 \lambda \sum_{l=1}^{L} \sum_{m=1}^{M_k} \sqrt{\sum_{j=1}^{p_{l,m}} \beta_j^{(l,m)}} \right.$$
$$\left. + (1 - \alpha_1 - \alpha_2) \lambda \sum_{k=1}^{q} |\beta_j| \right\}$$

avec α_1, α_2 et λ qui contrôlent la sparsité sur les groupes, les sous-groupes et l'ensemble des OTU de manière respective. Si $\alpha_1 = 1$, la seule contrainte porte sur les groupes. Si $\alpha_2 = 1$, la seule contrainte porte sur les sous-groupes. Lorsque $\alpha_1 = \alpha_2 = 0$, seuls les OTU sont contraints de manière individuels comme un critère Lasso classique. Pour des valeurs de α_1 et α_2 strictement comprise dans $]0, 1[$, des poids sont pris en compte sur l'ensemble des contraintes. Avec ces contraintes, ce n'est pas le group Lasso qui est adapté au contexte des relations phylogénétiques de grande dimension du microbiome, mais le overlapping Group Lasso [Yuan et al., 2011].

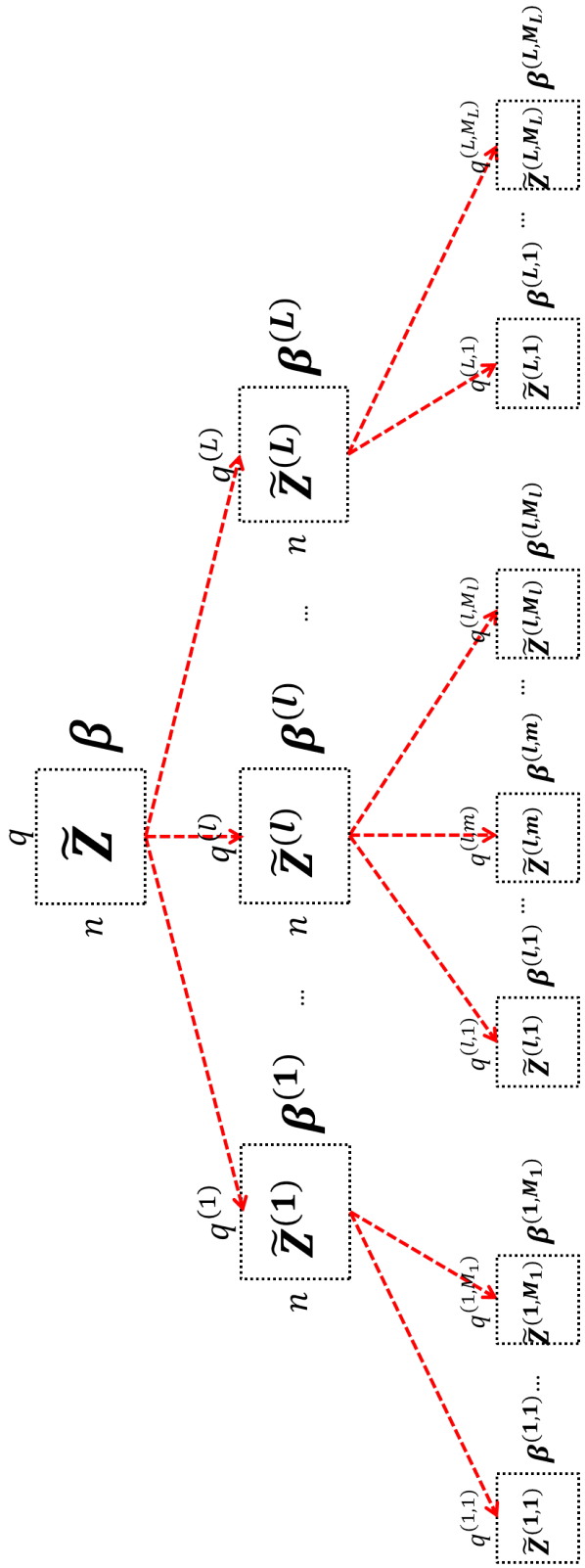

FIGURE 4.8 – Schéma de description des paramètres de la méthode SGSL.

SGSL prend en compte seulement deux rangs taxonomiques. La méthode **Phy-Lasso** proposée par Rush et al. [2016] permet d'intégrer l'ensemble de la lignée taxonomique de chacun des OTU. Cette approche considère l'ensemble des liens de parenté existants entre les OTU, du rang le plus bas au rang Royaume. Rush et al. [2016] se sont basés sur le modèle hiérarchique défini par Zhou and Zhu [2010] qui proposent de reparamétriser $\boldsymbol{\beta}^{(l)} = d_l \tilde{\boldsymbol{\beta}}^{(l)}$ avec d_l un paramètre agissant sur le groupe l. Le modèle hiérarchique s'écrit de la forme :

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\mathbf{Y} - \sum_{l=1}^{L} d_l \tilde{\mathbf{Z}}^{(l)} \tilde{\boldsymbol{\beta}}^{(l)}\|_2^2 - \sum_{l=1}^{L} d_l - \lambda \sum_{k=1}^{q} |\tilde{\beta}_j| \right\} \tag{15}$$

avec $d_l \geq 0$ qui contrôle la sélection du groupe l. Si $d_l = 0$, aucun OTU du groupe l n'est sélectionné. λ permet la régularisation dans la sélection des OTU à l'intérieur d'un groupe. Le modèle hiérarchique est en réalité un sparse GroupLasso paramétré différemment. Pour prendre en compte la structure phylogénétique entière des données, ils ont considéré un modèle hiérarchique à chaque rang taxonomique t avec $t = 1, \cdots, T$. $t = 1$ correspond au plus haut rang taxonomique (le Royaume) et T le plus bas (Espèce ou Genre). La Figure 4.9 décrit le schéma des différents paramètres. Le modèle Phy-Lasso reprend le modèle (15) en sommant sur l'ensemble des niveaux taxonomiques.

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\mathbf{Y} - \sum_{t=1}^{T} \sum_{l=1}^{L} d_l^t \tilde{\mathbf{Z}}^{(l,t)} \tilde{\boldsymbol{\beta}}^{(l,t)}\|_2^2 - \sum_{t=1}^{T} \sum_{l=1}^{L} d_l^t - \lambda \sum_{k=1}^{q} |\tilde{\beta}_j| \right\}$$

avec d_l^t qui contrôle la sélection du groupe l au niveau taxonomique t et λ permet la régularisation sur l'ensemble des OTU.

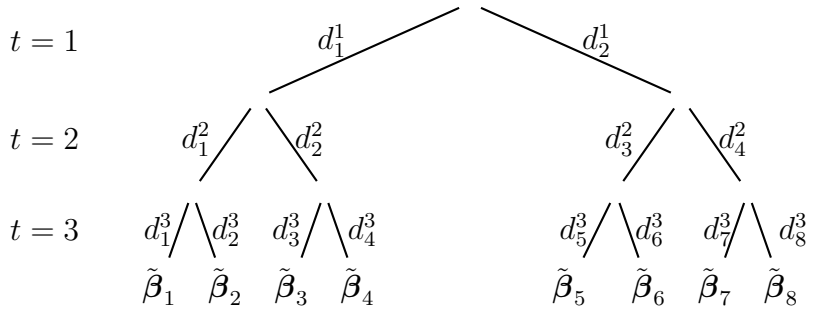

FIGURE 4.9 – **Paramétrisation de Phy-Lasso.** Exemple pour un arbre à $T = 3$ niveaux taxonomiques et $q = 8$ OTU.

Chen et al. [2015] a proposé **glmgraph** intégrant une structure sous-jacente de type réseau pour prendre en compte les liens phylogénétiques des données. Le problème optimisé est le suivant :

$$\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{Y} - \tilde{\mathbf{Z}}\boldsymbol{\beta}\|_2^2 - \lambda \sum_{k=1}^{q} |\tilde{\beta}_j| + \lambda_2 \boldsymbol{\beta}^\top \mathbf{L} \boldsymbol{\beta} \right\}$$

avec $\mathbf{L}$ une matrice de Laplace définie à partir du réseau établi, $\boldsymbol{\beta}^\top \mathbf{L} \boldsymbol{\beta} = \sum a_{jk} \left(\beta_j - \beta_k\right)^2$ et a_{jk} les éléments de la matrice $\mathbf{A}$ est définie dans la méthode ssCCA (sous section 4.4.4.3).

SMART-scan [Zhang et al., 2015] propose de réduire le nombre de paramètres β_j en regroupant des OTU entre eux selon leur position dans la taxonomie. Deux OTU provenant d'une même famille sont regroupés en sommant leurs abondances. Un grand nombre de combinaisons d'OTU étant possible, Zhang et al. [2015] ont proposé un nouveau critère AIC pour sélectionner le modèle.

La méthode de moindres carrés partiels (PLS) pour l'analyse discriminante a également été appliquée à des données de microbiote [Le Cao et al., 2016] mais ne sera pas détaillée dans cette thèse.

L'ensemble de ces méthodes présentées ci-dessus et leurs implémentations sont répertoriés dans le Tableau 4.7.

Article	Description	Implémentation
Lin et al. [2014]	Extension du modèle log-contrast à une pénalité Lasso	http://www.math.pku.edu.cn/teachers/linw/software.html
Shankar et al. [2015]	Evaluation systématique de méthode fréquentiste (régression Elastic-Net) et de méthode bayésienne (*a priori spike-and-slab*	http://github.com/openpencil/regeval
Garcia et al. [2014]	**SGSL** : Intègre deux niveaux taxonomiques au modèle de régression linéaire et applique une pénalité Lasso	http://www.stat.tamu.edu/~tpgarcia/publications.html
Chen et al. [2015]	**glmgraph** : Incorpore la structure phylogénétique dans une matrice de Laplace et applique une pénalité Lasso	glmgraph
Rush et al. [2016]	**Phy-Lasso** : Incorpore la structure phylogénétique dans un modèle hiérarchique et applique une pénalité Lasso	
Zhang et al. [2015]	**SMART-scan** : méthode de sélection de modèles basée sur le critère AIC pour sélectionner des groupes d'OTU.	https://dsgweb.wustl.edu/qunyuan/software/smartscan/
Le Cao et al. [2016]	**sPLS-DA** : Méthode de moindres partiels (PLS) pour l'analyse discriminante appliquée à des données du microbiote	http://www.mixomics.org/mixMC

Tableau 4.7 – Revue des méthodes de sélection d'OTU dans l'analyse des données de microbiote

5 Analyse des données du microbiote respiratoire chez des patients mucoviscidosiques (cohorte MucoFong)

Dans ce chapitre nous détaillons le protocole d'analyse statistique utilisé dans l'étude MucoFong. Une étude préalable de simulation a aidé à la prise de décision : parmi le large choix de méthodes récentes, pour lesquelles peu de comparaisons ont été faites dans la littérature, lesquelles utiliser dans l'analyse de ces données? En effet, parallèlement au développement des nouvelles techniques de séquençage du microbiome humain, on assiste à une émergence de méthodes statistiques spécifiques et d'outils informatiques. En raison de la nouveauté de ces approches, il est encore tôt pour évaluer l'applicabilité et la précision des méthodes disponibles décrites précédemment. Les simulations miment ces données, nous commençons ainsi par introduire les caractéristiques des données de l'étude du microbiote pulmonaire issues d'une étude cas-témoins nichée à la cohorte MucoFung.

Valorisation: L'étude de simulation a fait l'objet d'une communication poster à l'*International Biometric Conference* de 2018. Un article portant sur l'analyse des données et l'interprétation des résultats a été accepté pour publication dans le journal *Scientific Reports* en Octobre 2019.

5.1 La cohorte MucoFong

La mucoviscidose est une maladie génétique qui touche plus de 35 000 personnes en Europe [Society, 2014; Touati et al., 2014]. En France, on compte près de 6400 personnes souffrant de cette maladie orpheline [Society, 2014]. Il s'agit d'une maladie des glandes à sécrétion externe où les principaux organes touchés sont l'appareil respiratoire, le tube digestif et ses glandes annexes (pancréas, foie, voies biliaires). Cette maladie est due à la mutation du canal chlore CFTR (*Cystic Fibrosis Transmembrane conductance Regulator*). Cette mutation conduit à la sécrétion par les poumons d'un mucus visqueux et abondant propice aux infections bactériennes. Les plus fréquentes sont les infections

broncho-pulmonaires causées par des germes tels que *Haemophilus influenzae*, *Staphylococcus aureus* ou *Pseudomonas aeruginosa*. Les répétitions de ces poussées infectieuses et l'inflammation chronique conduisent à une insuffisance respiratoire sur le long terme. Elles constituent 9 fois sur 10 la cause de décès dans la mucoviscidose. Le traitement des infections se fait par l'administration d'antibiotiques, ce qui à terme engendre une multi-résistante des bactéries. Diverses espèces fongiques comme notamment *Aspergillus fumigatus* et *Scedosporium apiospermum* peuvent également coloniser les voies respiratoires et être responsables d'allergies broncho-pulmonaires.

De 2007 à 2014, un PHRC national multicentrique (PHRC national 1902006) a été conduit pour déterminer les micromycètes (des champignons eucaryiotes microscopiques) présents dans l'appareil respiratoire des patients atteints de mucoviscidose responsables d'une colonisation ou d'épisodes infectieux. L'objectif principal du PHRC était d'établir la prévalence de ces micromycètes dans le contexte de la mucoviscidose. L'ensemble des données mycologiques du projet ont permis de proposer pour la première fois des recommandations nationales pour la prise en charge mycologique des expectorations (Travaux du groupe MucoMicrobes coordonné par le Pr Plésiat, publication dans le Référentiel en microbiologie Médicale (REMIC) de la Société Française de Microbiologie version 5 de 2015).

De plus, une base de données informatisée exhaustive inclue une sous-population pour laquelle des données de métagénomique ciblée déterminant la flore procaryote et eucaryotique (microbiote et mycobiote) du tractus respiratoire sont disponibles. En effet, pour chacun des 300 patients inclus et suivis pendant 2 ans (3 bilans par patient), les données épidémiologiques, cliniques, radiologiques et biologiques ont été collectées, lors de chacun des 3 bilans selon les modalités préconisées par le registre national (`http://www.vaincrelamuco.org/face-la-mucoviscidose/registre-et-muco-en-chiffres`).

L'exhaustivité des données et leur caractère prospectif avec un suivi longitudinal des patients couplés à la caractérisation des microbiotes et mycobiotes respiratoires par séquençage haut-débit vise d'établir si une relation causale existe entre une dysbiose du tractus respiratoire, avec en particulier la présente d'*Aspergillus fumigatus* ou de *Candida albicans*, et la détérioration clinique et/ou respiratoire des patients.

L'objectif principal est donc de préciser le rôle des micromycètes dans la dégradation de la fonction pulmonaire des patients atteints de mucoviscidose en caractérisant les liens entre microbiotes et mycobiotes respiratoires, dysbiose et fonction respiratoire des patients. La capacité respiratoire est décrite par une variable quantitative appelée Volume Expiratoire Maximal par Seconde (VEMS). C'est le volume de gaz rejeté pendant la première seconde d'une expiration forcée et il est exprimé en pourcentages de valeurs prédites. De plus, le patient peut présenter un état stationnaire pendant de longues périodes puis tout

à coup développer une insuffisance respiratoire aiguë. On parle d'"exacerbation" lorsque l'ensemble des symptômes s'intensifie. Ce critère clinique est établi par le médecin lors d'une visite et peut être subjectif quant à l'état du patient.

La cohorte MucoFong regroupe 256 patients avec des dossiers complets (Figure 5.1 à gauche). Trois bilans cliniques ont été effectués dans le temps.

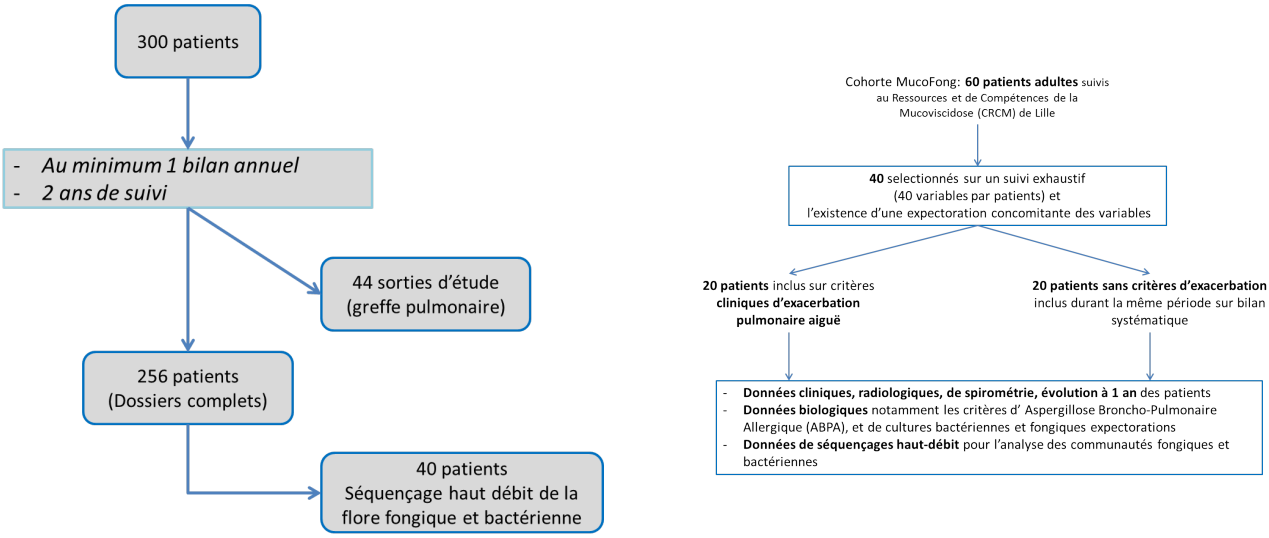

Figure 5.1 – Diagramme de flux de la cohorte MucoFong (gauche) et patients adultes sélectionnées pour comparer le microbiote et le mycobiote en présence ou non d'une exacerbation pulmonaire aiguë (droite).

Sur ces 256 patients, 40 ont été choisis et séquencés pour comparer le microbiote en présence ou non d'une exacerbation (Figure 5.1 à droite). Ces patients ont été sélectionnés afin d'avoir le même nombre de patients avec un statut exacerbé et un statut non-exacerbé.

L'ADN a été extrait avec le kit High Pure PCR Template Preparation kit (Roche Applied Science, Germany) selon les recommandations du fabricant avec allongement du temps de digestion par protéinase K à 1 h à 70C au lieu de 10 min. La zone utilisée pour cette métagénomique ciblée est V3-V4 de l'ADNr 16S et ITS2 pour les micromycètes. Les données brutes ont été analysées selon le protocole classiquement employé. Toutes les données (contrôle qualité du sequençage, OTU, analyse taxonomique et calcul de mesure de diversité) ont été effectuées avec le logiciel QIIME [Caporaso et al., 2010b]. Les données ont été assignées au rang Genre pour les bactéries et au rang Espèce pour les champignons.

En bref, les séquences avec des ambiguïtés de codes ont été supprimés. Les filtres 16rDNa et ITS2 ont ensuite été alignés respectivement sur la base de données GreenGene (V13.8 contenant 203 452 ADNr 16S) et la classification des gènes ITS2 à partir de la base de données UNITE (V 12.11 contenant 97 868 séquences ITS1-ITS2). Après avoir regroupé le prokaryote 16S rDNa et les pirosequences ITS2 de champignon avec 97 %

de similitude à l'aide d'Uclust (http://www.drive5.com/usearch/). Les courbes de Rarefaction ont ensuite été calculées.

Sur les 40 patients, trois échantillons n'ont pas été séquencées pour mauvaise condition de stockage. De plus, les échantillons 2, 7, 28 et 35 ont été écartés pour leur courbes de rarefaction correspondantes et/ou leurs nombre inadéquates de pyroséquences pour l'analyse de diversité. Les patients 12 et 14 ont également été exclus pour des erreurs de séquençage. Au final 31 patients ont été considérés dans l'analyse. Suite à la classification en OTU, 93 genres de bactéries et 132 espèces de champignons ont été relevés.

5.2 Protocole d'analyse de la cohorte MucoFong

La revue de la littérature des méthodes statistiques (4.4) nous a permis d'élaborer un plan d'analyse statistique cohérent pour répondre aux objectifs de l'étude. Différents critères d'intérêt ont été explorés : la présence ou non d'une exacerbation (CFPE), le VEMS, l'indice de masse corporelle (BMI) et le score de SHWACHMAN (Score-SK) qui représente la mesure du degré de sévérité de la mucoviscidose comprise entre 0 et 100 (score de 40 et moins interprété comme état sévère). L'article ci-après fait état des résultats seulement sur le CFPE et le VEMS.
Le protocole d'analyse statistique est décrit ci-dessous:

1. Analyse descriptive de la population et comparaison des populations par un test de Student (ou Wilcoxon) dans le cas de variables quantitatives et un test de Khi-2 (ou Fisher Exact) dans le cas de variables qualitatives.

2. Calcul des indices de diversité α (Shannon, Simpson, Chao1) et comparaison de la diversité entre les populations à l'aide d'un test de Student (ou test de rang de Wilcoxon).

3. Analyse de la diversité β à travers une représentation graphique reposant sur une PCoA et du test statistique ANOSIM.

4. Analyse d'abondance différentielle entre les populations.
   — Choix d'une distribution approprié à l'abondance de l'OTU : Zero-inflated Gaussian, Zero-inflated negative binomiale, Zero-Inflated Poisson ou aucun d'entre eux.
   — Choix du test statistique en fonction de la distribution choisie: DESeq et EdgeR pour les distributions discrètes, MetagenomeSeq pour une distribution gaussienne Zero-inflated et un test de Wilcoxon lorsqu'aucun des trois modèles n'est adapté.

5. Calcul des corrélations entre les bactéries et les champignons et construction d'un réseau. Les corrélations ont été calculées à partir de la méthode de ReBoot.
6. Sélection de la meilleur méthode multivariée pour de la sélection de variables à travers une étude de simulation.
7. Extension de Phy-Lasso à la régression linéaire (uniquement, l'implémentation de la régression logistique était disponible publiquement). Utilisation de la méthode sélectionnée Phy-Lasso pour l'analyse des données de MucoFong.

5.3 Etude de simulation pour l'analyse des données issues de la cohorte MucoFong

L'objectif de ces simulations est d'évaluer les capacités des différentes méthodes prédictives (décrites dans la sous section 4.4.4.4) afin d'utiliser une approche appropriée pour l'analyse des données issues de la cohorte MucoFong.

5.3.1 Protocole de simulation

Le protocole de simulation a été conduit suivant les données issues de la cohorte MucoFong.

Les abondances relatives ont été simulées à partir d'une distribution Dirichlet multinomiale (DM) telle que $\tilde{\mathbf{Z}} \sim \text{DM}(\boldsymbol{\pi}) \in \mathcal{S}_{n\times p}$. Le vecteur de paramètres $\boldsymbol{\pi}$ est choisi comme les moyennes des fréquences observées des OTUs calculées sur les abondances relatives des différentes bactéries relevées dans la cohorte MucoFong. Le nombre d'OTUs q est fixé à 75 (conforme au jeu de données réelles). Le nombre d'individus varie entre $n = 30$ (conforme au jeu de données réelles) et $n = 100$.

Le VEMS, $\mathbf{Y}$ dans nos simulations, est supposé gaussien. Un modèle de régression linéaire est utilisé pour simuler cette variable:

$$\mathbf{Y} = \beta_0 + \text{ALR}\left(\tilde{\mathbf{Z}}\right)\boldsymbol{\beta}^V + \boldsymbol{\varepsilon} \tag{16}$$

où β_0 représente la moyenne du VEMS. La valeur estimée de β_0 sur le jeu de données réelles est de 54%. $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2)$ et σ^2 est choisi afin que le ratio signal-bruit soit égal à 3. Supposons que tous les OTUs ne sont pas associés à $\mathbf{Y}$. Pour cela, certaines valeurs du vecteur $\boldsymbol{\beta}^V$ doivent être fixées à 0. Deux protocoles, illustrés par la Figure 5.2, ont été étudiés. Le protocole S1 (Figure 5.2 à gauche) prend en compte l'hypothèse suivante: les OTUs phylogénétiquement proches ont un effet similaire. Deux groupes d'OTUs issus de deux Classes taxonomiques différentes sont choisis pour être associés à $\mathbf{Y}$. Les paramètres

Analyse	**Exposure**	**Outcomes**		**Packages**
		CFPE clinical status (binary)	**VEMS, BMI or SK-Score (quantitative)**	
Network analyses	Interactions between bacterial and fungal community: Permutation-renormalization bootstrap (ReBoot) method			`ccrepe`
Bivariate analyses	Patient's characteristics at inclusion	Binary outcome × Quantitative characteristics: T-Student or Wilcoxon-signed-rank test (if T-Student conditions not fulfilled). Binary outcome × Qualitative characteristics: Khi-2 test or Fisher exact test (if Khi-2 test conditions not fulfilled).		`t.test`,`wilcox.test` and `chisq.test`
	Alpha diversity: Shannon, Simpson, Chao1	Binary outcome × Quantitative index: T-Student or Wilcoxon-signed-rank test (if T-Student conditions not fulfilled)	Quantitative outcome × Quantitative index: test on Pearson correlation coefficient or Spearman's rank correlation coefficient (if Pearson correlation conditions not fulfilled).	`vegan` and `fossil`
	Beta diversity: PCoA with Bray-Curtis similarities	Non-parametric Analysis Of SIMilarities (ANOSIM) test		`vegan` and `ape`
	Analysis in whole population: OTU-by-OTU statistical comparisons (genus level). Analysis in majority and minority population: OTU-by-OTU statistical comparisons (genus level)	Model Selection by Testing for the most appropriate distribution: Zero-inflated Gaussian, Zero-inflated negative binomial, Zero-inflated Poisson or non-parametric. Statistical comparisons are then performed using the most appropriate test: DESeq and EdgeR for discrete distribution, MetagenomeSeq for Zero-inflated Gaussian distribution, Wilcoxon-signed rank-test (in all other cases).		`wilcox.test`
Multivariate analyses	Selection of the most appropriate multivariate method among methods proposed for microbiome data analysis performed by a simulation study (*Soret et al. A simulation framework of high-dimensional phylogenetic microbiota data. In: 29th International Biometric Conference, Jul 2018, Barcelone, Spain*). Selected method: Phy-Lasso	Phy-Lasso Logistic regression: LOO-CV to tune the hyper parameters, Stability of selection of OTU: boostrap procedure.	Phy-Lasso Linear regression: LOO-CV to tune the hyper parameters, Stability of selection of OTU: boostrap procedure.	`PhyLasso`

Table 5.1 – Descriptif des analyses statistiques réalisées et les packages R utilisés.

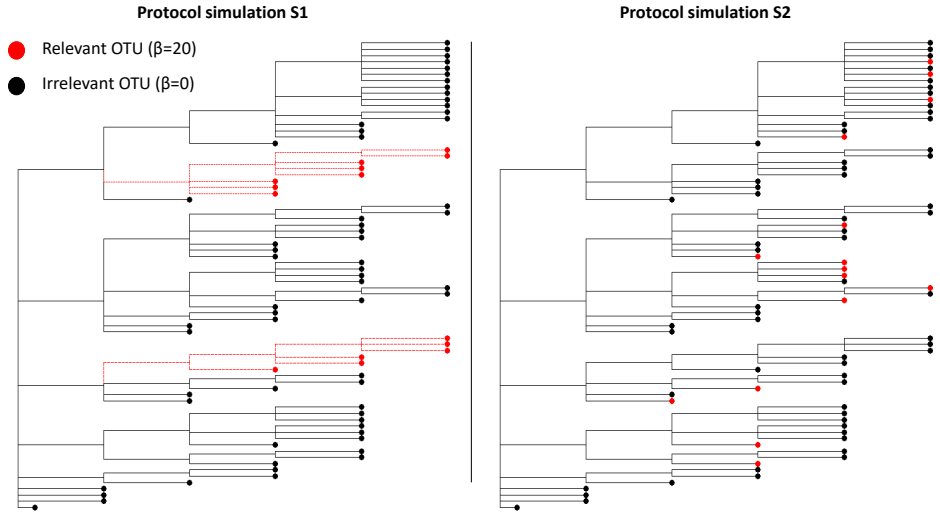

Figure 5.2 – Illustration des protocoles de simulations

$\boldsymbol{\beta}_j^V$ de ces OTUs sont fixés à 20. L'ensemble des autres paramètres $\boldsymbol{\beta}_j^V$ sont fixés à 0.

$$\boldsymbol{\beta}^V = (\underbrace{20, 20, 20, 20, 20}_{\text{rang Classe}}, 0, 0, \cdots, 0, 0, \underbrace{20, 20, 20, 20, 20}_{\text{rang Classe}})$$

Le protocole S2 (Figure 5.2 à droite) consiste à choisir un ensemble d'OTUs pertinents aléatoirement. Les valeurs des paramètres sont les mêmes que dans le protocole S1.

$$\boldsymbol{\beta}^V = (20, 0, 0, 0, 20, 0, \cdots, 0, 20, 0, 0, \cdots, 0, 20, 0, 0, 0)$$

5.3.1.1 Méthodes comparées pour l'étude de simulation

Le tableau 5.2 récapitule les méthodes comparées en précisant la forme de la pénalité, si celle-ci tient compte de la phylogénie, ainsi que la disponibilité du code R (package sur CRAN ou fonctions disponibles sur le site de l'auteur ou de la revue de publication de la méthode).

Méthodes	**Pénalité**	**Phylogénie**	**code R**
Lasso	$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{Y} - \tilde{\mathbf{Z}}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|$	Non	`glmnet`
Elastic-Net	$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{Y} - \tilde{\mathbf{Z}}\boldsymbol{\beta}\|_2^2 + \lambda\left(\alpha\|\boldsymbol{\beta}\| + (1-\alpha)\|\boldsymbol{\beta}\|_2^2\right)$	Non	`glmnet`
Bayesian	*a priori Spike-and-slab*	Non	Disponible
Log-Contrast	$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{Y} - \tilde{\mathbf{Z}}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\| \quad \sum_{j=1}^{p}\beta_j = 0$	Non	Disponible
SGSL	$\tilde{\mathbf{Z}} = \left(\tilde{\mathbf{Z}}^{(1)}, \cdots, \tilde{\mathbf{Z}}^{(k)}, \cdots, \tilde{\mathbf{Z}}^{(L)}\right)$ with $\tilde{\mathbf{Z}}^{(k)} \in \mathcal{M}_{(n\times p_k)}$ $\tilde{\mathbf{Z}}^{(k)} = \left(\tilde{\mathbf{Z}}^{(k,1)}, \cdots, \tilde{\mathbf{Z}}^{(k,m)}, \cdots, \tilde{\mathbf{Z}}^{(k,M_k)}\right)$ with $\tilde{\mathbf{Z}}^{(k,m)} \in \mathcal{M}_{(n\times p_{k,m})}$ $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left\{\frac{1}{n}\|\mathbf{Y} - \sum_{l=1}^{L}\tilde{\mathbf{Z}}^{(l)}\boldsymbol{\beta}^{(l)}\|_2^2 + \alpha_1\lambda\sum_{l=1}^{L}\sqrt{\sum_{m=1}^{M_k}\beta^{(l,m)}}\right.$ $\left. + \alpha_2\lambda\sum_{l=1}^{L}\sum_{m=1}^{M_k}\sqrt{\sum_{j=1}^{p_{l,m}}\beta_j^{(l,m)}} + (1-\alpha_1-\alpha_2)\lambda\sum_{k=1}^{q}\|\beta_j\|\right\}$	3 niveaux	Disponible
Phy-Lasso	$\tilde{\mathbf{Z}} = \left(\tilde{\mathbf{Z}}^{(1)}, \cdots, \tilde{\mathbf{Z}}^{(k)}, \cdots, \tilde{\mathbf{Z}}^{(L)}\right)$ with $\tilde{\mathbf{Z}}^{(k)} \in \mathcal{M}_{(n\times p_k)}$ $t = 1, \cdots, T+1$ taxonomic level $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left\{\frac{1}{n}\|\mathbf{Y} - \sum_{t=1}^{T}\sum_{l=1}^{L} d_l^t\tilde{\mathbf{Z}}^{(l,t)}\tilde{\boldsymbol{\beta}}^{(l,t)}\|_2^2 - \sum_{t=1}^{T}\sum_{l=1}^{L} d_l^t - \lambda\sum_{k=1}^{q}\|\tilde{\beta}_j\|\right\}$	Lignée taxonomique	Disponible
StructFDR	Approche tests multiples par modèle hiérarchique H_{0j}: le $j$ème OTU n'est pas associé à $\mathbf{Y}$ Algorithme de permutation qui contrôle le FDR	Lignée taxonomique	`StructFDR`

Table 5.2 – Liste des méthodes comparées dans les simulations.

5.3.1.2 Sélection des paramètres de régularisation et critère de comparaison

Le jeu de données $\mathbf{D}_k$ est aléatoirement coupé en 10 blocs disjoints de taille approximativement égale. $\mathbf{D}$ est considéré comme un jeu de données d'apprentissage, utilisé pour estimer les coefficients de régression $\boldsymbol{\beta}$. $\mathbf{D}_{\setminus k}$ représente le jeu de données test, qui n'est pas utilisé dans le processus d'estimation et qui est ensuite utilisé pour évaluer la fonction de perte. Ici, on considère la moyenne des erreurs de prédiction comme fonction de perte. La validation croisée peut donc s'écrire comme:

$$CV\left(\lambda\right) = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{n_k}\sum_{i\in\mathbf{D}_k}\left(Y_i - \text{CLR}\left(\tilde{\mathbf{Z}}_i\right)\hat{\boldsymbol{\beta}}(\lambda)_{\mathbf{D}_{\setminus k}}\right)^2 \tag{17}$$

avec n_k la taille de l'échantillon de $\mathbf{D}_k$.

L'objectif principal de l'étude de simulation étant la capacité de sélection des différentes méthodes, il est important de mettre en place une procédure de stabilité de sélection sur chacune des méthodes. Pour cela, nous avons repris le principe du BoLasso décrit dans la section 2.8. $B = 100$ échantillons bootstraps sont créés pour chacune des $S = 100$ simulations générées par le modèle (16). On note $\hat{\boldsymbol{\beta}}^{(s,b)}$ le vecteur de paramètres estimés sur l'échantillon bootstrap b du jeu de données simulé s. Une proportion de sélection notée SP est calculée pour chacun des OTUs et pour un jeu de simulation s :

$$SP\left(\hat{\beta}_j^{(s)}\right) = \frac{1}{B}\sum_{b=1}^{B}\mathbb{1}_{\hat{\beta}_j^{(s,b)}\neq 0} \quad \in \quad [0,1]$$

Une proportion de sélection élevée (supérieure à 70%, seuil choisi pour l'ensemble des méthodes en évaluant visuellement quelques résultats des simulations et les données réelles) pour un OTU signifie qu'il existe une association entre cet OTU et la variable $\mathbf{Y}$. Pour chacune des simulations générées, nous pouvons identifier quels OTUs ont été considérés comme pertinents. On note $\boldsymbol{\beta}_{SP}^{(s)}$, le vecteur binaire de taille $q \times 1$: 1 si l'OTU est identifié comme pertinent et 0 sinon. Cette approche est effectuée pour l'ensemble des méthodes comparées.

Afin d'évaluer la capacité de sélection de chacune des méthodes, nous comparons $\boldsymbol{\beta}_{SP}^{(s)}$ calculé par la stabilité de sélection avec le "vrai" $\boldsymbol{\beta}^V$ fixé pour les simulations. Le but est de déterminer si les méthodes identifient les OTUs réellement associés à la réponse. Pour cela, nous calculons le taux de faux positifs et le taux de faux négatifs pour chacune des méthodes. Un faux positif pour un OTU j est relevé lorsque $\boldsymbol{\beta}_{SP,j}^{(s)} = 1$ alors que $\boldsymbol{\beta}_j^V = 0$. Le taux de faux positifs est calculé comme la somme d'OTU faux positifs sur le nombre total d'OTUs non associés à la réponse. Un faux négatif pour un OTU j est relevé lorsque $\boldsymbol{\beta}_{SP,j}^{(s)} = 0$ alors que $\boldsymbol{\beta}_j^V \neq 0$. Le taux de faux négatifs est calculé comme la somme d'OTU faux négatifs sur le nombre total d'OTUs réellement associés à la réponse.

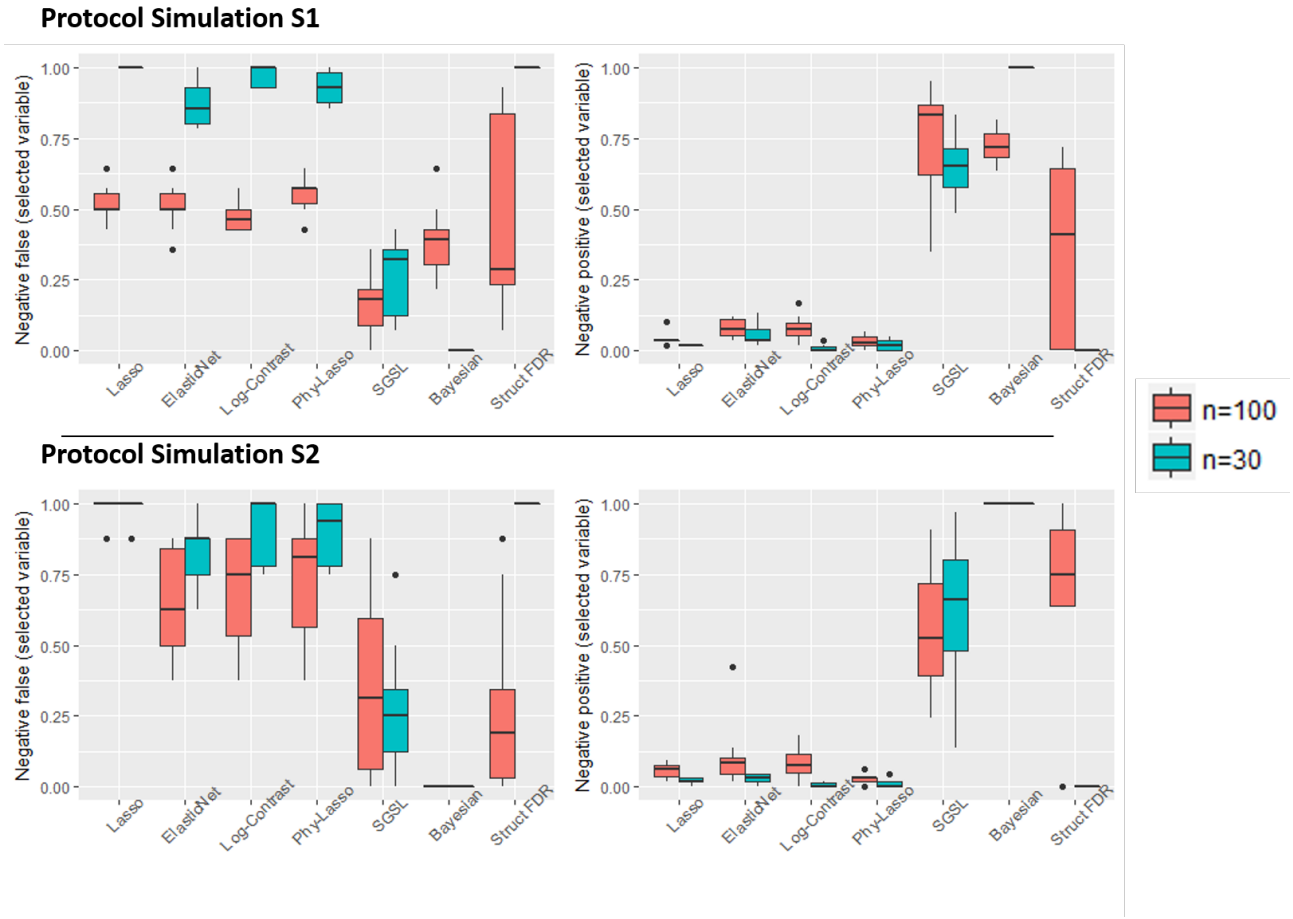

Figure 5.3 – Taux de faux positifs et taux de faux négatifs évaluées sur 100 simulations.

5.3.1.3 Résultats et discussion

Sur la Figure 5.2, les deux graphiques du haut représentent les résultats sur le protocole S1 et les deux graphiques du bas représentent les résultats du protocole S2. Les taux de faux négatifs sont à gauche et les taux de faux positifs sont à droite du panel. Les boxplots montrent la distribution des faux négatifs/positifs sur les 100 jeux de données simulées.

SGSL, Bayesian et StructFDR ont un comportement inverse à Lasso, ElasticNet, Log-Contrast et PhyLasso. En effet, leurs taux de faux positifs sont très élevés, ce qui signifie qu'ils sélectionnent un grand nombre de variables non pertinentes. A l'inverse, Lasso, ElasticNet, LogContrast et PhyLasso ont des taux de faux positifs relativement faibles. Ces approches conservent des variables pertinentes. Les résultats sont opposés pour les taux de faux négatifs. Lasso, ElasticNet, LogContrast et PhyLasso sont trop restrictifs dans leur identifications et ne sélectionnent pas toutes les variables pertinentes. Dans notre étude, parmi les méthodes présentant un faible nombre de faux positifs (Lasso, ElastiNet, LogContrast et PhyLasso), nous appliquerons la méthode qui prend en compte une plus grande complexité, PhyLasso. En effet, notre étude de simulation présente une

situation simple, où les variables proches phylogénétiquement seraient liées uniquement par le fait de donner lieu à des effets similaires. Dans les vraies données, nous nous attendons à que des relations sous-jacentes plus complexes.

5.4 Conclusion

Les poumons ont longtemps été considérés comme stériles mais les techniques de séquençage ont révélé l'existence d'une flore polymicrobienne. La mucoviscidose est une maladie génétique grave et fréquente, due à la mutation du canal chlore CFTR. Elle touche principalement les voies respiratoires et le système digestif et l'infection joue un rôle important dans cette maladie. Les évaluations actuelles sont enrichies par l'analyse de quelques souches microbiennes, basée sur l'analyse NGS d'une communauté polymicrobienne présente dans les voies aériennes [O'Brien and Fothergill, 2017; Botterel et al., 2017; Nguyen et al., 2016; Quinn et al., 2016a].

Ce chapitre montre les résultats d'une analyse de données métagénomiques, à la fois sur le microbiote et le mycobiote chez des patients atteints de la mucoviscidose. Ces recherches représentent un réel enjeu pour cette maladie. Même si l'antibiotique est à ce jour l'un des principaux traitements de la mucoviscidose, il ne cible qu'une seule bactérie à la fois. En revanche les données métagénomiques permettent de détecter des associations multiples avec la sévérité de la maladie. Une seule bactérie n'est peut être pas responsable de l'exacerbation du patient mais l'association de cette bactérie avec une autre ou un champignon.

Comme expliqué dans le chapitre précédent, ces données ont une structure particulière impliquant des challenges statistiques. La précision de la technique de séquençage ainsi que les bases de données NGS permettent d'attribuer un niveau phylogénétique à un OTU. Les connaissances sont encore faibles car seules quelques bactéries ou champignons ont été étudiées de manière précise. C'est pourquoi, on retrouve de gros déséquilibre dans la précision d'attribution. En effet, certains OTUs vont être identifié au niveau règne, ce qui signifie que la connaissance est minimale, alors que d'autres gènes vont être assignés à un niveau beaucoup plus précis, l'espèce. Nous avons donc le choix d'analyser l'ensemble des OTUs au même niveau phylogénétique, mais nous avons une perte d'information importante ou prendre en compte la structure phylogénique des données mais nous rencontrons des difficultés statistiques.

L'une des principales difficultés est de comprendre l'intérêt d'analyser les données en abondance relative et pas de manières brutes. Les patients 4 et 5 ont au total 12275 et 5111 abondances brutes de bactéries et parmi elles, 50 et 23 sont des bactéries du genre *Rothia*. Notre première conclusion serait de dire que la bactérie *Rothia* est deux fois plus abondante chez le patient 4 que chez le patient 5. Cependant, si nous relativisons au total d'abondances par patient, nous remarquons que cette bactérie est légèrement plus abondante chez le patient 5 (0,40%) chez dans le patient 4 (0,45%). Il est important de considérer le microbiote comme un ensemble de bactéries et non pas de considérer les

bactéries de manière indépendante.

L'excès de 0 dans les données d'abondance représente un autre challenge dans l'analyse de ces données [Paulson et al., 2013; McDonald et al., 2012]. Dans notre cas, le nombre de 0 représente plus de 65% pour les bactéries et plus de 80% pour les champignons. Dans notre cas, la règle des OTUs présents dans au moins 10% des patients a été appliquée, mais seulement 1 OTU pour les bactéries et 3 OTUs pour les champignons ont été supprimés. Le pré-traitement de ces données est une étape importante dans l'analyse des données.

Les données issues de cette étude représentaient un cas très précis de la grande dimension. Le nombre de variables était deux fois plus grand que le nombre de patients pour les bactéries et trois fois plus grand pour les champignons. Ce faible nombre de patients est courant dans les essais cliniques des maladies rares. Les données omics étant de plus en plus présentes dans les études, ce cas de grande dimension deviendra un modèle courant dans les analyses statistiques.

Le chapitre précédent présente une revue des nombreuses méthodes statistiques existantes. Ces méthodes sont cependant récentes et peu de comparaison ont été faites entre elles. Il n'existe pas encore de pipeline bien défini pour ce type d'analyse. C'est pourquoi nous avons proposé une stratégie d'analyse allant de l'analyse de la diversité à l'analyse de co-occurrence afin de répondre à la problématique. L'étude de simulation nous a permis de faire un choix de méthodes.

Ces nouvelles recherches représentent un enjeu pour les cliniciens. En effet, la recherche de biomarqueurs représente une partie importante des essais cliniques. Nous rentrons aujourd'hui dans l'ère de la médecine de précision. « Un traitement pour un patient ». La découverte de nouveaux biomarqueurs aide à comprendre la pathologie et permettrait de cibler une sous population répondant plus efficacement à un traitement. Dans notre cas, un niveau de diversité du microbiote, une abondance élevée de plusieurs bactéries ou champignons, une présence simultanée d'une bactérie et d'un champignon pourrait être des critères de sélection de patients. Certes, nous savons qu'il existe des associations entre les microbiotes et les maladies mais il est encore difficile de définir un lien cause-effet.

6 Conclusion générale

Cette thèse porte sur l'analyse des données biomédicales de grande dimension présentant des structures diverses. Des méthodes d'apprentissage statistique sont performantes dans les cas où le nombre de variables p est égale ou plus grand que le nombre d'individus n. A l'intersection de la statistique, de l'intelligence artificielle, de l'optimisation ou encore de l'automatique, l'apprentissage statistique joue de nos jours un rôle croissant dans de nombreux domaines d'application où le principal objectif est la prédiction.

Les régressions pénalisées sont devenues populaires et sont régulièrement appliquées à des données biomédicales. Le travail réalisé durant cette thèse est principalement axé sur la régression de type Lasso qui présente un bon comportement en prédiction dans des situations courantes. Cependant, en recherche médicale, on cherche généralement à développer des modèles explicatifs. L'objectif n'est pas (uniquement) l'obtention d'une prédiction précise de la réponse d'intérêt mais l'interprétation du lien entre les « variables explicatives » et la réponse.

Cette thèse s'est articulée autour de deux thématiques.

La première a porté sur le développement et l'implémentation d'un algorithme itératif dans le cas où la variable réponse est soumise à un seuil de quantification et le nombre de variables est égal ou plus grand que le nombre de patients. Basé sur un modèle linéaire, cette méthode alterne entre une imputation par l'estimateur de Buckley-James et une estimation par régression pénalisée. L'étude de simulation a révélé de meilleures performances de cet algorithme en termes de prédiction, comparée notamment à des méthodes d'imputation naïves.

L'algorithme a été développé dans un premier temps pour répondre à un problème de prédiction. En effet, cet objectif conduit à un problème plus simple. Cependant, l'interprétation des différentes mutations liées à la chute de la charge virale est intéressante. Pour cela, une stabilisation de la méthode, par exemple par bootstrap ou Lasso adaptatif, permettrait de s'adresser à un objectif d'identification. Cependant, l'algorithme étant itératif et présentant des problèmes de convergence dans certaines situations, une étape de rééchantillonnage ajouterait des problèmes computationnels. Le Lasso adaptatif quant à lui, soulève de questions sur la définition de la variance. Son comportement en fonction

de différentes possibilités serait ainsi à évaluer.

Ce travail a été motivé par des données issues de la recherche contre le VIH. En effet, la sévérité de cette maladie est contrôlée à partir de la charge virale. Lorsque la quantité du virus est basse, la charge virale devient très faible et les techniques de mesure manquent de sensibilité pour la détecter. On dit que la charge virale est en dessous du seuil de quantification. Cet algorithme a permis de construire un modèle prédictif de la charge virale à partir des nombreuses mutations du VIH.

Aujourd'hui, les techniques de mesure de la charge virale ont progressé et peuvent descendre jusqu'à un seuil de 20 copie/mL. La prise en compte de la censure a donc moins d'intérêt dans ce contexte. Cependant, des données censurées par limite de quantification sont très présentes dans d'autres domaines. En biologie, le dosage d'une protéine peut être soumis à des seuils de quantification. En écologie, la mesure de certaines matières organiques peut également faire face à un problème de seuil de quantification. Comment gérer ce type de censure reste ainsi un sujet méthodologique d'actualité [Wei et al., 2018; Keizer et al., 2015].

La deuxième partie de la thèse a été consacrée aux problématiques soulevées par l'analyse de données de microbiote. Le séquençage haut débit du gène 16S a donné un nouvel élan à la recherche microbienne et le nombre de publications s'est amplifié depuis les années 2000. Les données de microbiote ayant une structure compositionnelle, des méthodes ont été proposées afin de garantir des résultats statistiques valides. Dans cette thèse, un état de l'art des principales méthodes existantes a été dressé. Cette revue de la littérature liste un certain nombre de méthodes dans le même ordre qu'une stratégie d'analyse pour données issues du microbiote. Les méthodes pour l'analyse de diversité sont déjà bien établies et sont largement utilisées dans de nombreux domaines, principalement en écologie. Les analyses de corrélation entre les espèces présentes dans l'environnement étudié sont très demandées par les chercheurs. En effet, il est intéressant de savoir si l'abondance d'une espèce est liée à l'abondance d'une autre espèce. Notre état de l'art fait état de six méthodes existantes répondant à l'analyse des corrélations. Elles sont toutes adaptées à des données compositionnelles. L'analyse différentielle d'abondances permet de répondre à de nombreuses questions sur l'association entre des critères cliniques et la composition du microbiote. Nous avons détaillé l'adaptation à l'analyse de données du microbiote de quelques familles de méthodes (tests statistiques, analyse de covariance, méthodes de régression). Un manque d'évaluation et de comparaison des performances des différentes méthodes proposées dans la littérature a toutefois été constaté.

Ce travail de recensement a été motivé par l'analyse de données métagénomiques issues de patients mucoviscidosiques afin d'enrichir les connaissances sur l'association entre cette

maladie et le microbiote pulmonaire. Pour analyser ces données, un protocole d'analyse complet, répondant aux questions de recherche posées, a été mis en place. Une étude de simulation a également été conduite afin de faire un choix parmi les nombreuses méthodes d'apprentissage.

La génération de données inspirées du microbiote n'est cependant pas triviale. La distribution de Dirichlet multinomiale permet certes de générer des donnés d'abondance relative. En revanche, intégrer une structure de corrélation ou de dépendance spécifique à des données phylogénétiques reste une tâche complexe.

D'autre part, ce travail a suscité des débats sur l'utilisation des méthodes d'apprentissage statistique lorsque l'étude porte sur un faible nombre de patients. En effet, en raison de la taille réduite de l'échantillon, la puissance statistique est compromise pour certaines comparaisons. Notons que des pratiques courantes telles la sélection pas à pas ou celles consistant à effectuer des analyses multivariées en utilisant uniquement les variables issues d'une analyse bivariée préalable significative sont discutables [Heinze and Dunkler, 2017]. La modélisation séparée (ajustant un modèle par variable, en opposition a la modélisation multivariée) fournit, quant à elle, des estimations de précision optimiste et peut négliger des facteurs de confusion importants. Ainsi, Nous avons effectué des analyses par une approche multivariée pour prendre en compte simultanément les bactéries et les champignons en même temps et pour augmenter la probabilité de trouver des associations entre le statut clinique et les OTUs. L'un des avantages des méthodes multivariées est leur capacité à prendre en compte l'étendue des expositions pertinentes dans l'analyse. Encore une fois, en raison de la taille restreinte de l'échantillon, une approche multivariée standard ne peut pas être appliquée. Une approche régularisée, telle que la méthode Lasso, fournit une stratégie analytique alternative aux approches conventionnelles dans l'analyse de données de faible taille. De plus, nous avons ajouté une méthode de stabilisation de sélection pour s'assurer d'une certaine robustesse des résultats.

L'analyse de l'ensemble de ces données fait appel à de méthodes diverses, certaines très récentes ou en cours de développement, qui doivent être adaptées en fonction de la question posée. Ces données structurées reflètent un système complexe et nécessitent une modélisation sophistiquée. Cela ouvre des nouveaux domaines de recherche où le rôle du biostatisticien est central.

Bibliographie

Acosta N., Whelan F. J., Somayaji R., Poonja A., Surette M. G., Rabin H. R., and Parkins M. D. The evolving cystic fibrosis microbiome : A comparative cohort study spanning sixteen years. *Annals of the American Thoracic Society*, (ja), 2017. 18

Aitchison J. and Bacon-Shone J. Log contrast models for experiments with mixtures. *Biometrika*, 71(2) : 323–330, 1984. 19, 82

Aitchison J. The statistical analysis of compositional data. 1986. 64, 65, 66, 68, 82

Aitchison J., Kay J. W., et al. Possible solution of some essential zero problems in compositional data analysis. 2003. 67

Anders S., Reyes A., and Huber W. Detecting differential usage of exons from rna-seq data. *Genome research*, 22(10) : 2008–2017, 2012. 75

Anderson M. J., Crist T. O., Chase J. M., Vellend M., Inouye B. D., Freestone A. L., Sanders N. J., Cornell H. V., Comita L. S., Davies K. F., et al. Navigating the multiple meanings of β diversity : a roadmap for the practicing ecologist. *Ecology letters*, 14(1) : 19–28, 2011. 64, 71, 74

Andréjak C. and Delhaes L. Le microbiome pulmonaire en 2015-une fenêtre ouverte sur les pathologies pulmonaires chroniques. *médecine/sciences*, 31(11) : 971–978, 2015. 18, 54

Assoumou L., Houssaïna A., Corstagliola D., Flandre P., Standardization, and clinical relevance of HIV drug resistance testing project from the forum for collaborative HIV research. Relative contributions of baseline patient characteristics and the choice of statistical methods to the variability of genotypic resistance scores : the example of didanosine. *Journal antimicrop chemother*, 65(4) : 752–760, 2010. 16

Bach F. R. Bolasso : model consistent Lasso estimation through the bootstrap. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 33–40, New York, NY, USA, 2008. ACM. 15, 28

Bacon Shone J. et al. Discrete and continuous compositions. 2008. 68

Ban Y., An L., and Jiang H. Investigating microbial co-occurrence patterns based on metagenomic compositional data. *Bioinformatics*, 31(20) : 3322–3329, 2015. 69, 70

Beaume M., Köhler T., Greub G., Manuel O., Aubert J., Baerlocher L., Farinelli L., Buckling A., Van Delden C., and Study T. S. T. C. Rapid adaptation drives invasion of airway donor microbiota by pseudomonas after lung transplantation. *Scientific reports*, 7, 2017. 18

Beerenwinkel N., Montazeri H., Schuhmacher H., Knupfer P., von Wyl V., Furrer H., Battegay M., Hirschel B., Cavassini M., Vernazza P., Bernasconi E., Yerly S., Böni J., Klimkait T., Cellerai C., Günthard H. F., and Study T. S. H. C. The individualized genetic barrier predicts treatment response in a large cohort of HIV-1 infected patients. *PLOS Computational Biology*, 9(8) : 1–11, 2013. 16

Bhatt J. M. Treatment of pulmonary exacerbations in cystic fibrosis. *European Respiratory Review*, 22(129) : 205–216, 2013. 19

Bilton D., Canny G., Conway S., Dumcius S., Hjelte L., Proesmans M., Tümmler B., Vavrova V., and De Boeck K. Pulmonary exacerbation : towards a definition for use in clinical trials. report from the eurocarecf working group on outcome parameters in clinical trials. *Journal of Cystic Fibrosis*, 10 : S79–S81, 2011. 19

Bolker B. M., Brooks M. E., Clark C. J., Geange S. W., Poulsen J. R., Stevens M. H. H., and White J.-S. S. Generalized linear mixed models : a practical guide for ecology and evolution. *Trends in ecology & evolution*, 24(3) : 127–135, 2009. 75, 78

Botterel F., Angebault C., Cabaret O., Stressmann F. A., Costa J.-M., Wallet F., Wallaert B., Bruce K., and Delhaes L. Fungal and bacterial diversity of airway microbiota in adults with cystic fibrosis : Concordance between conventional methods and ultra-deep sequencing, and their practical use in the clinical laboratory. *Mycopathologia*, pages 1–13, 2017. 19, 20, 131

Boutin S. and Dalpke A. H. Acquisition and adaptation of the airway microbiota in the early life of cystic fibrosis patients. *Molecular and cellular pediatrics*, 4(1) : 1, 2017. 18

Bray J. R. and Curtis J. T. An ordination of the upland forest communities of southern wisconsin. *Ecological monographs*, 27(4) : 325–349, 1957. 62

Breiman L. et al. Statistical modeling : The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3) : 199–231, 2001. 14

Brun-Vezinet F., Costagliola D., Khaled M. A., Calvez V., Clavel F., Clotet B., Haubrich R., Kempf D., King M., Kuritzkes D., et al. Clinically validated genotype analysis : guiding principles and statistical concerns. *Antiviral therapy*, 9(4) : 465–478, 2004. 35

Buckley J. and James I. Linear regression with censored data. *Biometrika*, 66 : 429–436, 1979. 16, 34

Cai T., Huang J., and Tian L. Regularized estimation for the accelerated failure time model. *Biometrics*, 65 : 394–404, 2009. 17, 36

Caporaso J. G., Kuczynski J., Stombaugh J., Bittinger K., Bushman F. D., Costello E. K., Fierer N., Peña A. G., Goodrich J. K., Gordon J. I., et al. Qiime allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5) : 335–336, 2010a. 55

Caporaso J. G., Kuczynski J., Stombaugh J., Bittinger K., Bushman F. D., Costello E. K., Fierer N., Pena A. G., Goodrich J. K., Gordon J. I., et al. Qiime allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5) : 335, 2010b. 91

Carmody L. A., Zhao J., Schloss P. D., Petrosino J. F., Murray S., Young V. B., Li J. Z., and LiPuma J. J. Changes in cystic fibrosis airway microbiota at pulmonary exacerbation. *Annals of the American Thoracic Society*, 10(3) : 179–187, 2013. 19, 57

Chao A., Chazdon R. L., Colwell R. K., and Shen T.-J. A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology letters*, 8(2) : 148–159, 2005. 62

Charlson E. S., Bittinger K., Haas A. R., Fitzgerald A. S., Frank I., Yadav A., Bushman F. D., and Collman R. G. Topographical continuity of bacterial populations in the healthy human respiratory tract. *American journal of respiratory and critical care medicine*, 184(8) : 957–963, 2011. 18, 54

Charlson E. S., Diamond J. M., Bittinger K., Fitzgerald A. S., Yadav A., Haas A. R., Bushman F. D., and Collman R. G. Lung-enriched organisms and aberrant bacterial and fungal respiratory microbiota after lung transplant. *American journal of respiratory and critical care medicine*, 186(6) : 536–545, 2012. 19, 20

Chen J., Bushman F. D., Lewis J. D., Wu G. D., and Li H. Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*, 14, 2013. 80, 81, 83

Chen J. and Li H. *Kernel Methods for Regression Analysis of Microbiome Compositional Data*, pages 191–201. Springer New York, New York, NY, 2013a. 72, 73, 74

Chen J. and Li H. Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis. *The annals of applied statistics*, 7(1), 2013b. 79, 81

Chen J., Bittinger K., Charlson E. S., Hoffmann C., Lewis J., Wu G. D., Collman R. G., Bushman F. D., and Li H. Associating microbiome composition with environmental covariates using generalized unifrac distances. *Bioinformatics*, 28(16) : 2106–2113, 2012a. 63

Chen J., Bittinger K., Charlson E. S., Hoffmann C., Lewis J., Wu G. D., Collman R. G., Bushman F. D., and Li H. Associating microbiome composition with environmental covariates using generalized unifrac distances. *Bioinformatics*, 28(16) : 2106–2113, 2012b. 63

Chen L., Liu H., Kocher J.-P. A., Li H., and Chen J. glmgraph : an r package for variable selection and predictive modeling of structured genomic data. *Bioinformatics*, 31(24) : 3991–3993, 2015. 85, 87

Chiquet J., Mariadassou M., and Robin S. Variational inference for sparse network reconstruction from count data. In *Proceedings of the 36th International Conference on Machine Learning*, 2019. 69, 70

Chung M., Long Q., and Johnson B. A. A tutorial on rank-based coefficient estimation for censored data in small-and large-scale problems. *Statistics and computing*, 23(5) : 601–614, 2013. 17

Clarke K. R. Non-parametric multivariate analyses of changes in community structure. *Austral Ecology*, 18(1) : 117–143, 1993. 71, 74

Conrad D., Haynes M., Salamon P., Rainey P. B., Youle M., and Rohwer F. Cystic fibrosis therapy : a community ecology perspective. *American journal of respiratory cell and molecular biology*, 48(2) : 150–156, 2013. 20

Cox M. J., Turek E. M., Hennessy C., Mirza G. K., James P. L., Coleman M., Jones A., Wilson R., Bilton D., Cookson W. O., et al. Longitudinal assessment of sputum microbiome by sequencing of the 16s rrna gene in non-cystic fibrosis bronchiectasis patients. *PLOS ONE*, 12 (2) : e0170622, 2017. 18

Cozzi-Lepri A. Initiatives for developing and comparing genotype interpretation systems : external validation of existing rule-based interpretation systems for abacavir against virological response. *HIV medicine*, 9(1) : 27–40, 2008. 17, 36

Cozzi-Lepri A., Prosperi M. C. F., Kjær J., Dunn D., Paredes R., Sabin C. A., Lundgren J. D., Phillips A. N., Pillay D., for the EuroSIDA, and the United Kingdom CHIC/United Kingdom HDRD Studies. Can linear regression modeling help clinicians in the interpretation of genotypic resistance data ? an application to derive a lopinavir-score. *PLOS ONE*, 6(11) : 1–9, 2011. 16

Cribbs S. K. and Beck J. M. Microbiome in the pathogenesis of cystic fibrosis and lung transplant-related disease. *Translational Research*, 179 : 84–96, 2017. 18

Cryan J. F. and O'Mahony S. The microbiome-gut-brain axis : from bowel to behavior. *Neurogastroenterology & Motility*, 23(3) : 187–192, 2011. 18, 54

Datta S., Le-Rademacher J., and Datta S. Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and lasso. *Biometrics*, 63 : 259–271, 2007. 17

de Koff E. M., de Winter-de Groot K. M., and Bogaert D. Development of the respiratory tract microbiota in cystic fibrosis. *Current opinion in pulmonary medicine*, 22(6) : 623–628, 2016. 19

Del Greco M. F., Pattaro C., Minelli C., and Thompson J. R. Bayesian analysis of censored response data in family-based genetic association studies. *Biometrical Journal*, 58(5) : 1039–1053, 2016. 33

Delhaes L., Monchy S., Fréalle E., Hubans C., Salleron J., Leroy S., Prevotat A., Wallet F., Wallaert B., Dei-Cas E., et al. The airway microbiota in cystic fibrosis : a complex fungal and bacterial community—implications for therapeutic management. *PLOS ONE*, 7(4) : e36313, 2012. 13, 19, 20

Dickson R. P., Erb-Downward J. R., and Huffnagle G. B. The role of the bacterial microbiome in lung disease. *Expert review of respiratory medicine*, 7(3) : 245–257, 2013. 18, 54

Dinse G., Jusko A., Ho L., Annam K., Graubard B., Hertz-Picciotto I., Miller F., Gillespie B., and Weinberg C. Accomodating measurements below a limit of detection : A novel application of Cox regression. *American Journal of Epidemiology*, 179(8) : 1018–1024, 2014. 33

DiRienzo A. G. Parsimonious covariate selection with censored outcomes. *Biometrics*, 72 : 452–462, 2016. 17

Dray S., Dufour A.-B., et al. The ade4 package : implementing the duality diagram for ecologists. *Journal of statistical software*, 22(4) : 1–20, 2007. 65

Durack J., Lynch S. V., Nariya S., Bhakta N. R., Beigelman A., Castro M., Dyer A.-M., Israel E., Kraft M., Martin R. J., et al. Features of the bronchial bacterial microbiome associated with atopy, asthma, and responsiveness to inhaled corticosteroid treatment. *Journal of Allergy and Clinical Immunology*, 140(1) : 63–75, 2017. 18

Efron B., Hastie T., Johnstone I., and Tibshirani R. Least angle regression. *The Annals of Statistics*, 32 : 407–451, 2004. 26

Egozcue J. J., Pawlowsky-Glahn V., Mateu-Figueras G., and Barcelo-Vidal C. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3) : 279–300, 2003. 66

Eilers P. H., Röder E., Savelkoul H. F., and van Wijk R. G. Quantile regression for the statistical analysis of immunological data with many non-detects. *BMC Immunology*, 13 : 13–37, 2012. 33

Falissard B. Epistémologie de la statistique à l'heure du tout digital. Université de Bordeaux, France, 2018. Journées de l'Ecole Doctorale SP2. 14

Fan J. and Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456) : 1348–1360, 2001. 26

Fang H., Huang C., H.Zhao, and Deng M. Cclasso : correlation inference for compositional data through lasso. *Bioinformatics*, 31(19) : 3172, 2015. 68, 70

Faust K., Sathirapongsasuti J. F., Izard J., Segata N., Gevers D., Raes J., and Huttenhower C. Microbial co-occurrence relationships in the human microbiome. *PLOS Comput Biol*, 8(7) : e1002606, 2012. 68, 70

Feigelman R., Kahlert C. R., Baty F., Rassouli F., Kleiner R. L., Kohler P., Brutsche M. H., and von Mering C. Sputum dna sequencing in cystic fibrosis : non-invasive access to the lung microbiome and to pathogen details. *Microbiome*, 5(1) : 20, 2017. 18

Fernandes A. D., Macklaim J. M., Linn T. G., Reid G., and Gloor G. B. Anova-like differential expression (aldex) analysis for mixed population rna-seq. *PLOS ONE*, 8(7) : e67019, 2013. 76

Fernandes A. D., Reid J. N., Macklaim J. M., McMurrough T. A., Edgell D. R., and Gloor G. B. Unifying the analysis of high-throughput sequencing datasets : characterizing rna-seq, 16s rrna gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, 2(1) : 15, 2014. 78

Filkins L., Hampton T., Gifford A., Gross M., Hogan D., Sogin M., Morrison H., Paster B., and O'Toole G. Prevalence of streptococci and increased polymicrobial diversity associated with cystic fibrosis patient stability. *Journal of bacteriology*, 194(17) : 4709–4717, 2012. 19

Flandre P., Marcelin A.-G., Pavie J., Shmidely N., Wirden M., Lada O., Bernard M.-C., Molina J.-M., and Calvez V. Comparison of tests and procedures to build clinically relevant genotypic scores : application to the jaguar study. *Antivir Ther*, 10(4) : 479–87, 2005. 35

Fodor A. A., Klem E. R., Gilpin D. F., Elborn J. S., Boucher R. C., Tunney M. M., and Wolfgang M. C. The adult cystic fibrosis airway microbiota is stable over time and infection type, and highly resilient to antibiotic treatment of exacerbations. *PLOS ONE*, 7(9) : e45001, 2012. 19

Frank L. E. and Friedman J. H. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2) : 109–135, 1993. 26

Frayman K. B., Armstrong D. S., Grimwood K., and Ranganathan S. C. The airway microbiota in early cystic fibrosis lung disease. *Pediatric Pulmonology*, 2017. 18

Friedman J., Hastie T., and Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1) : 1–22, 2010. 25, 26

Friedman J., Alm E. J., and von Mering C. Inferring correlation networks from genomic survey data. *PLOS Computational Biology*, 8, 2012. 68, 70

Fu P., Hughes J., Zeng G., Hanook S., Orem J., Mwanda O., and Remick S. A comparative investigation of methods for longitudinal data with limits of detection through a case study. *Statistical Methods in Medical Research*, 25(1) : 153–166, 2016. 34

Fu W. J. Penalized regressions : the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3) : 397–416, 1998. 26

Fukuyama J., McMurdie P. J., LES DETHLEFSEN D. A. R., and HOLMES S. Comparisons of distance methods for combining covariates and abundances in microbiome studies. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 213. NIH Public Access, 2012. 63

Garcia T. P., Müller S., Carroll R. J., and Walzem R. L. Identification of important regressor groups, subgroups and individuals via regularization methods : application to gut microbiome data. *Bioinformatics*, 30(6) : 831–837, 2014. 83, 87

Gillespie B. W., Chen Q., Reichert H., Franzblau A., Hedgeman E., Lepkowski J., Adriaens P., Demond A., Luksemburg W., and Garabrant D. H. Estimating population distributions when some data are below a limit of detection by using a reverse kaplan-meier estimator. *Epidemiology*, 21 : 64–70, 2010. 33

Goleva E., Jackson L. P., Harris J. K., Robertson C. E., Sutherland E. R., Hall C. F., Good Jr J. T., Gelfand E. W., Martin R. J., and Leung D. Y. The effects of airway microbiome on corticosteroid responsiveness in asthma. *American journal of respiratory and critical care medicine*, 188(10) : 1193–1201, 2013. 18

Goss C. H. and Burns J. L. Exacerbations in cystic fibrosis · 1 : epidemiology and pathogenesis. *Thorax*, 62(4) : 360–367, 2007. 19

Guo P. and Hao Y. *SparseLearner : Sparse Learning Algorithms Using a LASSO-Type Penalty for Coefficient Estimation and Model Prediction*, 2015. R package version 1.0-2. 28

Hardcastle T. J. and Kelly K. A. Empirical bayesian analysis of paired high-throughput sequencing data with a beta-binomial distribution. *BMC bioinformatics*, 14(1) : 135, 2013. 76, 78

Healy B. C., DeGruttola V. G., and Hu C. Accommodating uncertainty in a tree set for function estimation. *Statistical applications in genetics and molecular biology*, 7(1), 2008. 36

Heinze G. and Dunkler D. Five myths about variable selection. *Transplant International*, 30 (1) : 6–10, 2017. 135

Heirali A. A., Workentine M. L., Acosta N., Poonja A., Storey D. G., Somayaji R., Rabin H. R., Whelan F. J., Surette M. G., and Parkins M. D. The effects of inhaled aztreonam on the cystic fibrosis lung microbiome. *Microbiome*, 5(1) : 51, 2017. 18

Helmbold D. P. and Long P. M. On the necessity of irrelevant variables. *Journal of Machine Learning Research*, 13(Jul) : 2145–2170, 2012. 14

Helsel D. R. More than obvious : Better methods for interpreting nondetect data. *Environmental Science & Technology*, 39(20) : 419A–423A, 2005. 33

Hilty M., Burke C., Pedro H., Cardenas P., Bush A., Bossley C., Davies J., Ervine A., Poulter L., Pachter L., et al. Disordered microbial communities in asthmatic airways. *PLOS ONE*, 5(1) : e8578, 2010. 18

Hirsch M. S., Günthard H. F., Schapiro J. M., Vézinet F. B., Clotet B., Hammer S. M., Johnson V. A., Kuritzkes D. R., Mellors J. W., Pillay D., et al. Antiretroviral drug resistance testing in adult HIV-1 infection : 2008 recommendations of an international aids society-USA panel. *Clinical Infectious Diseases*, 47(2) : 266–285, 2008. 16

Hoerl A. E. and Kennard R. W. Ridge regression : applications to nonorthogonal problems. *Technometrics*, 12(1) : 69–82, 1970. 15, 25

Hofstra L. M., Sauvageot N., Albert J., Alexiev I., Garcia F., Struck D., Van de Vijver D. A., Åsjö B., Beshkov D., Coughlan S., et al. Transmission of HIV drug resistance and the predicted effect on current first-line regimens in europe. *Clinical infectious diseases*, 62(5) : 655–663, 2016. 16

Hooks K. B. and O'Malley M. A. Dysbiosis and its discontents. *mBio*, 8(5), 2017. 59

Hosgood H. D., Sapkota A. R., Rothman N., Rohan T., Hu W., Xu J., Vermeulen R., He X., White J. R., Wu G., et al. The potential role of lung microbiota in lung cancer attributed to household coal burning exposures. *Environmental and molecular mutagenesis*, 55(8) : 643–651, 2014. 19

Huang J., Ma S., and Xie H. Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics*, 62 : 813–820, 2006. 17

Huang X., Pan W., Park S., Han X., Miller L. W., and Hall J. Modeling the relationship between lvad support time and gene expression changes in the human heart by penalized partial least squares. *Bioinformatics*, 20(6) : 888–894, 2004. 17, 36

Huang Y. J., Sethi S., Murphy T., Nariya S., Boushey H. A., and Lynch S. V. Airway microbiome dynamics in exacerbations of chronic obstructive pulmonary disease. *Journal of clinical microbiology*, 52(8) : 2813–2823, 2014. 18

Hughes J. P. Mixed effects models with censored data with application to HIV RNA levels. *Biometrics*, 55 : 625–629, 1999. 34

Huttenhower C., Gevers D., Knight R., Abubucker S., Badger J. H., Chinwalla A. T., Creasy H. H., Earl A. M., FitzGerald M. G., Fulton R. S., et al. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402) : 207, 2012. 53, 58, 59

Ishwaran H., Kogalur U. B., Blackstone E. H., and Lauer M. S. Random survival forests. *The Annals of Applied Statistics*, 2 : 841–860, 2008. 17

Jacqmin-Gadda H., Thiébaut R., Chêne G., and Commenges D. Analysis of left-censored longitudinal data with application to viral load in HIV infection. *Biostatistics*, 1(4) : 355–368, 2000. 34

Johnson B. A. Rank-based estimation in the 1-regularized partly linear model for censored outcomes with application to integrated analyses of clinical predictors and gene expression data. *Biostatistics*, 10 : 659–666, 2009a. 17

Johnson B. A. Variable selection in semiparametric linear regression with censored data. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 70 : 351–370, 2008. 17, 36

Johnson B. A. On lasso for censored data. *Electronic Journal of Statistics*, 3 : 485–506, 2009b. 17, 36

Johnson V. A., Brun-Vézinet F., Clotet B., Gunthard H., Kuritzkes D. R., Pillay D., Schapiro J. M., and Richman D. D. Update of the drug resistance mutations in **HIV**-1 : December 2009. *Top HIV Med*, 17(5) : 138–145, 2009. 36, 37

Keizer R. J., Jansen R. S., Rosing H., Thijssen B., Beijnen J. H., Schellens J. H., and Huitema A. D. Incorporation of concentration data below the limit of quantification in population pharmacokinetic analyses. *Pharmacology research & perspectives*, 3(2) : e00131, 2015. 51, 134

Koeth R. A., Wang Z., Levison B. S., Buffa J. A., Org E., Sheehy B. T., Britt E. B., Fu X., Wu Y., Li L., et al. Intestinal microbiota metabolism of l-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nature medicine*, 19(5) : 576–585, 2013. 18, 54

Koh H., Blaser M. J., and Li H. A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping. *Microbiome*, 5(1) : 45, 2017. 73, 74

Kong H. H., Oh J., Deming C., Conlan S., Grice E. A., Beatson M. A., Nomicos E., Polley E. C., Komarow H. D., Murray P. R., et al. Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. *Genome research*, 22(5) : 850–859, 2012. 54

Kraemer N., Schaefer J., and Boulesteix A.-L. Regularized estimation of large-scale gene regulatory networks using gaussian graphical models. *BMC Bioinformatics*, 10(384), 2009. 28

Kramer R., Sauer-Heilborn A., Welte T., Guzman C. A., Abraham W.-R., and Höfle M. G. Cohort study of airway mycobiome in adult cystic fibrosis patients : differences in community structure between fungi and bacteria reveal predominance of transient fungal elements. *Journal of clinical microbiology*, 53(9) : 2900–2907, 2015. 20

Krause R., Moissl-Eichinger C., Halwachs B., Gorkiewicz G., Berg G., Valentin T., Prattes J., Högenauer C., and Zollner-Schwetz I. Mycobiome in the lower respiratory tract – a clinical perspective. *Frontiers in Microbiology*, 7 : 2169, 2017. 20

Kurtz Z. D., Müller C. L., Miraldi E. R., Littman D. R., Blaser M. J., and Bonneau R. A. Sparse and compositionally robust inference of microbial ecological networks. *PLOS computational biology*, 11(5) : e1004226, 2015. 69, 70

LaRossa P., Brooks J. P., Deych E., Boone E. L., Edwards D. J., Wang Q., Sodergren E., Weinstock G., and Shannon W. D. Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLOS ONE*, 2012. 73, 74

Lauritzen S. L. *Graphical models*, volume 17. Clarendon Press, 1996. 69

Le Cao K.-A., Costello M.-E., Lakis V. A., Bartolo F., Chua X.-Y., Brazeilles R., and Rondeau P. Mixmc : a multivariate statistical framework to gain insight into microbial communities. *PLOS ONE*, 11(8) : e0160169, 2016. 86, 87

Lee M., Kong L., and Weissfeld L. Multiple imputation for left-censored biomarker data based on Gibbs sampling method. *Statistics in Medicine*, 31 : 1838—-1848, 2012. 33

Lee S. H., Sung J. Y., Yong D., Chun J., Kim S. Y., Song J. H., Chung K. S., Kim E. Y., Jung J. Y., Kang Y. A., et al. Characterization of microbiome in bronchoalveolar lavage fluid of patients with lung cancer comparing with benign mass like lesions. *Lung Cancer*, 102 : 89–95, 2016. 19

Leite C. d. C. F., Folescu T. W., de Cássia Firmida M., Cohen R. W. F., Leão R. S., de Freitas F. A. D., Albano R. M., da Costa C. H., and Marques E. A. Monitoring clinical and microbiological evolution of a cystic fibrosis patient over 26 years : experience of a brazilian cf centre. *BMC pulmonary medicine*, 17(1) : 100, 2017. 18

Li H. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual review of statistics and its application*, 2(1) : 73–94, 2015. 56

Lim Y. W., Schmieder R., Haynes M., Willner D., Furlan M., Youle M., Abbott K., Edwards R., Evangelista J., Conrad D., et al. Metagenomics and metatranscriptomics : windows on cf-associated viral and microbial communities. *Journal of Cystic Fibrosis*, 12(2) : 154–164, 2013. 19

Lin W., Shi P., Feng R., Li H., et al. Variable selection in regression with compositional covariates. *Biometrika*, 101(4) : 785–797, 2014. 82, 87

Liu D., Ghosh D., and Lin X. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC bioinformatics*, 9(1) : 292, 2008. 73

Liu H., Xu X., and Li J. J. A bootstrap lasso+ partial ridge method to construct confidence intervals for parameters in high-dimensional sparse linear models. *arXiv preprint arXiv :1706.02150*, 2017. 28

Love M. I., Huber W., and Anders S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12) : 550, 2014. 76, 78

Lozupone C. and Knight R. Unifrac : a new phylogenetic method for comparing microbial communities. *APPL. ENVIRON*, pages 71–8228, 2005. 63

Lozupone C. A., Hamady M., Kelley S. T., and Knight R. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Applied and environmental microbiology*, 73(5) : 1576–1585, 2007. 63

Lynn H. S. Maximum likelihood inference for left-censored HIV RNA data. *Statistics in Medicine*, 20 : 33–45, 2001. 34

MacDougall L. K., Broukhanski G., Simor A., Johnstone J., Mubareka S., McGeer A., Daneman N., Garber G., and Brown K. A. Comparison of qpcr versus culture for the detection and quantification of clostridium difficile environmental contamination. *PLOS ONE*, 13(8) : e0201569, 2018. 13

Mandal S., Van Treuren W., White R. A., Eggesbø M., Knight R., and Peddada S. D. Analysis of composition of microbiomes : a novel method for studying microbial composition. *Microbial ecology in health and disease*, 26, 2015. 77, 78

Marcon E. Mesures de la Biodiversité. Lecture, 2015. 60

Mariette J. and Villa-Vialaneix N. Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics*, 34(6) : 1009–1015, 2017. 64

Marks G., Gardner L. I., Craw J., Giordano T. P., Mugavero M. J., Keruly J. C., Wilson T. E., Metsch L. R., Drainoni M.-L., and Malitz F. The spectrum of engagement in hiv care : do more than 19% of hiv-infected persons in the us have undetectable viral load ? *Clinical infectious diseases*, 53(11) : 1168–1169, 2011. 67

Marri P. R., Stern D. A., Wright A. L., Billheimer D., and Martinez F. D. Asthma-associated differences in microbial composition of induced sputum. *Journal of Allergy and Clinical Immunology*, 131(2) : 346–352, 2013. 18

Marschner I., Betensky R., DeGruttola V., Hammer S., and Kuritzkes D. Clinical trials using HIV-1 RNA-based primary endpoints : Statistical analysis and potential biase. *Journal of acquired immune deficiency syndromes and human retrovirology*, 20(3) : 220–227, 1999. 33

Marsland B. J. and Gollwitzer E. S. Host-microorganism interactions in lung diseases. *Nature reviews. Immunology*, 14(12) : 827, 2014. 18, 54

Martín-Fernández J. and Thió-Henestrosa S. Rounded zeros : some practical aspects for compositional data. *Geological Society, London, Special Publications*, 264(1) : 191–201, 2006. 67

Martín-Fernández J.-A., Barceló-Vidal C., and Pawlowsky-Glahn V. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, 35(3) : 253–278, 2003. 67

Martın-Fernandez J. A., Palarea-Albaladejo J., and Olea R. A. Dealing with zeros. *Compositional data analysis : Theory and applications*, pages 43–58, 2011. 67

Marx V. Biology : The big challenges of big data. *Nature*, 498(7453) : 255–260, 2013. 13

McDonald D., Clemente J. C., Kuczynski J., Rideout J. R., Stombaugh J., Wendel D., Wilke A., Huse S., Hufnagle J., Meyer F., et al. The biological observation matrix (biom) format or : how i learned to stop worrying and love the ome-ome. *GigaScience*, 1(1) : 7, 2012. 56, 132

McMurdie P. J. and Holmes S. Waste not, want not : why rarefying microbiome data is inadmissible. *PLOS computational biology*, 10(4) : e1003531, 2014. 56

Meinshausen N. and Bühlmann P. Stability selection. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 72(4) : 417–473, 2010. 82

Molyneaux P. L., Mallia P., Cox M. J., Footitt J., Willis-Owen S. A., Homola D., Trujillo-Torralbo M.-B., Elkin S., Kon O. M., Cookson W. O., et al. Outgrowth of the bacterial airway microbiome after rhinovirus exacerbation of chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, 188(10) : 1224–1231, 2013. 18

Mooney S. J. and Pejaver V. Big data in public health : terminology, machine learning, and privacy. *Annual review of public health*, 39 : 95–112, 2018. 13

Morris A., Beck J. M., Schloss P. D., Campbell T. B., Crothers K., Curtis J. L., Flores S. C., Fontenot A. P., Ghedin E., Huang L., et al. Comparison of the respiratory microbiome in healthy nonsmokers and smokers. *American journal of respiratory and critical care medicine*, 187(10) : 1067–1075, 2013. 54

Müller P. and van de Geer S. Censored linear model in high dimensions. *TEST*, pages 75–92, 2015. 16

Nguyen L. D. N., Deschaght P., Merlin S., Loywick A., Audebert C., Van Daele S., Viscogliosi E., Vaneechoutte M., and Delhaes L. Effects of propidium monoazide (pma) treatment on mycobiome and bacteriome analysis of cystic fibrosis airways during exacerbation. *PLOS ONE*, 11(12) : e0168860, 2016. 18, 19, 20, 131

Nie L., Chu H., Liu C., Cole S. R., Vexler A., and Schisterman E. F. Linear regression with an independent variable subject to a detection limit. *Epidemiology*, 21 : 17–24, 2010. 34

O'Brien S. and Fothergill J. L. The role of multispecies social interactions in shaping pseudomonas aeruginosa pathogenicity in the cystic fibrosis lung. *FEMS Microbiology Letters*, 364 (15), 2017. 19, 20, 131

Oksanen J., Blanchet F. G., Friendly M., Kindt R., Legendre P., McGlinn D., Minchin P. R., O'Hara R. B., Simpson G. L., Solymos P., Stevens M. H. H., Szoecs E., and Wagner H. *vegan : Community Ecology Package*, 2017. R package version 2.4-4. 61, 62

on Antiretroviral Guidelines for Adults P. and Adolescents. Guidelines for the use of antiretroviral agents in adults and adolescents living with hiv, 2018. 34

Palarea-Albaladejo J., Martín-Fernández J. A., and Gómez-García J. A parametric approach for dealing with compositional rounded zeros. *Mathematical Geology*, 39(7) : 625–645, 2007. 67

Paliy O. and Shankar V. Application of multivariate statistical techniques in microbial ecology. *Molecular ecology*, 25(5) : 1032–1057, 2016. 64

Pan W., Kim J., Zhang Y., Shen X., and Wei P. A powerful and adaptive association test for rare variants. *Genetics*, 197(4) : 1081–1095, 2014. 72, 74

Paradis E., Claude J., and Strimmer K. APE : analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20 : 289–290, 2004. 64

Parks D. H., Tyson G. W., Hugenholtz P., and Beiko R. G. Stamp : statistical analysis of taxonomic and functional profiles. *Bioinformatics*, 30(21) : 3123–3124, 2014. 75, 78

Paulson J., Stine O., Corrada Bravo H., and Pop M. Robust methods for differential abundance analysis in marker gene surveys. *Nat Methods*, 10 : 1200–1202, 2013. 56, 76, 78, 132

Paxton W., Coombs R., McElrath M., Keefer M., Hughes J., Sinangil F., Chernoff D., Demeter L., B B. W., and Corey L. Longitudinal analysis of quantitative virologic measures in human immunodeficiency virus-infected subjects with > or = 400 cd4 lymphocytes : implications for applying measurements to individual patients. national institute of allergy and infectious diseases aids vaccine evaluation group. *Journal of Infectious Disease*, 175(2) : 247–254, 1997. 33

Pierotti M. E., Martín-Fernández J. A., and Seehausen O. Mapping individual variation in male mating preference space : multiple choice in a color polymorphic cichlid fish. *Evolution*, 63 (9) : 2372–2388, 2009. 67

Pittman J. E., Wylie K. M., Akers K., Storch G. A., Hatch J., Quante J., Frayman K. B., Clarke N., Davis M., Stick S. M., et al. Association of antibiotics, airway microbiome, and inflammation in infants with cystic fibrosis. *Annals of the American Thoracic Society*, 14(10) : 1548–1555, 2017. 18

Powell J. L. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, 25 : 303–325, 1984. 33

Powell J. L. Censored regression quantiles. *Journal of Econometrics*, 32 : 143–155, 1986. 33

Qin J., Li Y., Cai Z., Li S., Zhu J., Zhang F., Liang S., Zhang W., Guan Y., Shen D., et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418) : 55–60, 2012. 18, 54

Quinn R. A., Lim Y. W., Maughan H., Conrad D., Rohwer F., and Whiteson K. L. Biogeochemical forces shape the composition and physiology of polymicrobial communities in the cystic fibrosis lung. *MBio*, 5(2) : e00956–13, 2014. 20

Quinn R. A., Whiteson K., Lim Y.-W., Salamon P., Bailey B., Mienardi S., Sanchez S. E., Blake D., Conrad D., and Rohwer F. A winogradsky-based culture system shows an association between microbial fermentation and cystic fibrosis exacerbation. *The ISME journal*, 9(4) : 1024–1038, 2015. 19

Quinn R. A., Lim Y. W., Mak T. D., Whiteson K., Furlan M., Conrad D., Rohwer F., and Dorrestein P. Metabolomics of pulmonary exacerbations reveals the personalized nature of cystic fibrosis disease. *PeerJ*, 4 : e2174, 2016a. 19, 20, 131

Quinn R. A., Whiteson K., Lim Y. W., Zhao J., Conrad D., LiPuma J. J., Rohwer F., and Widder S. Ecological networking of cystic fibrosis lung infections. *NPJ Biofilms and Microbiomes*, 2 : 1, 2016b. 19, 20

Rabinowitz M., Myers L., Banjevic M., Chan A., Sweetkind-Singer J., Haberer J., McCann K., and Wolkowicz R. Accurate prediction of HIV-1 drug response from the reverse transcriptase and protease amino acid sequences using sparse models created by convex optimization. *Bioinformatics*, 22(5) : 541–549, 2006. 16

Ramette A. Multivariate analyses in microbial ecology. *FEMS microbiology ecology*, 62(2) : 142–160, 2007. 64

Rhee S., Taylor J., Wadhera G., Ben-Hur A., Brutlag D., and Shafer R. Genotypic predictos of human immunodeficiency cirus type 1 drug resistance. *Proc Natl Acad Sci U S A*, 103(46) : 17355–17360, 2006. 16

Robinson M. D. and Oshlack A. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3) : R25, 2010. 78

Robinson M. D., McCarthy D. J., and Smyth G. K. edger : a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1) : 139–140, 2010. 76

Romero R., Hassan S. S., Gajer P., Tarca A. L., Fadrosh D. W., Bieda J., Chaemsaithong P., Miranda J., Chaiworapongsa T., and Ravel J. The vaginal microbiota of pregnant women who subsequently have spontaneous preterm labor and delivery and those with a normal delivery at term. *Microbiome*, 2(1) : 18, 2014. 54

Rush S. T., Lee C. H., Mio W., and Kim P. T. The phylogenetic lasso and the microbiome. *arXiv preprint arXiv :1607.08877*, 2016. 85, 87

Schloss P. D., Westcott S. L., Ryabin T., Hall J. R., Hartmann M., Hollister E. B., Lesniewski R. A., Oakley B. B., Parks D. H., Robinson C. J., et al. Introducing mothur : open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 75(23) : 7537–7541, 2009. 55

Schumi J. and DeGruttola V. Resampling-based analyses of the effects of combinations of hiv genetic mutations on drug susceptibility. *Statistics in medicine*, 27(23) : 4740–4757, 2008. 36

Segata N., Izard J., Waldron L., Gevers D., Miropolsky L., Garrett W. S., and Huttenhower C. Metagenomic biomarker discovery and explanation. *Genome biology*, 12(6) : R60, 2011. 75, 78

Shafer R. W. and Schapiro J. M. **HIV**-1 drug resistance mutations : an updated framework for the second decade of haart. *AIDS reviews*, 10(2) : 67, 2008. 37

Shankar J., Szpakowski S., Solis N. V., Mounaud S., Liu H., Losada L., Nierman W. C., and Filler S. G. A systematic evaluation of high-dimensional, ensemble-based regression for exploring large model spaces in microbiome analyses. *BMC bioinformatics*, 16(1) : 31, 2015. regeval repository :`http://github.com/openpencil/regeval`. 81, 83, 87

Shmueli G. et al. To explain or to predict ? *Statistical science*, 25(3) : 289–310, 2010. 14

Shows J. H., Lu W., and Zhang H. H. Sparse estimation and inference for censored median regression. *Journal of Statistical Planning and Inference*, 140, 2010. 16

Simon N., Friedman J., Hastie T., and Tibshirani R. *SGL : Fit a GLM (or cox model) with a combination of lasso and group lasso regularization*, 2013a. R package version 1.1. 30

Simon N., Friedman J., Hastie T., and Tibshirani R. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2) : 231–245, 2013b. 30

Simpson J. L., Daly J., Baines K. J., Yang I. A., Upham J. W., Reynolds P. N., Hodge S., James A. L., Hugenholtz P., Willner D., et al. Airway dysbiosis : Haemophilus influenzae and tropheryma in poorly controlled asthma. *European Respiratory Journal*, 47(3) : 792–800, 2016. 18

Society E. C. F. *European Cystic Fibrosis Society Patient Registry Annual Report 2014*, 2014. 89

Stenbit A. E. and Flume P. A. Pulmonary exacerbations in cystic fibrosis. *Current opinion in pulmonary medicine*, 17(6) : 442–447, 2011. 19

Sverrild A., Kiilerich P., Brejnrod A., Pedersen R., Porsbjerg C., Bergqvist A., Erjefält J. S., Kristiansen K., and Backer V. Eosinophilic airway inflammation in asthmatic patients is associated with an altered airway microbiome. *Journal of Allergy and Clinical Immunology*, 140(2) : 407–417, 2017. 18

Thiébaut R., Hejblum B. P., and Richert L. L'analyse des «big data» en recherche clinique. *Epidemiology and Public Health/Revue d'Epidémiologie et de Santé Publique*, 62(1) : 1–4, 2014. 13

Thorsen J., Brejnrod A., Mortensen M., Rasmussen M. A., Stokholm J., Al-Soud W. A., Sørensen S., Bisgaard H., and Waage J. Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16s rrna gene amplicon data analysis methods used in microbiome studies. *Microbiome*, 4(1) : 62, 2016. 57, 75

Tibshirani R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 15, 23, 25

Tibshirani R. The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16 : 385–395, 1997. 16

Tobin J. Estimation of relationships for limited dependent variables. *Econometrica*, 26 : 24–36, 1958. 16, 34

Touati K., Delhaes L., et al. The airway colonization by opportunistic filamentous fungi in patients with cystic fibrosis : recent updates. *Current Fungal Infection Reports*, 8(4) : 302–311, 2014. 89

Tringe S. G. and Rubin E. M. Metagenomics : Dna sequencing of environmental samples. *Nature reviews genetics*, 6(11) : 805–814, 2005. 54

Tsilimigras M. C. and Fodor A. A. Compositional data analysis of the microbiome : fundamentals, tools, and challenges. *Annals of epidemiology*, 26(5) : 330–335, 2016. 69

Tunney M. M., Einarsson G. G., Wei L., Drain M., Klem E. R., Cardwell C., Ennis M., Boucher R. C., Wolfgang M. C., and Elborn J. S. Lung microbiota and bacterial abundance in patients with bronchiectasis when clinically stable and during exacerbation. *American journal of respiratory and critical care medicine*, 187(10) : 1118–1126, 2013. 19

Ueki M. A note on automatic variable selection using smooth-threshold estimating equations. *Biometrika*, pages 1005–1011, 2009. 17

van den Boogaart K. G., Tolosana R., and Bren M. Compositions : compositional data analysis. *R package version*, pages 1–10, 2010. 66

Van der Burgh H. K., Schmidt R., Westeneng H.-J., de Reus M. A., van den Berg L. H., and van den Heuvel M. P. Deep learning predictions of survival based on mri in amyotrophic lateral sclerosis. *NeuroImage : Clinical*, 13, 2017. 17

Vavrek M. J. fossil : palaeoecological and palaeogeographical analysis tools. *Palaeontologia Electronica*, 14(1) : 1T, 2011. R package version 0.3.0. 60

Wadsworth W. D., Argiento R., Guindani M., Galloway-Pena J., Shelburne S. A., and Vannucci M. An integrative bayesian dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC bioinformatics*, 18(1) : 94, 2017. 79, 81

Wang H. J., Zhu Z., and Zhou J. Quantile regression in partially linear varying coefficient models. *The Annals of Statistics*, 37(6B) : 3841–3866, 2009. 33

Wang H. J., Zhou J., and Li Y. Variable selection for censored quantile regression. *Statistica Sinica*, 23(1) : 145–167, 2013. 16

Wang J., Lesko M., Badri M. H., Kapoor B. C., Wu B. G., Li Y., Smaldone G. C., Bonneau R., Kurtz Z. D., Condos R., et al. Lung microbiome and host immune tone in subjects with idiopathic pulmonary fibrosis treated with inhaled interferon-γ. *ERJ Open Research*, 3(3) : 00008–2017, 2017. 19

Wang S., Nan B., Zhu J., and Beer D. G. Doubly penalized Buckley – James method for survival data with high-dimensional covariates. *Biometrics*, 64(1) : 132–140, 2008. 17

Wang Y., Chen T., and Zeng D. Support vector hazards machine : A counting process framework for learning risk scores for censored outcomes. *Journal of Machine Learning Research*, 17 (167) : 1–37, 2016. 17

Wang Z., Wu Y., and Zhao L. A LASSO-type approach to variable selection and estimation for censored regression model. *Chinese Journal of Applied Probability and Statistics*, 26(1) : 66–80, 2010. 16

Wang Z. and Wang C. Buckley-James boosting for survival analysis with high-dimensional biomarker data. *Statistical Applications in Genetics and Molecular Biology*, 9(1) : 24, 2010. 17, 36

Wei R., Wang J., Jia E., Chen T., Ni Y., and Jia W. Gsimp : A gibbs sampler based left-censored missing value imputation approach for metabolomics studies. *PLoS computational biology*, 14 (1) : e1005973, 2018. 51, 134

Weinreich U. and Korsgaard J. Bacterial colonisation of lower airways in health and chronic lung disease. *The clinical respiratory journal*, 2(2) : 116–122, 2008. 18

Weiss S., Xu Z. Z., Peddada S., Amir A., Bittinger K., Gonzalez A., Lozupone C., Zaneveld J. R., Vázquez-Baeza Y., Birmingham A., Hyde E. R., and Knight R. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1) : 27, 2017. 57, 75

Wensing A. M., Calvez V., Günthard H. F., Johnson V. A., Paredes R., Pillay D., Shafer R. W., and Richman D. D. 2017 update of the drug resistance mutations in HIV-1. *Topics in antiviral medicine*, 24(4) : 132, 2017. 16

White J. R., Nagarajan N., and Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLOS computational biology*, 5(4) : e1000352, 2009. 75, 78

Whiteson K. L., Bailey B., Bergkessel M., Conrad D., Delhaes L., Felts B., Harris J. K., Hunter R., Lim Y. W., Maughan H., et al. The upper respiratory tract as a microbial source for pulmonary infections in cystic fibrosis. parallels from island biogeography. *American journal of respiratory and critical care medicine*, 189(11) : 1309–1315, 2014. 19, 20

Whittaker R. H. Vegetation of the siskiyou mountains, oregon and california. *Ecological Monographs*, 30(3) : 279–338, 1960. 61

Wiegand R. E., Rose C. E., and Karon J. M. Comparison of models for analyzing two-group, cross-sectional data with a gaussian outcome subject to a detection limit. *Statistical Methods in Medical Research*, pages 2733–2749, 2016. 17, 34, 36

Willger S. D., Grim S. L., Dolben E. L., Shipunova A., Hampton T. H., Morrison H. G., Filkins L. M., George A., Moulton L. A., Ashare A., et al. Characterization and quantification of the fungal microbiome in serial samples from individuals with cystic fibrosis. *Microbiome*, 2(1) : 40, 2014. 19, 20

Willner D., Haynes M. R., Furlan M., Hanson N., Kirby B., Lim Y. W., Rainey P. B., Schmieder R., Youle M., Conrad D., et al. Case studies of the spatial heterogeneity of dna viruses in the cystic fibrosis lung. *American journal of respiratory cell and molecular biology*, 46(2) : 127–131, 2012. 19

Wittkop L., Günthard H., de Wolf F., Dunn D., Cozzi-Lepri A., de Luca A., Kücherer C., Obel N., von Wyl V., Masquelier B., Stephan C., Torti C., Antinori A., Garcia F., Judd A., Porter K., Thiébaut R., Castro H., van Sighem A., Colin C., Kjaer J., Lundgren J., Paredes R., Pozniak A., Clotet B., philipps A., Pillay D., Chêne G., and study group E.-C. Effects of transmitted drug resistance on virological and immunological response to initial combination antiretroviral therapy for HIV (euro-coord-chain joint project) : a european multicohort study. *The lancet infectious diseases*, 11(5) : 363–371, 2011. 16

Wittkop L., Commenges D., Pellegrin I., Breilh D., Neau D., Lacoste D., Pellegrin J.-L., Chêne G., Dabis F., and Thiébaut R. Alternative methods to analyse the impact of HIV mutations on virological response to antiviral therapy. *BMC Medical Research Methodology*, 8(1) : 68, 2008. 16, 35

Wold S., Martens H., and Wold H. The multivariate calibration problem in chemistry solved by the pls method. In *Matrix pencils*, pages 286–293. Springer, 1983. 15

Xia F., Chen J., Fung W. K., and Li H. A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, 69(4) : 1053–1063, 2013. 80, 81

Xia Y. and Sun J. Hypothesis testing and statistical analysis of microbiome. *Genes & Diseases*, 4(3) : 138–148, 2017. 75

Xue X., Xie X., and Strickler H. D. A censored quantile regression approach for the analysis of time to event data. *Statistical Methods in Medical Research*, 0(0), 2016. 76

Yang Y. and Zou H. A fast unified algorithm for computing group-lasso penalized learning problems. *Statistics and Computing, Accepted*, 2013. 30

Yang Y. and DeGruttola V. Resampling-based multiple testing methods with covariate adjustment : Application to investigation of antiretroviral drug susceptibility. *Biometrics*, 64(2) : 329–336, 2008. 36

Yu G., Gail M. H., Consonni D., Carugno M., Humphrys M., Pesatori A. C., Caporaso N. E., Goedert J. J., Ravel J., and Landi M. T. Characterizing human lung tissue microbiota and its relationship to epidemiological and clinical features. *Genome biology*, 17(1) : 163, 2016. 19

Yuan L., Liu J., and Ye J. Efficient methods for overlapping group lasso. In *Advances in Neural Information Processing Systems*, pages 352–360, 2011. 83

Yuan M. and Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 68 : 49–67, 2006. 28, 30

Zakharkina T., Heinzel E., Koczulla R. A., Greulich T., Rentz K., Pauling J. K., Baumbach J., Herrmann M., Grünewald C., Dienemann H., et al. Analysis of the airway microbiota of healthy individuals and patients with chronic obstructive pulmonary disease by t-rflp and clone sequencing. *PLOS ONE*, 8(7) : e68302, 2013. 18

Zemanick E. T., Harris J. K., Wagner B. D., Robertson C. E., Sagel S. D., Stevens M. J., Accurso F. J., and Laguna T. A. Inflammation and airway microbiota during cystic fibrosis pulmonary exacerbations. *PLOS ONE*, 8(4) : e62917, 2013. 19

Zhan X., Tong X., Zhao N., Maity A., Wu M., and Chen J. A small-sample multivariate kernel machine test for microbiome association studies. *Genetic Epidemiology*, 41(5) : 210–220, 2017. 73

Zhang Q., Cox M., Liang Z., Brinkmann F., Cardenas P. A., Duff R., Bhavsar P., Cookson W., Moffatt M., and Chung K. F. Airway microbiota in severe asthma and relationship to asthma severity and phenotypes. *PLOS ONE*, 11(4) : e0152724, 2016. 18

Zhang Q., Abel H., Wells A., Lenzini P., Gomez F., Province M. A., Templeton A. A., Weinstock G. M., Salzman N. H., and Borecki I. B. Selection of models for the analysis of risk-factor trees : leveraging biological knowledge to mine large sets of risk factors with application to microbiome data. *Bioinformatics*, 31(10) : 1607–1613, 2015. 86, 87

Zhao J., Schloss P. D., Kalikin L. M., Carmody L. A., Foster B. K., Petrosino J. F., Cavalcoli J. D., VanDevanter D. R., Murray S., Li J. Z., et al. Decade-long bacterial community dynamics in cystic fibrosis airways. *Proceedings of the National Academy of Sciences*, 109 (15) : 5809–5814, 2012. 19

Zhao S. D., Lee D., and Li Y. The Dantzig selector for censored linear regression models. *Statistica Sinica*, 24(1) : 251–268, 2014. 17

Zhou N. and Zhu J. Group variable selection via a hierarchical lasso and its oracle property. *arXiv preprint arXiv :1006.2871*, 2010. 85

Zou H. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476) : 1418–1429, 2006. 27

Zou H. and Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67 : 301–320, 2005. 27

Activités liées à la thèse

Publications [1]

Articles dans des revues internationales avec comité de lecture :

► P. Soret, M. Avalos, L. Wittkop, D. Commenges, R. Thiébaut, Lasso regularization for left-censored Gaussian outcome and high-dimensional predictors, *BMC Medical Research methodology*, 18 :159. DOI : 10.1186/s12874-018-0609-4

► P. Soret*, L.E. Vandenborght*, F. Francis, N. Coron, R. Enaud, The Mucofong Investigation Group, M. Avalos, T. Schaeverbeke, P. Berger,M. Fayon, R. Thiébaut L. Delhaes, Respiratory mycobiome and suggestion of inter-kingdom network during acute pulmonary exacerbation in cystic fibrosis. *Accepté dans Scientific reports.*

Communications orales à des conférences nationales avec comité de lecture :

► P. Soret†, M. Avalos, C. Ong, R. Thiébaut, High-dimensional compositional microbiota data : state-of-the-art of methods and software implementations. In : *GDR, "Statistique et Santé*, Bordeaux, France, 2017.

► P. Soret†, M. Avalos, L. Wittkop, D. Commenges, R. Thiébaut. Lasso pour données censurées à gauche : une comparaison par simulation d'algorithmes proposés dans la littérature. In : $47^{ème}$ Journées de la Statistique, Lille, France, 2015.

1.

Légende

* Contribution égale

† Présentateur

► Censure due à une limite de détection en grande dimension, VIH

► Données compositionnelles de grande dimension, Microbiome

Communications écrites (poster) à des conférences internationales avec comité de lecture :

▶ P. Soret, M. Avalos†, L. Delhaes, R. Thiébaut, A simulation framework of high-dimensional phylogenetic microbiota data. In : *29*th International Biometric Conference, Barcelone, Espagne, Juillet, 2018

Communications écrites (poster) à des conférences nationales ou internationales sans comité de lecture :

▶ L.-E. Vandenborght†, P. Soret, N. Coron, R. Enaud, F. Francis, M. Avalos, T. Schaeverbeke, P. Berger, M. Fayon, R. Thiébaut, L. Delhaes, Caractérisation du mycobiote et du microbiote respiratoire dans la mucoviscidose : importance du dialogue interrègnes pendant un phénomène d'exacerbation, Congrès de la Société Française de Mycologie Médicale et la Société Française de Parasitologie, Nice, France, 2018.

Bourses obtenues

▶ Thèse financée par une bourse doctorale ministérielle

▶ Bourse de mobilité de l'Idex (Initiative d'excellence) 2017 de l'Université de Bordeaux et Bourse de mobilité Zellidja 2017. Séjour de recherche du 1er février 2017 au 30 Avril 2017 au sein du laboratoire de *machine learning* Data61 (CSIRO), Canberra, Australie, à l'invitation de Cheng Soon Ong.

Logiciels

▶ Implémentation des méthodes pour données censurées par limite de quantification `https://github.com/psBiostat/left-censored-Lasso`

▶ Code des analyses statistiques réalisées sur l'étude MucoFong et étude de simulation `https://github.com/psBiostat/MucoFong-study.git`

Autres activités réalisées

Encadrements de stage de recherche

► Co-encadrement d'un stage de Master 2 Fouille de données, Univ de Compiègne (Hao Ren), 2016. Évaluation des différentes implémentations liées au GroupLasso

► Encadrement d'un stage de M1 Santé publique, option biostatistique, Univ de Bordeaux (Maéva Kyheng), 2015. Étude de l'évolution de la charge viral chez des patients issu d'un essai pour un vaccin thérapeutique contre le VIH (Dalia-1) : Comparaison de deux approches statistiques

Activités liées à l'enseignement

► Participation à la mise en place du projet "Begin R" : séquence pédagogique autour de la prise en main de R, de la manipulation de données et de la mise en œuvre des principales approches de statistiques descriptives et inférentielles. Ce projet a été financé par l'Idex (Université de Bordeaux) sur la période 2014-2015.

► Master1 Santé Publique : enseignement sur la régression linéaire (14h)

► Master1 Santé Publique : supervision d'un projet tutoré

► Master1 Santé Publique : Référente de trois étudiants à distance

► Intervention au sein du Workshop "Big Data in clinical research" (4h)

Responsabilités administratives et implications dans l'organisation d'événements

► 2015-2016 : Présidente de l'association des doctorants de l'EDSP2

► 2014-2017 : Membre du comité des doctorants de l'EDSP2 et Membre de l'association des doctorants de l'EDSP2

► 26 Mai 2016 : Organisation des journées de l'EDSP2

► 11 Octobre 2016 : Participation à la fête de la science, INRIA, Bordeaux

► 28 Mai 2015 : Aide à l'organisation des journées de l'EDSP2

Régression pénalisée de type Lasso pour l'analyse de données biologiques de grande dimension : application à la charge virale du VIH censurée par une limite de quantification et aux données compositionnelles du microbiote

Résumé : Dans les études cliniques et grâce aux progrès technologiques, la quantité d'informations recueillies chez un même patient ne cesse de croître conduisant à des situations où le nombre de variables explicatives est plus important que le nombre d'individus. La méthode Lasso s'est montrée appropriée face aux problèmes de sur-ajustement rencontrés en grande dimension. Cette thèse est consacrée à l'application et au développement des régressions pénalisées de type Lasso pour des données cliniques présentant des structures particulières.

Premièrement, chez des patients atteints du virus de l'immunodéficience humaine des mutations dans les gènes du virus peuvent être liées au développement de résistances à tel ou tel traitement. La prédiction de la charge virale à partir des mutations (potentiellement grand) permet d'orienter le choix des traitements. En dessous d'un seuil, la charge virale est indétectable, on parle de données censurées à gauche. Nous proposons deux nouvelles approches Lasso de l'algorithme itératif Buckley-James consistant à imputer les valeurs censurées par une espérance conditionnelles. En inversant la réponse, on peut se ramener à un problème de censure à droite, pour laquelle des estimations non-paramétriques de l'espérance conditionnelle ont été proposées en analyse de survie. En deuxième, nous proposons une estimation paramétrique qui repose sur une hypothèse Gaussienne.

Deuxièmement, nous nous intéressons au rôle du microbiote dans la détérioration de la santé respiratoire. Les données du microbiote sont sous forme d'abondances relatives (proportion de chaque espèce par individu, dites données de compositions) et elles présentent une structure phylogénétique. Nous avons dressé un état de l'art des méthodes d'analyses statistiques de données du microbiote. En raison de la nouveauté, peu de recommandations existent sur l'applicabilité et l'efficacité des méthodes proposées. Une étude de simulation nous a permis de comparer la capacité de sélection des méthodes de pénalisation proposées spécifiquement pour ce type de données. Puis nous appliquons ces recherches à l'analyse de l'association entre les bactéries/champignons et le déclin de la fonction pulmonaire chez des patients atteints de la mucoviscidose du projet MucoFong.

Mots clés : Apprentissage statistique, Sélection de variables, Prédiction, Algorithme, Limite de détection, Données structurées, Mucoviscidose

Penalized Lasso regression for the analysis of high-dimensional biological data: application to HIV viral load censored by a limit of quantification and to microbiota compositional data

Abstract: In clinical studies and thanks to technological advances, the amount of information collected from the same patient is constantly increasing, leading to situations where the number of explanatory variables is greater than the number of individuals. The Lasso method has proven to be appropriate in the face of over-adjustment problems encountered in high-dimensional settings. This thesis is devoted to the application and development of penalized Lasso-type regressions for clinical data with particular structures.

First, in patients with human immunodeficiency virus, mutations in the genes of the virus may be related to the development of resistance to particular treatments. Viral load prediction based on (potentially large number of) mutations helps to guide the choice of treatments. Below a threshold, the viral load is undetectable, we are talking about left-censored data. We propose two new Lasso approaches to the iterative Buckley-James algorithm consisting in imputing censored values with a conditional expectation. By reversing the answer, we can reduce this to a problem of right-censorship, for which non-parametric estimates of conditional expectation have been proposed in survival analysis. Second, we propose a parametric estimate based on a Gaussian hypothesis.

Secondly, we are interested in the role of the microbiota in the deterioration of respiratory health. The microbiota data are in the form of relative abundances (proportion of each species per individual, called compositional data) and they have a phylogenetic structure. We have established state of the art methods of statistical analysis of microbiota data. Due to the novelty, few recommendations exist on the applicability and effectiveness of the proposed methods. A simulation study allowed us to compare the selection capacity of penalization methods proposed specifically for this type of data. Then we apply this research to the analysis of the association between bacteria / fungi and the decline in lung function in cystic fibrosis patients of the MucoFong project.

Key words: Statistical machine learning, Variable selection, Prediction, Algorithm, Limit of detection, Structured data, Cystic fibrosis

Discipline : Santé publique – option : Biostatistiques
Laboratoire : Unité INSERM U1219, Bordeaux Population Health Center - INRIA - Université de Bordeaux, 146 rue Léo Saignat 33000 Bordeaux, FRANCE