Table des matières

Liste des tableaux

Table des figures

Liste des abréviations

ACP Analyse en Composantes Principales

AIC Akaike Information Criterion

ANRT Association Nationale de la Recherche et de la Technologie

ARB Absolute Relative Bias

BIC Bayesian Information Criterion

CD Covariates dependent

CIC Correlation Information Criterion

CIFRE Convention Industrielle de Formation par la Recherche

CNAM Conservatoire National des Arts et Metiers

CV Cross Validation

DOM Distribution Of Missingness

EPB Expected Predictive Bias

FNIH Foundation for the National Institutes of Health

GCV Generalized Cross Validation

GEE Generelized Estimating Equations

GLIC Generalized Longitudinal Information Criterion

GLM Generelized Linear Models

IPW Inverse Probabibility Weighting

IRIS Institut de Recherches Internationales Servier

IRM Imagerie par Résonance Magnétique

LARS Least Angle Regression

LASSO Least Absolute Shrinkage and Selection Operator

LOCF Last Observation Carried Forward

LOD Limit Of Detection

LOO Leave One Out

LOQ Limit Of Quantification

LQA Local Quadratic Approximation

MAR Missing At Random

MCAR Missing Completely At Random

MCMC Monte Carlo par Chaînes de Markov

MI-LASSO Multiple Imputation Least Absolute Shrinkage and Selection Operator

MI-PGEE Multiple Imputation Penalized Generalized Estimating Equations

MLICC Missing Longitudinal Information Criterion for Correlation

MLIC Missing Longitudinal Information Criterion

MNAR Missing Not At Random

MNR Modified Newton Raphson

MOAKS MRI OsteoArthritis Knee Score

NIAMS National Institute of Arthritis and Musculoskeletal and Skin Diseases

NIA National Institute on Aging

OAI Initiative OsteoArthritis

PA Population-Average

PGEE Penalized Generalized Estimating Equations

PMSE Predictive Mean Squared Error

QGCV Quasi-Generalized Cross Validation

QIC Quasi-likelihood under Independance model Criterion

RSS Residuals Sum of Squares

SCAD Smoothly Clipped Absolute Deviation

SEKOIA Strontium ranelate Efficacy in Knee OsteoarthItis triAL

SS Subject-Specific

WGEE Weighted Generalized Estimating Equations

WORMS Whole-Organ Magnetic Resonance Imaging Score

WRSS Weighted Residuals Sum of Squares

Introduction

Un biomarqueur est une caractéristique mesurable objectivement qui représente un indicateur des processus biologiques. Il peut prendre différentes formes : données de protéomique, de génétique ou de transcriptomique comme des données d'imagerie médicale. Les biomarqueurs présentent un intérêt dans la recherche de candidats médicaments pour aider à caractériser les pathologies ; ce qui explique qu'ils soient de plus en plus mesurés dans les études cliniques. Par ailleurs, les nouvelles technologies de mesure permettent d'obtenir des bases de données composées de nombreuses variables. L'objectif est donc de dissocier le signal porté par les biomarqueurs informatifs, des biomarqueurs non informatifs qui sont des variables de bruit.

Les modèles linéaires généralisés (GLM) sont souvent utilisés pour analyser ces relations et modéliser le lien entre la réponse clinique et les biomarqueurs. Cependant, la multitude de marqueurs mesurés pose le problème de l'interprétation. Il est donc nécessaire de choisir les biomarqueurs à intégrer dans le modèle. Une première approche de sélection consiste à calculer un critère de qualité, comme le BIC ou l'AIC pour chaque modèle plausible. Malheureusement, dés que le nombre de variables dépasse $p = 30$, ces méthodes ne sont plus applicables car explorer l'ensemble des 2^p modèles est un problème combinatoire. La méthode de sélection par régularisation, comme par exemple la méthode LASSO est une alternative capable de sélectionner un sous-groupe de variables d'intérêt sans explorer tous les modèles.

Lorsque le suivi longitudinal du patient est observé, le critère clinique d'intérêt et certains biomarqueurs sont disponibles à plusieurs instants. La mise en relation de ces quantités, dans ce contexte de mesures répétées dans le temps, nécessite des méthodes particulières. Nous pouvons raisonnablement penser que les observations issues du même

sujet sont plus semblables que les observations inter-sujets. De ce fait, les corrélations intra-patient dues au temps doivent être intégrées au modèle. Les Generalized Estimating Equations (GEE) sont une extension des GLM pour les données corrélées. Cette méthode marginale propose d'estimer les coefficients de régression par des équations généralisées où la matrice de corrélation dite de *travail* est fixée par l'utilisateur. On évite ainsi la spécification de la vraisemblance jointe en utilisant uniquement des hypothèses sur les deux premiers moments de la réponse.

La plupart des critères de sélection de modèles repose sur deux quantités : une mesure d'ajustement du modèle comme la vraisemblance ou la somme des résidus au carré (RSS) et une pénalité discrète sur la complexité du modèle comme le nombre de coefficients à estimer, ou plus généralement le degré de liberté du modèle. Dans un contexte de données corrélées, ces quantités doivent être adaptées. La vraisemblance se transforme en quasi-vraisemblance ; la RSS en somme pondérée par les corrélations entre observations et les degrés de liberté prennent en compte les relations intra-patients. De la même manière, il est possible d'étendre les méthodes de sélection par régularisation en pénalisant directement les GEE (PGEE). De nombreuses méthodes ont ainsi été proposées pour les GEE comme le LASSO, les pénalités Ridge ou plus généralement Bridge, ainsi que les pénalités combinées de type Elastic-Net.

Ces méthodes supposent que les covariables et la réponse sont observées aux mêmes instants pour tous les patients. Malheureusement, les données réelles présentent souvent des données manquantes qui peuvent apparaître sous différentes formes. Une étude longitudinale peut souffrir d'attrition : certains patients sortent de l'étude. Il est possible d'avoir des visites manquantes intermittentes : un patient manque une visite mais se présente à la visite suivante. Une autre problématique est celle des données non renseignées : lorsqu'un patient effectue une visite, il peut ne pas répondre à tous les items d'un questionnaire. Par ailleurs, certaines covariables peuvent souffrir de problèmes de mesure. Ce dernier cas soulève la question des covariables soumises à un seuil de détection. Dans la recherche de biomarqueurs, cette problématique est souvent rencontrée. Se pose alors la question de l'intégration de ces données manquantes aux analyses sans biaiser les résultats.

L'objectif de ces travaux de recherche est double. Dans un premier, temps le but est d'étudier l'impact des données manquantes sur l'estimateur obtenu par GEE. Nous proposons des solutions non biaisées pour l'analyse du lien entre biomarqueurs et réponse. Dans un contexte clinique où de nombreux biomarqueurs sont mesurés, la problématique des variables à seuil est récurrente. Nous proposons une méthodologie permettant d'imputer les variables explicatives soumises à un seuil de détection. Dans un second temps, l'objectif est de développer de nouvelles méthodes statistiques nécessaires pour intégrer les données manquantes dans la sélection de variables pour données longitudinales.

La thèse est composée de cinq parties : les deux premières parties proposent un état de l'art sur les méthodes de sélection de modèles et de variables pour les GEE. Dans une première partie, nous considérons la problématique de l'estimation des paramètres de la moyenne dans un contexte longitudinal. La méthode des GEE est introduite. La généralisation de critères de qualité de modèle comme l'AIC, le BIC ou les C_p de Mallows aux GEE est présentée ainsi que les critères de sélection pour la matrice de corrélation de travail. La deuxième partie traite le problème de sélection de variables dans le cadre des régressions pénalisées. Les pénalités Bridge, LASSO, Ridge et Group-LASSO sont présentées. Leur extension aux GEE nécessite d'affiner les algorithmes existants. La méthode d'approximation quadratique locale est présentée ainsi que des critères pour la sélection du paramètre de régularisation.

La troisième partie de ce travail de thèse propose d'étudier l'impact de données manquantes sur les GEE. Les méthodes de pondération sont proposées pour prendre en compte l'attrition des études longitudinales. Une étude par simulations est proposée afin d'explorer l'effet de visites manquantes intermittentes sur l'estimateur obtenu par GEE. Nous nous intéressons ensuite au cas particulier des variables soumises à un seuil de détection. Dans ce contexte, les données non observées sont non aléatoires et donc non-ignorables. Nous proposons une fonction d'imputation pour ces données particulières à l'aide du modèle Tobit pour l'algorithme d'imputation multivariée par équations en chaîne (mice). Pour l'inférence, les règles de Rubin peuvent être utilisées afin de prendre en compte la variabilité

induite par l'imputation. Les bonnes propriétés de la méthode sont vérifiées par simulations.

La quatrième partie de ce travail de thèse propose une nouvelle méthode de sélection de variables pour données corrélées qui intègre les données manquantes : le Multiple Imputation Penalized Generalized Estimating Equations (MI-PGEE). Cette méthode est une extension du MI-LASSO, opérateur développé pour une réponse continue indépendante. Le MI-PGEE utilise des GEE pénalisées par une pénalité Ridge et des poids qui sont communs à l'ensemble des coefficients de régression estimés de la même variable sur les échantillons imputés. Nous présentons un nouveau critère de type BIC pour le choix du paramètre de régularisation. Notre méthode fournit une sélection consistante sur l'ensemble des imputations ; ce qui en fait une méthode de sélection pour données longitudinales capable d'intégrer les données manquantes et les corrélations intra-sujets.

Ces travaux de thèse ont été motivés par des recherches sur l'arthrose du genou. Nous proposons cette application en cinquième et dernière partie de thèse. Nous étudions l'histoire naturelle de la maladie au travers du critère principal qu'est la largeur de l'espace articulaire mesurée à plusieurs temps. L'objectif est de sélectionner le sous-groupe de biomarqueurs qui explique au mieux les différences de largeur de l'espace articulaire entre patients au cours du temps. Deux bases sont étudiées, le groupe placebo d'une étude clinique Servier : Strontium ranelate Efficacy in Knee OsteoarthrItis triAl (SEKOIA) et le projet Foundation for the National Institutes of Health (FNIH) de l'étude OsteoArthritis Initiative (OAI).

Notations et rappels

Nous utiliserons tout au long de ce document les notations usuelles de la littérature. Les matrices sont désignées par des lettres en majuscule à caractère gras (par exemple $\mathbf{X}$), les vecteurs par des lettres en minuscule à caractère gras (par exemple $\mathbf{y}$). Les éléments qui composent ces matrices et vecteurs sont désignés par des lettres minuscule en italique avec les indices appropriés (par exemple $x_{i,j}$ est un élément de $\mathbf{X}$). La matrice identité de taille $n \times n$ est définie par $\mathbf{I}_n$, le vecteur colonne dont les composantes sont égales à un par $\mathbf{1}$, la transposée d'une matrice^T et son inverse^{-1}. Appliqué à une matrice carrée, l'opérateur diagonale noté $\mathbf{diag}(.)$ prend les éléments de la diagonale de la matrice et les stocke dans un vecteur colonne ; lorsqu'il est appliqué à un vecteur, l'opérateur $\mathbf{diag}(.)$ stocke les éléments du vecteur sur la diagonale d'une matrice. L'opérateur trace noté $\mathbf{tr}(.)$ appliqué à une matrice carré calcule la somme des éléments de sa diagonale. Le produit standard entre deux matrices $\mathbf{X} \times \mathbf{Y}$ est noté $\mathbf{XY}$. L'espérance et la variance d'une variable aléatoire u se notent $\mathbb{E}(u)$ et $\mathbb{V}(u)$. Pour un vecteur aléatoire $\mathbf{u}$, la matrice de variance covariance se note $\mathbb{C}ov(\mathbf{u})$. Soit $\mathbf{v}$, un vecteur de taille p contenant q coefficients non nuls. La norme L_0 est le nombre de coefficients non nul d'un vecteur, $||\mathbf{v}||_0 = \sum_{j=1}^{p} \mathbb{1}_{|v_j|>0} = q$, pour $\gamma > 0$, la norme L_γ représente la quantité suivante : $||\mathbf{v}||_\gamma = \left(\sum_{j=1}^{p} |v_j|^\gamma\right)^{1/\gamma}$.

Les analyses effectuées portent sur les mesures répétées de K patients observés à T_i temps pour $i \in \{1, \ldots, K\}$ soit $N = \sum_{i=1}^{K} T_i$ observations au total. Les bases de données sont composées d'une réponse clinique $\mathbf{y}$, vecteur de taille N, et d'une matrice $\mathbf{X}$ de taille $N \times p$ composée de p covariables observées. Nous notons $y_{i,t}$ et $\mathbf{x}_{i,t}$ la réponse et le vecteur de covariables observés au temps $t \in \{1, \ldots, T_i\}$ pour l'individu i. Les données sont structurées en K matrices $\mathbf{X}_i$ de taille $T_i \times p$ pour les covariables et K vecteurs $\mathbf{y}_i$ de taille T_i pour la

réponse que l'on concatène de la manière suivante :

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_i \\ \vdots \\ \mathbf{y}_K \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_i \\ \vdots \\ \mathbf{X}_K \end{bmatrix}$$

Où pour chaque individu $i \in \{1, \ldots, K\}$:

$$\mathbf{y_i} = \begin{pmatrix} y_{i,1} \\ \vdots \\ y_{i,t} \\ \vdots \\ y_{i,T_i} \end{pmatrix} \qquad \mathbf{X_i} = \begin{pmatrix} x_{i,1,1} & \cdots & x_{i,j,1} & \cdots & x_{i,p,1} \\ \vdots & \ddots & & & \vdots \\ x_{i,1,t} & & x_{i,j,t} & & x_{i,p,t} \\ \vdots & & & \ddots & \vdots \\ x_{i,1,T_i} & \cdots & x_{i,j,T_i} & \cdots & x_{i,p,T_i} \end{pmatrix}$$

Lorsque la variable réponse $\mathbf{y}$ est continue, on peut supposer une relation linéaire entre les covariables et cette dernière. On définit un modèle linéaire de la forme :

$$\mathbb{E}[y_{i,t}] = \mu_{i,t} = \mathbf{x}_{i,t}^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i,1,t} + \beta_2 x_{i,2,t} + \ldots + \beta_p x_{i,p,t} \tag{1}$$

où $y_{i,t} \sim \mathcal{N}(\mu_{i,t}, \sigma^2)$ sont des réalisations indépendantes. Le paramètre β_0 représente la valeur moyenne de $y_{i,t}$ lorsque toutes les autres variables sont mises à zéro. L'objectif principal de ce type de modèle est de déterminer comment la moyenne de la réponse évolue en fonction des prédicteurs à l'aide des coefficients de régression correspondant $\beta_1, \beta_2, \ldots, \beta_p$. Le livre de Rao and Toutenburg [1995] est une bonne introduction aux modèles linéaires et à leur propriétés géométriques. Pour une interprétation biostatistique, les lecteurs peuvent se référer au Chapitre 4 du livre de Vittinghoff et al. [2011].

Chapitre 1

La sélection de modèles dans les Generalized Estimating Equations : état de l'art

En présence de données longitudinales (i.e. répétées dans le temps), les corrélations intra-sujets doivent être prises en compte car nous pouvons raisonnablement penser que les observations issues du même sujet sont plus semblables que les observations inter-sujets. Ignorer cet aspect des données conduit à des estimations incorrectes de la variance ce qui peut conduire à des inférences fallacieuses (Fitzmaurice et al. [2012]).

Deux grandes catégories de méthodes permettent de prendre en compte cet aspect des observations : les méthodes dites spécifiques au sujet (SS) et les méthodes sur la moyenne de la population (PA). Les modèles spécifiques au sujet proposent de modéliser l'hétérogénéité entre individus, alors que les modèles marginaux, proposent de s'intéresser aux effets moyens sur la population. La corrélation intra-sujet est décrite, mais ses sources ne sont pas expliquées. Notre objectif étant d'étudier l'impact de différentes covariables sur la réponse, les modèles marginaux sont donc appropriés pour évaluer le lien entre prédicteurs et moyenne de la population.

Les Generalized Estimating Equations sont une méthode marginale qui permet l'analyse de données longitudinales où la réponse peut être de différents types (i.e. continue, binaire, de comptage...). Cette méthode est souvent présentée comme une extension des modèles linéaires généralisés (GLM) aux données corrélées. La première section de ce chapitre pré-

sente la famille exponentielle, les GLM et ses extensions afin d'introduire la nouvelle classe d'équations d'estimation que sont les Generalized Estimating Equations.

En présence de multiples covariables mesurées, la modélisation de la fonction moyenne doit faire l'objet d'une réflexion. En pratique, il est courant d'inclure uniquement un sous-ensemble de variables importantes dans le modèle afin d'améliorer la prédictibilité et la parcimonie du modèle. Dans la littérature, on trouve de nombreux critères de sélection de modèles faisant intervenir la notion de vraisemblance, de degré de liberté et de mesure d'ajustement, quantités ambiguës dans le cadre des GEE. La deuxième section de ce chapitre propose une liste non-exhaustive de critères adaptés à la bonne spécification du modèle pour les GEE.

1.1 Generalized Estimating Equations

Une première généralisation du modèle linéaire présentée dans l'équation (1) a été proposée par Nelder and Baker [1972]. Les GLM permettent d'étendre les modèles linéaires à une plus grande classe de variables, mais la spécification de la vraisemblance jointe de $\mathbf{y}$ reste un problème, surtout en présence de données corrélées. Les GEE permettent d'éviter cette complication en ne spécifiant que les deux premiers moments de la réponse.

1.1.1 GLM et famille exponentielle

Les GLM proposent des relations plus complexes que le modèle linéaire de l'équation (1) entre covariables mesurées et réponse afin de convenir à des variables de type binaire ou de comptage. Pour plus de flexibilité, cette méthode suppose que l'espérance est reliée aux covariables par une composante linéaire $\mathbf{x}_{i,t}^T\boldsymbol{\beta}$ grâce à la fonction de lien $g(.)$:

$$\mathbb{E}(y_{i,t}) = \mu_{i,t} = g^{-1}(\mathbf{x}_{i,t}^T\boldsymbol{\beta}) \tag{1.1}$$

Afin de permettre à la réponse de prendre différentes formes, l'hypothèse de normalité est relâchée au profit de l'appartenance à la famille exponentielle. La densité de $y_{i,t}$ sera alors de la forme :

$$f(y_{i,t}, \theta_{i,t}, \phi) = \exp\left\{\frac{y_{i,t}\theta_{i,t} - a(\theta_{i,t})}{\phi} + c(y_{i,t}, \phi)\right\} \tag{1.2}$$

pour tout individu $i \in \{1, \ldots, K\}$ à tout temps $t \in \{1, \ldots, T_i\}$ où $a(.)$ et $c(.,.)$ sont des fonctions connues. On montre que :

$$\begin{cases} \mathbb{E}(y_{i,t}) = a'(\theta_{i,t}) \\ \mathbb{V}(y_{i,t}) = a''(\theta_{i,t})\phi \end{cases} \tag{1.3}$$

où ϕ est le paramètre de dispersion que l'on considère comme un paramètre de nuisance et $V(\mu_{i,t}) = a''(\theta_{i,t})$ est appelée fonction de variance. Le paramètre naturel $\theta_{i,t}$ dépend du prédicteur linéaire $\mathbf{x}_{i,t}^T\boldsymbol{\beta}$ à travers la fonction $h(.) : \theta_{i,t} = h(\boldsymbol{\beta})$. La plupart des distributions peuvent être écrites sous cette forme, parmi lesquelles les classiques Gaussienne, Bernoulli et Poisson :

Distribution $f(y;\theta)$	Fonction de lien $g(.)$	Variance $V(.)$	Dispersion ϕ	Naturel θ
Gaussien (μ, σ^2)	identité	1	σ^2	μ
Bernoulli (p)	logit	$p(1-p)$	1	$\log(\frac{p}{1-p})$
Poisson (λ)	logarithme	λ	1	$\log(\lambda)$

TABLE 1.1 – Fonctions de lien et de variance, paramètres de dispersion et naturels pour trois distributions classiques de la famille exponentielle

La vraisemblance partielle $L(\theta_{i,t}, \phi|y_{i,t})$ de l'individu i au temps t est simplement une re-formulation de la densité où l'on considère la réponse $y_{i,t}$ connue et le paramètre $\theta_{i,t}$ inconnu, sa log-vraisemblance partielle s'écrit :

$$l(\theta_{i,t}, \phi|y_{i,t}) = \frac{y_{i,t}\theta_{i,t} - a(\theta_{i,t})}{\phi} + c(y_{i,t}, \phi) \tag{1.4}$$

En supposant que l'ensemble des observations du vecteur $\mathbf{y}$ sont indépendantes, nous pouvons définir la log-vraisemblance complète comme la somme des log-vraisemblances partielles. En notant $\boldsymbol{\theta}$ le vecteur des paramètres $\theta_{i,t}$:

$$l(\boldsymbol{\theta}, \phi|\mathbf{y}) = \sum_{i,t}(l(\theta_{i,t}, \phi|y_{i,t})) = \sum_{i,t}\left(\frac{y_{i,t}\theta_{i,t} - a(\theta_{i,t})}{\phi} + c(y_{i,t}, \phi)\right) \tag{1.5}$$

La fonction de score $G_{i,t}$ pour l'individu i au temps t, dérivée partielle de sa log-vraisemblance, s'écrit :

$$G_{i,t} = \frac{\partial}{\partial \theta_{i,t}}\Big[l(\theta_{i,t}, \phi, y_{i,t})\Big] = \frac{y_{i,t} - \mu_{i,t}}{\phi} \tag{1.6}$$

Elle vérifie les propriétés suivantes :

$$\begin{cases} \mathbb{E}(G_{i,t}) = 0 \\ \mathbb{V}(G_{i,t}) = \mathbb{E}(G_{i,t}^2) = -\mathbb{E}(\frac{\partial G_{i,t}}{\partial \mu_{i,t}}) \end{cases} \tag{1.7}$$

Nous pouvons utiliser la règle de la chaîne pour obtenir des équations d'estimation pour le vecteur $\boldsymbol{\beta}$:

$$\frac{\partial}{\boldsymbol{\beta}} = (\frac{\partial}{\partial \theta_{i,t}})(\frac{\partial \theta_{i,t}}{\partial \mu_{i,t}})(\frac{\partial \mu_{i,t}}{\partial \boldsymbol{\beta}}) \tag{1.8}$$

Le maximum de vraisemblance $\hat{\boldsymbol{\beta}}$ est alors solution de :

$$\frac{\partial l(\boldsymbol{\beta}, \phi|\mathbf{y})}{\boldsymbol{\beta}} = \sum_{i,t} \mathbf{diag}\left(\frac{\partial \mu_{i,t}}{\partial \boldsymbol{\beta}}\right) \frac{y_{i,t} - \mu_{i,t}}{\mathbb{V}(y_{i,t})} = \sum_{i} \mathbf{D}_i^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = 0 \tag{1.9}$$

Ou $\mathbf{D}_i$ est la matrice des dérivées composée des éléments $\partial \mu_{i,t}/\partial \boldsymbol{\beta}$ pour $t \in \{1, \ldots, T_i\}$ et $\mathbf{V}_i$ est la matrice diagonale des variances marginales $\mathbb{V}(y_{i,t})$. Ces équations peuvent être non linéaires en $\boldsymbol{\beta}$. Les méthodes itératives faisant intervenir le Hessien pour la méthode de Newton-Raphson ou la matrice d'information pour la méthode de Fisher peuvent alors être utilisées.

Les GLM supposent que la fonction de variance $V(.)$, le paramètre de dispersion ϕ et la fonction de lien $g(.)$ sont correctement spécifiés. Une hypothèse forte sur laquelle reposent les équations d'estimation (1.9) est l'indépendance des observations. Le déroulement des équations dépend du fait que les observations sont indépendantes et identiquement distribuées (*iid*). Il existe cependant beaucoup de situations pour lesquelles cette hypothèse n'est pas respectée, à commencer par le cas des mesures répétées au cours du temps pour chaque patient. Dans un contexte longitudinal, les corrélations intra-patient ne sont pas

ignorables et passer de la vraisemblance individuelle à la vraisemblance complète n'est plus une évidence.

1.1.2 Quasi-Vraisemblance et Pseudo-Vraisemblance

L'enjeu principal avec les approches utilisant la vraisemblance jointe complète de la réponse $\mathbf{y}$ est la complexité algorithmique. Lorsque les données ne correspondent pas exactement à une distribution connue (Gaussienne, Binomiale...), établir sa fonction de vraisemblance devient complexe. Pour ces raisons, lorsque l'on s'intéresse plus particulièrement au vecteur de paramètres $\boldsymbol{\beta}$ modélisant l'espérance marginale, deux méthodes initiatrices des GEE ont été proposées : la quasi-vraisemblance (Wedderburn [1974]; McCullagh [1983]) et la pseudo-vraisemblance (Arnold and Strauss [1991]; Geys et al. [1999]).

Quasi-Vraisemblance La méthode de Quasi-Vraisemblance utilise une fonction paramétrique des covariables pour modéliser l'espérance, la variance étant supposée être une fonction de la moyenne. Cette méthode permet d'utiliser des fonctions qui ne font pas partie de la famille exponentielle, dans le cas contraire, elle coïncide avec la méthode de la vraisemblance. La méthode repose sur le fait que la variable aléatoire $u_{i,t} = (y_{i,t} - \mu_{i,t})/(\phi V(\mu_{i,t}))$ partage un certain nombre de propriétés avec la dérivée de la log-vraisemblance :

$$\begin{cases} \mathbb{E}(u_{i,t}) = 0 \\ \mathbb{V}(u_{i,t}) = \mathbb{E}(u_{i,t}^2) = -\mathbb{E}\left[\dfrac{\partial u_{i,t}}{\partial \mu_{i,t}}\right] \end{cases} \tag{1.10}$$

Il semble raisonnable que, si elle existe, l'intégrale

$$Q(\mu_{i,t}|\phi, y_{i,t}) = \int_{y_{i,t}}^{\mu_{i,t}} \frac{y_{i,t} - v}{\phi V(v)} dv \tag{1.11}$$

se comporte comme une fonction de log-vraisemblance pour $\mu_{i,t}$. Si l'on suppose que toutes les observations sont indépendantes nous pouvons écrire :

$$Q(\boldsymbol{\mu}|\phi, \mathbf{y}) = \sum_{i=1}^{K} \sum_{t=1}^{T_i} Q(\mu_{i,t}|\phi, y_{i,t}) \tag{1.12}$$

Cette quantité, appelée quasi-vraisemblance ou log-quasi-vraisemblance, est l'analogue de la log-vraisemblance usuelle. Le paramètre de nuisance ϕ n'intervient que de façon multiplicative, il ne perturbe pas l'estimation de $\mu_{i,t}$ ni de $\boldsymbol{\beta}$ (Davis [2002]). Pour plus de simplicité, nous noterons cette quantité $Q(\boldsymbol{\beta}|\phi, \mathbf{y})$ pour l'estimation.

Pseudo-Vraisemblance Cette méthode a été introduite afin de simplifier l'expression de la vraisemblance. L'idée sous-jacente, qui rejoint les motivations des GEE, est de modifier les équations du score par des équations d'estimation plus simples qui permettent néanmoins de conserver la consistance et la normalité asymptotique des estimateurs. Plus de détails et exemples sur cette approximation peuvent être trouvés dans le chapitre 9 de Verbeke and Molenberghs [2005].

1.1.3 La méthode des GEE

Pour les données en cluster ou répétées, Liang and Zeger [1986] proposent les Generalized Estimating Equations. Ces équations d'estimation généralisées évitent le détail de la vraisemblance jointe et ne requièrent que la spécification correcte de l'espérance marginale et de sa variance. Les corrélations intra-cluster, ou intra-individu sont modélisées par une matrice de corrélation dite de *travail* que l'analyste peut choisir.

1.1.3.1 Estimation de β

L'estimateur proposé est solution du système d'équations :

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{K} \mathbf{D}_i{}^T \mathbf{V}_i{}^{-1} [\mathbf{y}_i - \boldsymbol{\mu}_i] = 0 \tag{1.13}$$

Où $\mathbf{D}_i$ représente la matrice des dérivées de $\boldsymbol{\mu}_i$ par rapport à $\boldsymbol{\beta}$ et $\mathbf{V}$ est la matrice de variance-covariance de *travail* définie par :

$$\mathbf{V}_i = \mathbf{A}_i{}^{1/2} \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}_i{}^{1/2} \tag{1.14}$$

$\mathbf{A}_i$ est la matrice diagonale composée des variances marginales $\phi V(\mu_{i,t})$. La matrice $\mathbf{R}_i(\boldsymbol{\alpha})$ représente une matrice de corrélation de travail (working correlation matrix) qui

dépend d'un vecteur de paramètres $\boldsymbol{\alpha}$ à déterminer. $\mathbf{R}_i(\boldsymbol{\alpha})$ n'est pas la vraie matrice de corrélation de la variable aléatoire $\mathbf{y}_i$, car si tel était le cas la matrice de variance-covariance $\mathbf{V}_i$ serait la vraie matrice de variance $\mathbb{V}(\mathbf{y}_i)$. La structure de la matrice de corrélation est commune à tous les individus ce qui fait que le vecteur $\boldsymbol{\alpha}$ n'est pas indexé par i, la dimension de la matrice dépend par contre du nombre d'observations effectuées par le patient.

Pour chaque individu, l'expression $U_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \mathbf{D}_i{}^T \mathbf{V}_i{}^{-1}[\mathbf{y}_i - \boldsymbol{\mu}_i]$ est similaire aux fonctions issues des approches de quasi-vraisemblance. La principale différence vient du fait qu'ici $\mathbf{V}_i$ ne dépend pas que de la moyenne mais aussi d'un paramètre de corrélation $\boldsymbol{\alpha}$. Le système d'équations (1.13) peut être écrit à l'aide de matrices combinées et diagonales par bloc :

$$U(\boldsymbol{\beta}) = \mathbf{D}^T \mathbf{V}^{-1}[\mathbf{y} - \boldsymbol{\mu}] = 0 \tag{1.15}$$

où $\mathbf{D} = (\mathbf{D}_1{}^T, ..., \mathbf{D}_K{}^T)^T$ et $\mathbf{V}$ est une matrice diagonale de taille $N \times N$ dont les blocs sont les matrices $\mathbf{V}_i$, $i \in \{1, \ldots, K\}$. Nous pouvons remarquer que ces équation supposent que les covariables sont mesurées aux mêmes instants sans données manquantes. Dans le cas contraire, μ ne serait pas défini pour chaque individu en tout temps. Par ailleurs, ces équations reposent sur l'hypothèse de nullité de l'espérance du score. Cependant, si l'étude souffre d'attrition, les sortis d'étude peuvent être informatifs. Dans ce contexte, l'espérance du score n'est pas nulle et les estimateurs obtenus par résolution de l'équation (1.15) peuvent être biaisés. La section 3.2, présente plus en détail cette problématique et les solutions non biaisées.

1.1.3.2 Estimation des paramètres de nuisance

Les paramètres $\boldsymbol{\alpha}$ et ϕ sont considérés comme des paramètres de nuisance que Liang and Zeger [1986] proposent d'estimer par une méthode consistante.

Méthode d'estimation de ϕ On utilise une méthode analogue à la statistique de Pearson (Wedderburn [1974]; McCullagh [1983]) où les résidus de Pearson sont définis par :

$$r_{i,t} = \frac{y_{i,t} - \mu_{i,t}}{\sqrt{V(\mu_{i,t})}} \tag{1.16}$$

Où $\mu_{i,t}$ peut être remplacé par son estimateur $\hat{\mu}_{i,t}$ ce qui permet d'obtenir $\hat{r}_{i,t}$. Un estimateur consistant de ϕ est alors donné par :

$$\hat{\phi} = \sum_{i=1}^{K} \sum_{t=1}^{T_i} \frac{\hat{r}_{i,t}^2}{N-p} \tag{1.17}$$

$\hat{r}_{i,t}$ dépend de $\hat{\boldsymbol{\beta}}$ à travers le lien entre $\mu_{i,t}$ et $\boldsymbol{\beta}$ mais aussi à travers le lien entre $V(\mu_{i,t})$ et $\boldsymbol{\beta}$.

Méthode d'estimation de $\boldsymbol{\alpha}$ La méthode d'estimation dépendra de la forme choisie pour la matrice de corrélation $\mathbf{R}_i(\boldsymbol{\alpha})$. Comme pour le paramètre de dispersion, le vecteur $\boldsymbol{\alpha}$ peut être estimé par n'importe quelle méthode consistante. Si tous les individus ont le même nombre de visites, alors $T_i = T$ et $\mathbf{R}_i(\boldsymbol{\alpha}) = \mathbf{R}(\boldsymbol{\alpha})$, $\forall i \in \{1, ..., K\}$. L'estimateur empirique peut être utilisé pour estimer chaque élément R_{uv} :

$$\hat{R}_{u,v} = \sum_{i=1}^{K} \frac{\hat{r}_{i,u}\hat{r}_{i,v}}{N-p} \tag{1.18}$$

Dans le cas de données non-équilibrées, le calcul de $\hat{R}_{uv}$ se fait sur l'ensemble des individus présents aux temps u et v. Tous les coefficients de corrélation n'étant pas estimés sur la même population, les matrices de corrélation redeviennent spécifiques à l'individu et à ses temps d'observation. Cette méthode est très générale ce qui permet de ne pas faire de suppositions sur la structure de corrélation des observations. Cependant, il peut être intéressant de pouvoir choisir entre différentes structures possibles afin de mieux renseigner la structure de corrélation, réduire le nombre de paramètres à estimer et s'ajuster au mieux aux données. Les méthodes associées et le choix de la matrice de corrélation seront plus amplement détaillés dans la section 1.1.5.

1.1.4 Algorithme de calcul

Afin de résoudre le système d'équations (1.13), les paramètres de nuisance sont remplacés par leurs estimations ce qui permet d'obtenir des équations qui ne dépendent que de $\boldsymbol{\beta}$:

$$\sum_{i=1}^{K} U_i[\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}\{\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta})\}] = 0 \tag{1.19}$$

Ces équations peuvent être résolues par un algorithme itératif de type Newton modifié. Supposons qu'à l'itération l nous connaissions la valeur de $\hat{\boldsymbol{\beta}}^{(l)}$, la formule itérative suivante est utilisée :

$$\hat{\boldsymbol{\beta}}^{(l+1)} = \hat{\boldsymbol{\beta}}^{(l)} - \left(-\sum_{i=1}^{K} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i \right)^{-1} \sum_{i=1}^{K} \mathbf{D}_i^T \mathbf{V}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(l)}} \tag{1.20}$$

En utilisant les matrices combinées et diagonales par blocs :

$$\hat{\boldsymbol{\beta}}^{(l+1)} = \hat{\boldsymbol{\beta}}^{(l)} - \left(-\mathbf{D}^T \mathbf{V}^{-1} \mathbf{D} \right)^{-1} \mathbf{D}^T \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(l)}} \tag{1.21}$$

Cette équation associée aux formules d'estimation des paramètres $\boldsymbol{\alpha}$ et ϕ données dans les équations (1.18) et (1.17) permettent de mettre en place l'algorithme suivant :

Algorithm 1 Generalized Estimating Equations

Initialize:
- $l = 0$
- $\boldsymbol{\beta}^{(l)}$, $\boldsymbol{\alpha}^{(l)}$, $\phi^{(l)}$
- $diff = 10$
- **while** $diff > \epsilon$ and $l \in \{0, \ldots, l_{max}\}$ **do**
  - $\mathbf{U}^{(l)} = \mathbf{D}^T \mathbf{V}^{(l)-1} (\mathbf{y} - \boldsymbol{\mu}^{(l)})$
  - $\dot{\mathbf{U}}^{(l)} = -\mathbf{D}^T \mathbf{V}^{(l)-1} \mathbf{D}^T$
  - $\boldsymbol{\beta}^{(l+1)} = \boldsymbol{\beta}^{(l)} - \dot{\mathbf{U}}^{(l)-1} \mathbf{U}^{(l)}$
  - $diff = ||\boldsymbol{\beta}^{(l+1)} - \boldsymbol{\beta}^{(l)}||_2^2$
  - $l = l + 1$
  - mise à jour des paramètres $\boldsymbol{\alpha}^{(l)}$ et $\phi^{(l)}$
- **end while**

La valeur initiale $\boldsymbol{\beta}^{(0)}$ est donnée par un estimateur de maximum de vraisemblance ou des moindres carrés en supposant les observations indépendantes. Ce premier estimateur

permet de calculer $\boldsymbol{\alpha}^{(0)}$ et $\phi^{(0)}$ par une méthode consistante comme celles proposées dans la section 1.1.3.2.

1.1.5 Structure de corrélation

La matrice de corrélation de travail $\mathbf{R}_i(\boldsymbol{\alpha})$ permet de spécifier la structure des corrélations intra-individus. L'utilisateur peut choisir différentes structures parmi lesquels certaines très classiques sont largement utilisées. La plus connue reste la matrice identité, qui suppose l'indépendance entre toutes les observations. La liste présentée, non-exhaustive, permet de décrire les structures usuelles.

Corrélation échangeable ou symétrique La structure échangeable ou symétrique, suppose que toutes les corrélations sont identiques, quel que soit l'intervalle de temps entre deux observations. Dans ce contexte, α est un scalaire et la matrice de corrélation de taille $T_i \times T_i$ s'écrit :

$$\mathbf{R}_i(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha & \dots & \alpha \\ \alpha & 1 & \alpha & \dots & \alpha \\ \alpha & \alpha & 1 & \dots & \alpha \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \alpha & \dots & 1 \end{pmatrix} \quad \alpha \geq 0$$

Cette structure dépend d'une hypothèse forte. Elle est souvent utilisée pour des observations organisées en groupes.

Corrélation auto-régressive d'ordre un : AR(1) Dans un contexte longitudinal, il semble raisonnable d'assumer une dépendance du temps. Cette structure suppose que les corrélations sont d'autant plus fortes que les observations sont rapprochées dans le temps et inversement.

$$\mathbf{R}_i(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha^2 & \dots & \alpha^{|T_i-1|} \\ \alpha & 1 & \alpha & \dots & \alpha^{|T_i-2|} \\ \alpha^2 & \alpha & 1 & \dots & \alpha^{|T_i-3|} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha^{|T_i-1|} & \alpha^{|T_i-2|} & \alpha^{|T_i-3|} & \dots & 1 \end{pmatrix}$$

Ce qui peut être écrit $corr(y_{i,t}, y_{i,t'}) = \alpha^{|t-t'|}$ où $|.|$ est la fonction valeur absolue.

Corrélation M-dependent Une alternative possible part de l'hypothèse qu'à partir d'un certain temps, les corrélations sont inexistantes mais que pour les observations rapprochées, l'hypothèse d'auto-corrélation est plausible. Dans ce cas on suppose que :

$$R_{uv} = \begin{cases} \alpha^{|u-v|} & \text{si } |u-v| < M \\ 0 & \text{sinon} \end{cases}$$

Corrélation non structurée Cette structure - moins restrictive - propose que chaque coefficient soit estimé par une méthode des moments. On retrouve alors la méthode générale définie dans l'équation (1.18). Il n'est pas garanti que la matrice de corrélation ainsi construite soit inversible surtout dans le cas de données non équilibrées (Hardin and Hilbe [2003]). Dans un contexte longitudinal, lorsque l'étude souffre de perte de suivi, cette structure est donc déconseillée.

D'autres structures sont envisageables comme une matrice fixe (si l'utilisateur connaît la structure et les valeurs *a priori*), ou une spécification libre si la forme est connue d'une autre source. Pour choisir entre plusieurs structures plausibles, des critères de sélection de matrice de corrélation de travail ont été développés. Ils sont détaillés dans la section 1.2.5.

1.1.6 Matrice de variance covariance

L'estimateur $\hat{\boldsymbol{\beta}}$ obtenu par GEE est consistant et asymptotiquement Gaussien multivarié même si $\mathbf{R}_i(\boldsymbol{\alpha})$ est mal spécifiée (Liang and Zeger [1986]). Supposons que $\mathbf{R}_i(\boldsymbol{\alpha})$ soit la vraie matrice de corrélation, alors la variance de $\hat{\boldsymbol{\beta}}$ serait estimée par :

$$\Omega_{\mathbf{R}} = \left(\sum_{i=1}^{K} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i \right)^{-1} \tag{1.22}$$

Supposons une structure indépendante où $\mathbf{R}_i(\boldsymbol{\alpha}) = \mathbf{I}$, on a alors :

$$\Omega_{\mathbf{I}} = \left(\sum_{i=1}^{K} \mathbf{D}_i^T \mathbf{A}_i^{-1} \mathbf{D}_i \right)^{-1} \tag{1.23}$$

Cette matrice est appelée estimateur *naïf*. En général, la matrice de corrélation de

travail n'est *pas bien renseignée* au sens où ce n'est pas la vraie matrice de corrélation. Il est nécessaire d'utiliser un estimateur robuste, consistant, pour la matrice de variance covariance de $\hat{\boldsymbol{\beta}}$. Pour cela, Liang and Zeger [1986] proposent d'utiliser l'estimateur robuste de White [1982] et Huber [1967] décrit pour les GEE par Carroll et al. [1998] :

$$\mathbf{V_R} = \Omega_{\mathbf{R}} \left(\sum_{i=1}^{K} \mathbf{D}_i^t \mathbf{V}_i^{-1} \mathbb{C}ov(\mathbf{y}_i) \mathbf{V}_i^{-1} \mathbf{D}_i \right) \Omega_{\mathbf{R}} \tag{1.24}$$

Où $\mathbb{C}ov(\mathbf{y}_i)$ peut être estimée par $(\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^T$, ou comme Pan [2001c] le propose par $\mathbf{A}_i^{\frac{1}{2}} \left[\sum_{i=1}^{K} \mathbf{A}_i^{-\frac{1}{2}} (\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^T \mathbf{A}_i^{-\frac{1}{2}} \right] \mathbf{A}_i^{\frac{1}{2}}$.

1.2 La sélection de modèles pour GEE

Supposons que l'on souhaite modéliser le lien entre les covariables mesurées $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ et la réponse $\mathbf{y}$ en ajustant un modèle linéaire généralisé. Nous ne sommes pas sûrs du sous-groupe de covariables à inclure dans le modèle mais nous cherchons le modèle associé au paramètre $\boldsymbol{\beta}^*$ composé de $q \leq p$ coefficients non nuls. Les coefficients de tous les autres prédicteurs sont mis à zéro. Si plusieurs sous-modèles sont plausibles, l'enjeu devient de choisir le *meilleur* modèle, celui qui permet le meilleur ajustement sans ajouter de variables inutiles.

Les plus intuitifs cherchent à optimiser les capacités de prédiction du modèle tout en évitant le sur-ajustement, en minimisant la somme des carrés des résidus (RSS) ou l'erreur moyenne de prédiction (PMSE). Ce type de méthode fait souvent intervenir des techniques robustes comme la validation croisée (CV) (Stone [1974]; Geisser [1975]) ou le bootstrap Efron [1979] afin d'obtenir un modèle reproductible.

Certains utilisent le maximum de vraisemblance associé à une pénalité discrète sur la complexité du modèle. Parmi les critères les plus classiques, nous pouvons citer l'Information Criterion (AIC) de Akaike [1973], le Bayesian Information Criterion (BIC) de Schwarz et al. [1978] ou le C_p de Mallows [1973] qui font intervenir une pénalisation à l'aide des degrés de liberté du modèle. On choisit alors de conserver le modèle qui optimise le critère de sélection choisi.

La sélection de modèles est un domaine très étudié, y compris pour les données longitudinales et les GEE. Cette section propose un état de l'art des critères de sélection de modèles pour GEE de façon non-exhaustive.

1.2.1 Critères de type somme des carrés des résidus

Il est possible de définir différentes fonctions de perte à minimiser en utilisant les données longitudinales ou les informations données par les GEE. Dziak and Li [2007] suggèrent que la façon la plus simple de définir une mesure de qualité d'ajustement d'un modèle est d'utiliser la somme des résidus au carré. Dans le cadre d'une réponse continue ou binaire, le critère à minimiser est :

$$RSS(\hat{\boldsymbol{\beta}}^*) = \sum_{i,t}(y_{i,t} - \hat{y}_{i,t})^2 = (\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}}) \tag{1.25}$$

où $\hat{y}_{i,t} = g^{-1}(\mathbf{x}_{i,t}^T\hat{\boldsymbol{\beta}}^*)$, et $\hat{\boldsymbol{\beta}}^*$ est l'estimateur de $\boldsymbol{\beta}^*$ obtenu par GEE. Afin de prendre en compte la variabilité des observations, nous pouvons pondérer les observations par leur variance :

$$WRSS_{\mathbf{A}}(\hat{\boldsymbol{\beta}}^*) = \sum_{i,t}\frac{(y_{i,t} - \hat{y_{i,t}})^2}{\hat{\mathbb{V}}(y_{i,t})} = \sum_{i=1}^{K}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^T\hat{\mathbf{A}}_i^{-1}(\mathbf{y}_i - \hat{\mathbf{y}}_i) \tag{1.26}$$

Les observations les moins sures, associées à une plus grande variance auront moins d'impact sur la quantité finale. Pour adapter cette définition aux données longitudinales, il faut intégrer la corrélation intra-individuelle. La matrice $\hat{\mathbf{A}}_i$ peut être remplacée par la matrice de corrélation ou de variance-covariance :

$$\begin{aligned} WRSS_{\mathbf{R}}(\hat{\boldsymbol{\beta}}^*) &= \sum_{i=1}^{K}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^T\hat{\mathbf{R}}_i^{-1}(\mathbf{y}_i - \hat{\mathbf{y}}_i) \\ WRSS_{\mathbf{V}}(\hat{\boldsymbol{\beta}}^*) &= \sum_{i=1}^{K}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^T\hat{\mathbf{V}}_i^{-1}(\mathbf{y}_i - \hat{\mathbf{y}}_i) \end{aligned} \tag{1.27}$$

Les matrices estimées $\hat{\mathbf{A}}_i$, $\hat{\mathbf{R}}_i$ et $\hat{\mathbf{V}}_i$ peuvent être estimées une première fois à l'aide du modèle complet composé de p variables et être utilisées pour tous les autres modèles.

Cantoni et al. [2005] remarquent qu'il est possible d'améliorer cette définition en ajoutant des poids définis par l'analyste. Selon leur définition ces poids pourraient prendre en compte l'hétéroscédasticité ou permettre la robustesse de la fonction de perte :

$$WRSS_{\mathbf{\Omega}}(\hat{\boldsymbol{\beta}}^*) = \sum_{i,t} \omega_{i,t}^2 \frac{(y_{i,t} - \hat{y_{i,t}})^2}{\hat{\mathbb{V}}(y_{i,t})} = \sum_{i=1}^{K} (\mathbf{y}_i - \hat{\mathbf{y}}_i)^T \mathbf{\Omega}_i \hat{\mathbf{A}}_i^{-1} \mathbf{\Omega}_i (\mathbf{y}_i - \hat{\mathbf{y}}_i) \tag{1.28}$$

où $\mathbf{\Omega}_i = diag\{\omega_{i,t}\}$ est la matrice des poids définie pour chaque modèle.

La méthode de validation croisée (CV) en *leave-one-out* (LOO) peut être utilisée pour chacun de ces critères en laissant de côté un individu plutôt qu'une observation pour ne pas diviser un groupe d'observations (Cantoni et al. [2007]). Pour réduire le temps de calcul induit par cette méthode, il est possible d'utiliser la validation croisée en k-blocs. Cantoni et al. [2007] proposent d'estimer la $WRSS_{\mathbf{V}}$ par validation croisée et utilisent la méthode Monte Carlo par Chaînes de Markov (MCMC) afin de ne pas calculer le critère pour tous les modèles. Une autre approche de validation croisée suggérée par Pan [2001b], propose de minimiser une combinaison linéaire du biais de prédiction espéré (EPB) :

$$EPB(\hat{\boldsymbol{\beta}}^*) = \mathbb{E}_{\mathbf{X}} \mathbb{E}_y \left[\mathbf{U}(\mathbf{y}|\mathbf{X}, (\hat{\boldsymbol{\beta}}^*) \right] \tag{1.29}$$

Pan [2001b] prédit une fonction de risque sur des données futures à l'aide de la validation croisée lissée par bootstrap pour réduire la variance.

Dans le contexte particulier des observations indépendantes et du modèle linéaire, la validation croisée par LOO est asymptotiquement équivalente à l'AIC et au C_p de Mallows (Efron [1986]; Stone [1977]). Pour étendre ces deux critères aux GEE, nous devons définir le degré de liberté pour l'ajustement du modèle et une mesure de qualité d'ajustement.

1.2.2 Généralisation du C_p de Mallows

Cantoni et al. [2005] ont proposé une extension du C_p de Mallows (Mallows [1973]) pour les GEE qui intègre une fonction de perte pondérée. Ces poids peuvent être ajustés

pour prendre en compte les corrélations intra-sujets, ou permettre d'obtenir une fonction de perte robuste aux valeurs aberrantes. Supposons toutes les observations indépendantes, le C_p de Mallows pour le modèle associé au vecteur $\boldsymbol{\beta}^*$ est défini par :

$$C_p(\hat{\boldsymbol{\beta}}^*) = WRSS_{\mathbf{A}}(\hat{\boldsymbol{\beta}}^*) - N + 2q \tag{1.30}$$

Ce critère de qualité ne permet pas de prendre en compte les corrélations dues au temps. Pour ce faire, Cantoni et al. [2005] propose une extension : les GC_p. Leur forme robuste aux données aberrantes peut être complexe mais dans le contexte *classique*, ils prennent la forme suivante :

$$GC_p(\hat{\boldsymbol{\beta}}^*) = WRSS_{\mathbf{A}}(\hat{\boldsymbol{\beta}}^*) - N + 2df_C \tag{1.31}$$

où $\hat{\mathbf{A}}_i$ est estimé sous le modèle complet à p variables. Le degré de liberté df_C est utilisé par Dziak and Li [2007] comme une nouvelle définition du degré de liberté pour les GEE avec $df_C = tr(\mathbf{H}^{-1}\mathbf{Q})$, $\mathbf{H} = K^{-1}\sum_i \mathbf{D}_i^T\mathbf{V}_i^{-1}\mathbf{D}_i$ et $\mathbf{Q} = K^{-1}\sum_i \mathbf{D}_i^T\mathbf{A}_i^{-1}\mathbf{D}_i$. Lorsque l'on utilise une matrice de corrélation indépendante alors $df_C = q$. Les calculs peuvent être longs si l'on veut estimer des poids robustes puisqu'il faut faire appel aux approches de Monte Carlo.

1.2.3 Généralisation de l'AIC

Pan [2001a] introduit une extension de l'AIC (Akaike [1973]) aux données longitudinales. L'AIC est une mesure de la qualité d'un modèle, basée sur la vraisemblance pénalisée par le nombre de covariables incluses dans le modèle :

$$AIC(\hat{\boldsymbol{\beta}}^*) = -2l(\boldsymbol{\beta}^*) + 2q \tag{1.32}$$

Où $l(.)$ est la log-vraisemblance. La définition de l'AIC fait intervenir la mesure de Kullback-Leiber entre la vraisemblance du modèle et celle du vrai modèle inconnu. Pour les données longitudinales analysées par GEE, ces quantités ne sont pas définies. Pan [2001a] propose de les remplacer par des log-quasi-vraisemblances (quantité définie dans

l'équation (1.12)). On obtient alors le Quasi-likelihood under Independance model Criterion (QIC) :

$$QIC(\hat{\boldsymbol{\beta}}^*) = -2Q(\hat{\boldsymbol{\beta}}_I^*) + 2\mathbf{tr}(\hat{\boldsymbol{\Omega}}_I \hat{\mathbf{V}}_{\mathbf{R}}) \tag{1.33}$$

où $\hat{\boldsymbol{\beta}}_I^*$ est l'estimateur de $\boldsymbol{\beta}^*$ sous l'hypothèse d'indépendance (i.e $\mathbf{R}_i(\boldsymbol{\alpha}) = \mathbf{I}_{T_i}$) et $\boldsymbol{\Omega}_I$ est la matrice de variance-covariance naïve sous hypothèse d'indépendance définie dans l'équation (1.23). $\hat{\mathbf{V}}_{\mathbf{R}}$ est l'estimateur sandwich robuste de $\mathbb{C}ov(\boldsymbol{\beta}^*)$ défini dans l'équation (1.14). Si le modèle est adéquat et si les réponses sont toutes indépendantes alors $tr(\hat{\boldsymbol{\Omega}}_I \hat{\mathbf{V}}_{\mathbf{R}}) \approx q$. Pan [2001a] propose une version simplifiée de ce critère :

$$QIC_u(\hat{\boldsymbol{\beta}}^*) = -2Q(\hat{\boldsymbol{\beta}}_I^*) + 2q \tag{1.34}$$

Cependant, le QIC repose sur l'hypothèse d'indépendance ce qui peut nuire à ses performances en cas de fortes corrélations. Pour y remédier Pan [2001a] propose une extension naïve de l'AIC :

$$QIC_{\mathbf{R}}(\hat{\boldsymbol{\beta}}^*) = -2Q(\hat{\boldsymbol{\beta}}^*) + 2\mathbf{tr}(\hat{\boldsymbol{\Omega}}_I \hat{\mathbf{V}}_{\mathbf{R}}) \tag{1.35}$$

où $\hat{\boldsymbol{\beta}}^*$ est estimé par GEE avec une matrice de corrélation $\mathbf{R}_i(\boldsymbol{\alpha})$ qui n'est pas la matrice identité de taille $T_i \times T_i$.

1.2.4 Critères de type BIC

En supposant l'indépendance de toutes les réponses, le Bayesian Information Criterion (BIC) de Schwarz et al. [1978] est défini par :

$$BIC(\hat{\boldsymbol{\beta}}^*) = -2l(\hat{\boldsymbol{\beta}}^*) + \log(N)q \tag{1.36}$$

Comme pour la généralisation de l'AIC, la log-vraisemblance est remplacée par une quasi-log-vraisemblance. Une deuxième quantité est ambiguë pour les données corrélées : le nombre d'observations. Deux extensions naïves utilisant respectivement le nombre de sujet K ou le nombre d'observations N sont possibles :

$$BIC_N(\hat{\boldsymbol{\beta}}^*) = -2Q(\hat{\boldsymbol{\beta}}_I^*) + df\log(N)$$
$$BIC_K(\hat{\boldsymbol{\beta}}^*) = -2Q(\hat{\boldsymbol{\beta}}_I^*) + df\log(K) \tag{1.37}$$

où le degré de liberté df peut être estimé par la dimension q du modèle ou par une formule plus générale comme pour les GC_p où $df_C = tr(\mathbf{H}^{-1}\mathbf{Q})$. Dziak and Li [2006] proposent des critères de type BIC en remplaçant la quasi-vraisemblance par une somme des carrés des résidus pondérés comme utilisée par Wang et al. [2006]. Il est alors possible de définir plusieurs critères de la forme :

$$BIC(\hat{\boldsymbol{\beta}}^*) = N\log(WRSS_{\mathbf{Q}}(\hat{\boldsymbol{\beta}}^*)/N) + \log(\tilde{n})df \tag{1.38}$$

où $Q \in \{A, R, V\}$ peut être estimé sur le modèle complet et $\tilde{n}$ est à définir. Puisque le nombre de sujets n'est plus égal au nombre d'observations le choix de $\tilde{n}$ est ambigu. Il est possible de considérer un critère *léger* en utilisant $\tilde{n} = K$ ou un critère *lourd* en utilisant $\tilde{n} = N$. Fu [2003] propose une *taille effective de l'échantillon* qui prend en compte l'importance de la corrélation intra-sujet :

$$\tilde{n} = \sum_i \frac{T_i^2}{\sum_{i,t} R_{i,t}} \tag{1.39}$$

On observe alors que $K \leq \tilde{n} \leq N$, diminue à mesure que les observations deviennent plus corrélées. Nous utiliserons par la suite le critère de type BIC défini par :

$$BIC_n(\hat{\boldsymbol{\beta}}^*) = N\log(WRSS_{\mathbf{R}}/N) + \log(\tilde{n})df \tag{1.40}$$

Une étude par simulation effectuée par Dziak and Li [2006], a montré que le choix parmi les critères des équations (1.37) et (1.38) impacte peu la performance de la sélection.

1.2.5 Critères de sélection de structure de matrice corrélation

Le choix de la matrice de corrélation est stratégique, il permet d'améliorer les estimations des coefficients de régression (Wang and Carey [2003]) et permet de mieux représenter

les données observées et leur structure. Cependant, si la structure de la moyenne est mal spécifiée il n'y a que peu d'intérêt à chercher une matrice de corrélation optimale puisque la structure des erreurs sera modifiée par l'écart entre le vrai modèle et celui utilisé. Par contre, une fois la sélection des variables à inclure dans le modèle réalisée, autrement dit une fois que l'on juge la fonction moyenne bien spécifiée, sélectionner la matrice de corrélation qui représente le mieux les corrélations intra-sujet améliore les estimations, les intervalles de confiance et donc l'inférence. On trouve une littérature riche concernant la sélection de la matrice de corrélation de travail, nous présentons ici certains critères largement utilisés. Pour une fonction moyenne fixée, il est possible d'utiliser le QIC_R, défini dans l'équation (1.33). Le critère devient alors :

$$QIC_{\mathbf{R}}(\mathbf{R}) = -2Q(\hat{\boldsymbol{\beta}}_{\mathbf{R}}) + 2\mathbf{tr}(\hat{\boldsymbol{\Omega}}_I \hat{\mathbf{V}}_{\mathbf{R}}) \tag{1.41}$$

Nous choisirons la matrice de corrélation associée au QIC_R le plus faible. Les performances de ce critère largement utilisé ont été étudiées par Hin et al. [2012] qui montrent son bon taux de sélection. Hin and Wang [2009] proposent une amélioration du critère, le Correlation Information Information (CIC) :

$$CIC(\mathbf{R}) = \mathbf{tr}\left(\hat{\boldsymbol{\Omega}}_I \hat{\mathbf{V}}_{\mathbf{R}}\right) \tag{1.42}$$

où les matrices $\hat{\boldsymbol{\Omega}}_I$ et $\hat{\mathbf{V}}_{\mathbf{R}}$ sont évaluées en $\hat{\boldsymbol{\beta}}_{\mathbf{R}}$. Une autre classe de critères se déduit des critères de Rotnitzky and Jewell [1990]. Ils proposent d'étudier la distance entre la matrice de corrélation étudiée et la vraie matrice de corrélation à l'aide des définitions suivantes :

$$\begin{aligned}
\mathbf{Q}_0 &= \frac{1}{K}\sum_{i=1}^{K} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i \\
\mathbf{Q}_1 &= \frac{1}{K}\sum_{i=1}^{K} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbb{C}ov(\mathbf{y}_i) \mathbf{V}_i^{-1} \mathbf{D}_i \\
\mathbf{Q} &= \mathbf{Q}_0^{-1}\mathbf{Q}_1
\end{aligned} \tag{1.43}$$

Ces trois matrices permettent de définir deux mesures d'ajustement :

$$C_1 = \mathbf{tr}(\mathbf{Q})/p$$
$$C_2 = \mathbf{tr}(\mathbf{Q}^T\mathbf{Q})/p \tag{1.44}$$

Pour une matrice de corrélation de travail $\mathbf{R}$ proche de la vraie matrice de corrélation de la réponse, le couple (C_1, C_2) doit être proche du couple $(1, 1)$. Hin et al. [2012] proposent d'étudier le critère :

$$RJ = \sqrt{(C_1 - 1)^2 + (C_2 - 1)^2} \tag{1.45}$$

qui mesure la distance au point optimal. On choisira donc la matrice associée au RJ le plus faible. Les travaux de Rotnitzky and Jewell [1990] ont inspiré d'autres critères parmi lesquels nous pouvons citer les travaux de Shults et al. [2009] et Wang and Carey [2003].

Conclusion Les critères de sélection classiques de type AIC et BIC ont été étendus aux données corrélées grâce aux notions de quasi-vraisemblance et aux degrés de liberté adaptés aux GEE. Cependant, certains de ces critères ne prennent pas bien en compte les corrélations intra-individu (exemple du QIC) et ces critères ne sont pas toujours implémentés dans les logiciels statistiques. Ces critères sont à calculer pour tous les modèles plausibles ce qui peut être coûteux en temps de calcul puisque pour p variables observées il y a 2^p modèles à tester, sans compter les interactions possibles. De ce fait, il devient impossible d'explorer l'ensemble des 2^p modèles possibles dès que le nombre de covariables potentielles p est supérieur à 30. Par ailleurs, ces méthodes supposent que toutes les covariables et la réponse sont mesurées aux même instants. Dans le cas contraire, la moyenne, la log-quasi-vraisemblance et les matrices de covariances et corrélations pourraient ne pas être définies.

Chapitre 2

Sélection de variables et régularisation dans les GEE : état de l'art

La technologie de nos outils de mesures (séquençage, imagerie médicale, dosage de protéines...) permet d'obtenir des bases de données riches en informations. Il n'est pas rare aujourd'hui d'observer de trente à quelques milliers de variables pour une base de données. Lorsque l'objectif est de les relier à un critère clinique d'intérêt, il est crucial de pouvoir sélectionner un sous-groupe de variables qui expliquent ou prédisent au mieux la réponse observée. Dans ce contexte, les outils de sélection de modèles présentés dans le chapitre 1 ne sont pas suffisants. Les méthodes de régularisation permettent de parer aux limites de la sélection de modèles et font face à la problématique de multi-colinéarité inévitable en présence de nombreuses variables. Ces méthodes, historiquement développées pour une réponse continue font intervenir des analogies avec les moindres carrés. Elle peuvent cependant être étendues aux GLM en utilisant la log-vraisemblance négative ainsi qu'aux GEE grâce à la log-quasi-vraisemblance négative.

La plupart des critères détaillés dans le chapitre 1 sont de la forme :

$$G(\boldsymbol{\beta}) + \mathcal{P}(\boldsymbol{\beta}) \tag{2.1}$$

Où $G(\boldsymbol{\beta})$ est une fonction de perte qui peut être une RSS, une WRSS, une log-

vraisemblance ou quasi-vraisemblance négative. Les critères de sélection de modèles font intervenir une mesure de complexité de modèle $\mathcal{P}(.)$ discrète. Cette mesure comprend souvent une pénalité L_0 sur le vecteur de paramètre $\boldsymbol{\beta}$ ce qui revient au nombre de paramètres non nuls du modèle. On appelle cela la méthode du *Best Subset Selection.* Un simple changement dans la base de données peut conduire à un modèle final différent, ce qui apporte du biais et de l'instabilité (Breiman et al. [1996]; Miller [2002]; Dziak and Li [2006]). Motivé par cette faiblesse, les méthodes de régularisation proposent d'utiliser une fonction pénalité continue en $\boldsymbol{\beta}$.

2.1 Les régressions pénalisées

L'objectif des régressions pénalisées est de minimiser la RSS en imposant une contrainte sur les coefficients du modèle afin de combiner bonne prédiction et sparsité du modèle.

2.1.1 La régression Bridge

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

$$\min_{\boldsymbol{\beta}} \left\{ ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 \right\}, \quad \sum_j |\beta_j|^\gamma < c \tag{2.2}$$

On cherche alors à minimiser la RSS sous contrainte où pour chaque $c \geq 0$ il existe un $\lambda \geq 0$ tel que le problème d'optimisation (2.2) puisse s'écrire :

$$\min_{\boldsymbol{\beta}} \left\{ ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda \sum_j |\beta_j|^\gamma \right\} \tag{2.3}$$

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 $\lambda = \infty$ toutes les variables sont associées à un coefficient nul. Le paramètre γ permet de définir différentes régions de contraintes (figure 2.1). Fu [1998] étudie les propriétés de l'estimateur pour $\gamma \geq 1$ et propose l'algorithme du Modified Newton Raphson (M-N-R) pour $\gamma > 1$.

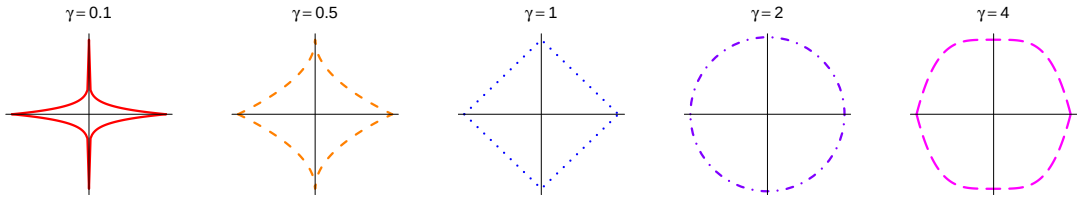


FIGURE 2.1 – Régions de contraintes en fonction de γ pour deux paramètres (β_1, β_2)

La figure 2.2 représente la pénalité Bridge pour différentes valeurs de γ. Les zones de contraintes $\sum_j |\beta_j|^\gamma < c$ associées sont représentées dans la figure 2.1. Plus le paramètre γ est faible, plus la zone de contrainte est restreinte à une faible zone de possibilités pour l'estimateur. L'objectif est alors de trouver le compromis entre une zone de contrainte restreinte et un estimateur le moins biaisé possible. Pour $\gamma < 2$, les zones de contraintes admettent des sommets, c'est cette particularité qui permet la sélection, si l'optimum se trouve sur un sommet de la zone alors un paramètre est estimé exactement à zéro, plus il y a de sommets, plus il y a de chances d'avoir des coefficients nuls. Pour $\gamma \geq 1$ les pénalités sont convexes ce qui permet d'obtenir des propriétés intéressantes, utiles à la résolution du problème d'optimisation.

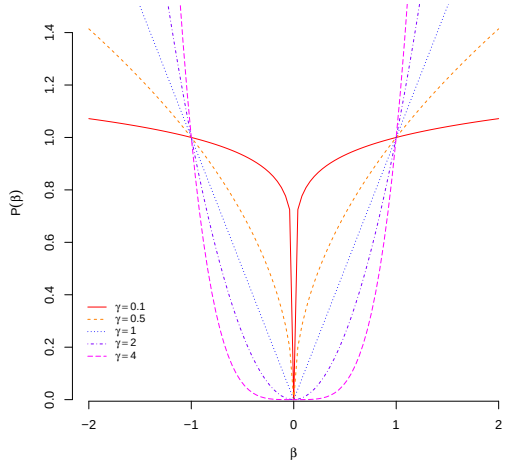


FIGURE 2.2 – Différentes pénalités Bridge en fonction de γ

Le choix de γ peut être arbitraire ou par optimisation d'un critère basé sur les données, cependant ce paramètre supplémentaire augmente la complexité de la méthode (Friedman et al. [2001]). Deux cas particuliers de cette pénalité ont été largement étudiés, la régression Ridge pour $\gamma = 2$ et le Least Absolute Shrinkage and Selection Operatot (LASSO) pour $\gamma = 1$ de Tibshirani [1996].

2.1.2 Ridge

La régression Ridge utilise une pénalité en norme L_2, qui permet de rétrécir les coefficients (ce que on appelle le *shrinkage*) mais ne les met pas à zéro. Elle permet de stabiliser l'estimation des coefficients dans un contexte de fortes corrélations inter-variables. L'estimateur est solution de :

$$\min_{\boldsymbol{\beta}} \left\{ ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_2^2 \right\} \tag{2.4}$$

Il existe un λ à partir duquel l'inverse de la matrice $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$ existe. On trouve donc un problème d'optimisation convexe qui possède une solution analytique :

$$\hat{\boldsymbol{\beta}}_{Ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \tag{2.5}$$

Cet estimateur fut introduit par Hoerl and Kennard [1970] afin d'améliorer les capacités de prédiction du modèle. Bien que biaisé, l'estimateur $\hat{\boldsymbol{\beta}}_{Ridge}$ possède une variance plus faible que l'éstimateur des moindre carré, problème bien connu du compromis biais-variance. Le fait d'ajouter un terme sur la diagonale de la matrice de Gram $\mathbf{X}^T\mathbf{X}$ permet de stabiliser les calcul de l'inverse, ce qui rend cet estimateur plus robuste au problème de multi-colinéarité.

2.1.3 LASSO

Introduit par Tibshirani [1996], le LASSO propose une norme L_1 sur le vecteur de paramètres $\boldsymbol{\beta}$ qui peut combiner les capacités d'une norme L_0 à supprimer automatiquement les petits coefficients, et les capacités d'une norme L_2 à stabiliser les coefficients (*shrinkage*). Le LASSO est solution du problème d'optimisation suivant :

$$\min_{\boldsymbol{\beta}} \left\{ ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1 \right\} \tag{2.6}$$

L'utilisation de la norme L_1 permet de réduire les petits coefficients à zéro et donc de simplifier les modèles. Cette norme a l'avantage d'être convexe et donc de garantir l'unicité de la solution. L'inconvénient de cette pénalité est que sa non différentiabilité en zéro, représentée dans les figures 2.1 et 2.2, complique son calcul, il n'existe plus de solution analytique linéaire en $\boldsymbol{\beta}$ comme pour la pénalité Ridge. De nombreux algorithmes ont été développés pour trouver la solution au problème d'optimisation (2.6). Tibshirani [1996] propose un programme linéaire, Fu [1998] propose l'algorithme du "shooting LASSO", Fan and Li [2001] propose une méthode de Newton modifiée et l'algorithme Least Angle Regression (LARS) est détaillé par Efron et al. [2004]. Le LASSO peut biaiser les estimations en *écrasant* trop fortement les coefficients. Pour remédier à cette problématique Fan and Li [2001] proposent la pénalité (SCAD) qui pénalise plus fortement les coefficients les plus élevés (en valeurs absolues).

D'autres pénalités ont été proposées pour adapter la régularisation à divers contextes comme l'Elastic-Net, de Zou and Hastie [2005], composé d'une pénalité Ridge et d'une pénalité LASSO ou l'adaptive LASSO de Zou [2006] qui utilise des poids spécifiques à chaque variable. Dans le contexte de données issues de génétique le smooth-LASSO (Tibshirani et al. [2005]) et le fused-LASSO (Hebiri et al. [2011]) utilisent une pénalité LASSO associée à une pénalité sur la différence entre deux paramètres consécutifs du vecteur $\boldsymbol{\beta}$.

2.1.4 Group-LASSO

Initialement, le LASSO fut introduit pour des covariables continues ou binaire. Lorsqu'une des covariables est qualitative à d modalités, cette dernière est convertie en un bloc de $d-1$ indicatrices. Une sélection qui ne ferait apparaitre qu'une partie du bloc qui définit la variable qualitative a peu de sens. Sélectionner une variable n'est pas sélectionner une ou plusieurs de ces modalités, il faut pouvoir sélectionner le bloc des indicatrices. Pour pallier cette difficulté, Yuan and Lin [2006] ont introduit le Group-LASSO 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. la variable avec toutes ses modalités) soit ne pas sélectionner le groupe (i.e. aucune des modalités). Supposons que l'on dispose de p variables chacune composée de M modalités, alors l'éstimateur est solution de :

$$\min_{\boldsymbol{\beta}} \left\{ ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda \sum_{j=1}^{p} \sqrt{\sum_{m=1}^{M} \beta_{m,j}^2} \right\} \tag{2.7}$$

Ce problème d'optimisation est généralisable aux cas où le nombre de modalités diffère d'une variable à l'autre. De cette manière, en présence de variables qualitatives et quantitatives, nous pourrons considérer les qualitatives comme des groupes à $M = d - 1$ variables et les quantitatives comme un groupe d'une seule variable. Introduit pour le cas des variables discrètes, cette pénalité a été utilisée pour les variables continues organisées en groupes comme par exemple les données d'expressions génétiques organisées par gènes. L'optimum du problème (2.7) peut être calculé grâce à l'algorithme Group-LARS de Yuan and Lin [2006].

Remarques Les régressions pénalisées peuvent être étendues au contexte des GLM en remplaçant la RSS par la log-vraisemblance changée de signe. On cherche alors à minimiser cette quantité combinée à une pénalité continue. Pour trouver la solution de ce nouveau problème d'optimisation, Tibshirani [1996] propose un algorithme de type Newton modifié et Friedman et al. [2010] proposent l'algorithme de Coordinate Descent. Malheureusement, la vraisemblance n'est pas définie pour les GEE ce qui pose le problème du critère à minimiser pour obtenir des GEE pénalisés.

2.2 Les GEE pénalisées

Les GEE évitent de spécifier la vraisemblance jointe en utilisant des hypothèses sur les deux premiers moments de la réponse ce qui transforme le problème d'optimisation défini dans l'équation (2.1) en système d'équations pénalisées.

2.2.1 Équations pénalisées

Si l'on considère la matrice de variance covariance de travail $\mathbf{V}_i$ définie dans l'équation (1.14) comme fixe, nous pouvons considérer une extension des régressions pénalisées aux GEE comme suit :

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{K} (\mathbf{y}_i - \boldsymbol{\mu}_i)^T \mathbf{V}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) + \mathcal{P}(\boldsymbol{\beta}) \right\} \tag{2.8}$$

On cherchera alors à trouver quand sa dérivée s'annule, ce qui donne des équations d'estimation généralisées pénalisées (PGEE) :

$$U_P(\boldsymbol{\beta}) = \sum_{i=1}^{K} {\mathbf{D}_i}^T {\mathbf{V}_i}^{-1} [\mathbf{y}_i - \boldsymbol{\mu}_i] - \frac{\partial}{\partial \boldsymbol{\beta}} \mathcal{P}(\boldsymbol{\beta}) = 0 \tag{2.9}$$

On retrouve en première partie le score défini dans l'équation (1.13) qui définit les GEE et la dérivée de la pénalité utilisée. Les équations d'estimations pénalisées et leurs propriétés asymptotiques (existence et unicité) ont été étudiées par Fu [2003]. Il s'intéresse à la pénalité Bridge et ses deux cas particuliers : le LASSO et la pénalité Ridge. Le concept d'équations pénalisées peut se généraliser à toutes formes de pénalité convexe. Blommaert et al. [2014] proposent d'utiliser une combinaison de deux pénalités pour intégrer le problème de multicolinéarité dans la sélection de variables dans le cas particulier d'une réponse gaussienne. La première est de type L_1 (i.e LASSO ou SCAD) ce qui permet de sélectionner un sous-groupe de variables et la deuxième est une pénalité Ridge comme proposé pour l'Elastic-Net de Zou and Hastie [2005].

Wang et al. [2012] proposent la pénalité SCAD dans un contexte de grande dimension où le nombre de covariables p augmente avec le nombre d'individus et construisent des résultats asymptotiques en supposant uniquement que cette divergence est du même ordre. Les auteurs montrent que l'estimateur est consistant même si la matrice de corrélation de travail est mal spécifiée.

Ces équations pénalisées reposent sur les mêmes hypothèses que les GEE. Par conséquent elles n'accepetent pas de données manquantes sur la réponse ni sur les covariables. Pour ce qui est des sortis d'études, l'estimateur obtenu sera biaisé s'il ne respecte pas des

conditions restrictives (section 3.2).

2.2.2 La méthode du LQA

Fan and Li [2001] propose une méthode générale basée sur la vraisemblance pénalisée qui utilise un Newton modifié. Leur algorithme utilise une méthode d'approximation locale quadratique (LQA) afin de contourner le problème de non-différentiabilité des pénalités. Appliquée aux GEE, cette méthode propose d'estimer $U_P(\boldsymbol{\beta})$ par un développement de Taylor. Soit $\boldsymbol{\beta}^{(0)}$ un estimateur initial proche de la solution :

$$\begin{aligned} U_P(\boldsymbol{\beta}) &\approx U_P(\boldsymbol{\beta^{(0)}}) + \dot{U}_P(\boldsymbol{\beta^{(0)}})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}) = 0 \\ U_P(\boldsymbol{\beta}) &\approx (U(\boldsymbol{\beta^{(0)}}) - \dot{\mathcal{P}}(\boldsymbol{\beta}^{(0)})) + (\dot{U}(\boldsymbol{\beta^{(0)}}) - \ddot{\mathcal{P}}(\boldsymbol{\beta}^{(0)}))(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}) = 0 \end{aligned} \tag{2.10}$$

Où $U(\boldsymbol{\beta^{(0)}})$ est deux fois différentiable mais la pénalité $\mathcal{P}(\boldsymbol{\beta})$ ne l'est pas forcement en tous points (la pénalité Bridge pour $\gamma < 2$ est non différentiable en zéro). Fan and Li [2001] proposent d'écrire cette pénalité comme une somme de p pénalités qui dépendent de la valeur absolue afin d'obtenir une formule générale qui correspond aux différentes pénalités Bridge $\gamma > 0$:

$$\mathcal{P}(\boldsymbol{\beta}) = \sum_{j=1}^{p} p_\lambda(|\beta_j|) \tag{2.11}$$

Dans le but de contourner le problème de dérivabilité de la pénalité en zéro, les auteurs proposent une estimation locale à l'aide de l'estimateur initial $\boldsymbol{\beta}_j^{(0)}$. La pénalité $p_\lambda(|\beta_j|)$ est estimée de la manière suivante :

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^{(0)}|) + \frac{1}{2}\dot{p}_\lambda(|\beta_j^{(0)}|)(\beta_j^2 - \beta_j^{(0)2}) \quad j \in 1, \ldots, p \tag{2.12}$$

Où $p_\lambda(|\beta_j^{(0)}|)$ et $\beta_j^{(0)}$ sont des constantes, ce qui permet d'obtenir une équation simplifiée, à une constante C prés :

$$\begin{aligned} p_\lambda(|\beta_j|) &\approx \frac{1}{2}\dot{p}_\lambda(|\beta_j^{(0)}|)\beta_j^2 + C \\ \dot{p}_\lambda(|\beta_j|) &\approx \dot{p}_\lambda(|\beta_j^{(0)}|)\beta_j \end{aligned} \tag{2.13}$$

Comme $\dot{p}_\lambda(|\beta_j^{(0)}|) = \partial p_\lambda(|\beta_j^{(0)}|)/\partial\beta_j^{(0)} = \{\partial p_\lambda(|\beta_j^{(0)}|)/\partial|\beta_j^{(0)}|\}/|\beta_j^{(0)}|$, on obtient l'estimation suivante pour la dérivée de la pénalité :

$$\dot{p}_\lambda(|\beta_j|) = \left\{ \frac{\partial p_\lambda(|\beta_j^{(0)}|)}{\partial|\beta_j^{(0)}|} \times \frac{1}{|\beta_j^{(0)}|} \right\} \beta_j \quad j \in 1, \ldots, p \tag{2.14}$$

En notation matricielle où $\Delta^{(0)} = \mathbf{diag}(\left\{ \frac{\partial p_\lambda(|\beta_1^{(0)}|)}{\partial|\beta_1^{(0)}|} \times \frac{1}{|\beta_1^{(0)}|} \right\}, ..., \left\{ \frac{\partial p_\lambda(|\beta_p^{(0)}|)}{\partial|\beta_p^{(0)}|} \times \frac{1}{|\beta_p^{(0)}|} \right\})$, on obtient :

$$\begin{aligned} \mathcal{P}(\boldsymbol{\beta}) &\approx \frac{1}{2}\boldsymbol{\beta}^T\Delta^{(0)}\boldsymbol{\beta} \\ \dot{\mathcal{P}}(\boldsymbol{\beta}) &\approx \Delta^{(0)}\boldsymbol{\beta} \\ \ddot{\mathcal{P}}(\boldsymbol{\beta}) &\approx \Delta^{(0)} \end{aligned} \tag{2.15}$$

L'équation (2.10) devient alors :

$$U_P(\boldsymbol{\beta}) \approx (U(\boldsymbol{\beta^{(0)}}) - \Delta^{(0)}\boldsymbol{\beta}^{(0)}) + (\dot{U}(\boldsymbol{\beta^{(0)}}) - \Delta^{(0)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}) = 0 \tag{2.16}$$

Plus particulièrement, la j-ème composante de la pénalité Bridge s'écrit $\dot{p}_\lambda(|\beta_j|) = \lambda|\beta_j|^\gamma$. La matrice diagonale $\Delta^{(0)}$ s'écrit alors $\Delta^{(0)} = \mathbf{diag}(\left\{ \lambda\gamma|\beta_1^{(0)}|^{\gamma-2}, \ldots, \lambda\gamma|\beta_p^{(0)}|^{\gamma-2} \right\})$.

2.2.3 Algorithme de calcul

L'équation (2.16) permet d'obtenir une formule récursive pour l'algorithme. Soit $\boldsymbol{\beta}^{(l)}$ l'estimateur connu à l'itération l de l'algorithme, l'estimateur à l'itération suivante est calculé de la façon suivante :

$$\boldsymbol{\beta}^{(l+1)} = \boldsymbol{\beta}^{(l)} - (\dot{U}(\boldsymbol{\beta^{(l)}}) - \Delta^{(l)})^{-1}(U(\boldsymbol{\beta^{(l)}}) - \Delta^{(l)}\boldsymbol{\beta}^{(l)}) \tag{2.17}$$

Si l'on réécrit cette formule pour faire apparaître en bleu les différences avec l'équation (1.21) des GEE on obtient :

$$\hat{\boldsymbol{\beta}}^{(l+1)} = \hat{\boldsymbol{\beta}}^{(l)} - \left(-\mathbf{D}^T\mathbf{V}^{-1}\mathbf{D}\textcolor{blue}{-\Delta}\right)^{-1} \left(\mathbf{D}^T\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu})\textcolor{blue}{-\Delta\boldsymbol{\beta}}\right)\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(l)}} \tag{2.18}$$

Une limite de la méthode LQA est que $\Delta^{(l)}$ n'est pas défini si l'un des coefficients $\beta_j^{(l)}$ est nul et est instable pour $\beta_j^{(l)}$ très proche de zéro. Fan and Li [2001] proposent alors de rejeter à chaque étape de l'algorithme les coefficients vérifiant $|\beta_j^{(l)}| < \eta$ où η est de l'ordre de 0.001. Les coefficients associés sont mis à zéro, les colonnes de $\mathbf{X}$, les lignes de $\boldsymbol{\beta}^{(l)}$, $\boldsymbol{\beta}^{(l+1)}$ et $\Delta^{(l)}$ sont enlevées afin d'éviter les instabilités numériques. Un inconvénient majeur est qu'une fois qu'une variable est associée à un coefficient nul, elle ne peut pas être ré-introduite dans le modèle. Fan and Li [2001] montrent cependant que la méthode réduit considérablement les charges de calcul et obtient de bon résultats sur simulations. Nous utiliserons cette méthode pour les PGEE avec pénalité LASSO.

Cette équation associée aux formules d'estimation des paramètres de corrélations et de variance détaillées dans le chapitre 1 conduisent à l'algorithme du LQA pour PGEE :

Algorithm 2 Penalized Generalized Estimating Equations

Initialize:
$l = 0$
$\boldsymbol{\beta}^{(l)}$, $\boldsymbol{\alpha}^{(l)}$, $\phi^{(l)}$
$diff = 10$
while $diff > \epsilon$ and $l \in \{0, \ldots, l_{max}\}$ **do**
  if $|\beta_j^{(l)}| < \eta$ **then**
    $\beta_j^{(l)} = 0$ et on enlève la variable associée
  end if
  $\mathbf{U}_P^{(l)} = \mathbf{D}^T \mathbf{V}_{(l)}^{-1}(\mathbf{y} - \boldsymbol{\mu}^{(l)}) - \dot{\mathcal{P}}(\boldsymbol{\beta}^{(l)})$
  $\dot{\mathbf{U}}_P^{(l)} = -\mathbf{D}^T \mathbf{V}^{-1(l)} \mathbf{D}^T - \ddot{\mathcal{P}}(\boldsymbol{\beta}_{(l)})$
  $\boldsymbol{\beta}^{(l+1)} = \boldsymbol{\beta}^{(l)} - \dot{\mathbf{U}}_P^{(l)-1} \mathbf{U}_P^{(l)}$
  $diff = ||\boldsymbol{\beta}^{(l+1)} - \boldsymbol{\beta}^{(l)}||_2^2$
  $l = l + 1$
  mise à jour des paramètres $\boldsymbol{\alpha}^{(l)}$ et $\phi^{(l)}$
end while

La valeur initiale $\boldsymbol{\beta}^{(0)}$ est donnée par un estimateur de maximum de vraisemblance ou des moindres carrés en supposant les observations indépendantes. Ce premier estimateur permet de calculer $\boldsymbol{\alpha}^{(0)}$ et $\phi^{(0)}$ par une méthode consistante comme celles proposées dans la section 1.1.3.2.

2.3 Choix du paramètre de régularisation

On cherche à ajuster au mieux les données observées et à obtenir un modèle simple et interprétable. Lorsque beaucoup de variables explicatives sont observées, il est facile d'obtenir un bon ajustement en utilisant un modèle composé de nombreuses covariables. On parle alors de sur-ajustement et de modèle non reproductible. Le modèle ainsi choisi s'ajuste parfaitement aux données observées mais possède de mauvaises capacités de prédiction puisque la structure des données est mal identifiée.

Afin d'éviter cette problématique, le paramètre de régularisation λ doit être choisi minimisant un critère spécifique. L'objectif est de trouver le paramètre qui correspond au meilleur modèle au sens d'un critère choisi.

2.3.1 Critères de type somme des résidus au carré

Les critères de la section 1.2.1 pour une réponse continue ou binaire, estimés par validation croisée en LOO s'écrivent :

$$LOO_Q(\lambda) = \sum_{i=1}^{K} \frac{(\mathbf{y}_i - \hat{\mathbf{y}}_i^{-i})^T \hat{\mathbf{Q}}_i^{-1} (\mathbf{y}_i - \hat{\mathbf{y}}_i^{-i})}{T_i} \tag{2.19}$$

Où $\hat{\mathbf{y}}^{-i}$ représente le vecteur de réponse de l'individu i estimé sans utiliser l'individu i. La matrice $\hat{\mathbf{Q}}_i$ peut être une matrice identité de taille $T_i \times T_i$, la matrice diagonale des variances marginales $\hat{\mathbf{A}}_i$, la matrice de corrélation $\hat{\mathbf{R}}_i$ ou de variance covariance $\hat{\mathbf{V}}_i$ de travail estimée sur le modèle complet (i.e. non pénalisé). Le calcul de ce critère sur une grille de λ peut être long. Afin de raccourcir les temps de calcul il est possible d'utiliser la validation croisée en 5 groupes.

Une autre façon de réduire les temps de calcul est d'utiliser la méthode de validation croisée généralisée (Tibshirani [1996]; Fu [1998]). Ce critère utilise le nombre effectif de paramètres du modèle défini par :

$$p(\lambda, \gamma) = p^* \times s \tag{2.20}$$

Il se compose du nombre de coefficients non nuls p^* multiplié par le taux de réduction s introduit par Tibshirani [1996] pour le cas particulier du LASSO. Généralisé aux pénalités Bridge par Fu [2003, 2005] il devient :

$$s = \frac{||\hat{\boldsymbol{\beta}}_{(\lambda,\gamma)}||_\gamma}{||\hat{\boldsymbol{\beta}}_{GEE}||_\gamma} \tag{2.21}$$

où $||.||_\gamma$ représente la norme L_γ pour $\gamma \geq 1$. $\hat{\boldsymbol{\beta}}_{(\lambda,\gamma)}$ est obtenue par PGEE alors que l'estimateur $\hat{\boldsymbol{\beta}}_{GEE}$ est obtenu par GEE classique. Cette quantité correspond à l'écart entre le modèle classique et le modèle pénalisé. La validation croisée généralisée adaptée aux GEE s'écrit alors :

$$GCV(\lambda) = \frac{WRSS_{\mathbf{R}}(\lambda)}{K(1 - p(\lambda, \gamma)/N)^2} \tag{2.22}$$

où la WRSS peut être remplacée par une déviance pondérée pour les réponses discrètes. On utilise alors les résidus de déviance $\tilde{r}_{i,t}$:

$$WDev = \sum_{i=1}^{K} \tilde{\mathbf{r}}_i^T \mathbf{R}_i^{-1} \tilde{\mathbf{r}}_i \tag{2.23}$$

Fu [2005] propose une quasi-validation croisée généralisée (QGCV) :

$$QGCV(\lambda) = \frac{WDev(\lambda)}{K\{1 - p(\lambda, \gamma)/\tilde{n}\}^2} \tag{2.24}$$

$\tilde{n}$ représente le nombre d'observations corrigé par l'importance des corrélations : $\tilde{n} = \sum_i \frac{T_i^2}{\sum_{i,t} \mathbf{R}_{i,t}}$ comme défini dans la section 1.2.4. Plus les corrélations sont importantes, plus le nombre effectif d'observations $\tilde{n}$ est proche du nombre d'individus K, plus les corrélations sont faibles plus l'on se rapproche du nombre total d'observations N. Si les corrélations sont négatives, $\tilde{n}$ peut dépasser le nombre d'observations mais ce cas est rare.

2.3.2 Critères de type AIC et BIC

Un autre type de critère peut être utilisé afin de minimiser les temps de calcul pour le choix du paramètre λ : les critères de type AIC et BIC. La version généralisée de ces critères pour les GEE est détaillée dans la section 1.2. Le degré de liberté qu'ils font intervenir doit être adapté à la pénalité choisie. Il peut être estimé par $p(\lambda, \gamma)$ comme dans le cadre de la validation croisée généralisée, ou par les formules proposées par Pan [2001a] et Cantoni et al. [2005] (section 1.2). On obtient alors une multitude de critères possibles pour choisir le paramètre de régularisation λ. La performance des différents critères ainsi obtenus est testée empiriquement par Dziak and Li [2006] qui n'observent pas de grandes différences entre les définitions.

Conclusion Les régressions pénalisées adaptées aux GEE permettent de sélectionner le *meilleur* sous-groupe de variables au sens du critère de qualité choisi sans tester tous les modèles possibles. Elles proposent une méthode de sélection (pour $1 \leq \gamma < 2$) plus stable ainsi qu'un gain de temps pour le choix du sous groupe le plus en lien avec la réponse. Malheureusement, ces méthodes supposent que les variables sont observées aux mêmes instants. Les données réelles présentent souvent des données non renseignées, une extension des ces méthodes aux données manquantes est donc nécessaire.

Chapitre 3

Étude de l'impact de données manquantes sur l'estimateur des GEE

Les données manquantes font référence à des données qui auraient dû être collectées mais qui ne l'ont pas été en opposition aux designs déséquilibrés pour lesquels l'absence d'information est prévue. Les raisons pour lesquelles les données n'ont pas pu être collectées doivent être étudiées et le mécanisme des données manquantes doit être pris en compte dans les analyses statistiques afin de comprendre pourquoi ces données sont manquantes et leur impact sur les inférences, interprétations et conclusions.

Lorsqu'une partie des données est manquante, l'analyste fait inévitablement face à une perte d'informations et une réduction de la précision avec laquelle les paramètres d'intérêt peuvent être estimés. Cette réduction de la précision est directement liée à la quantité de données manquantes et est influencée, jusqu'à un certain point, par la méthode d'analyse (Molenberghs et al. [2014]).

Nous proposons dans ce chapitre d'étudier l'impact de différents types de données manquantes sur les GEE. Notre première partie propose un état de l'art succinct des méthodes pour prendre en compte les sorties d'étude, tandis que notre deuxième partie étudie l'effet de visites manquantes intermittentes sur l'estimateur. La troisième partie de ce chapitre présente les méthodes d'imputation pour le traitement des données manquantes ponctuelles. Plus particulièrement, la quatrième partie étudie le cas des variables soumises

à un seuil de détection pour lesquelles nous proposons une nouvelle fonction d'imputation à utiliser avec le package `mice` de R (Van Buuren and Groothuis-Oudshoorn [2011]). Une étude par simulations est proposée afin d'évaluer l'estimateur obtenu.

3.1 Taxonomie des données manquantes

Le mécanisme des valeurs manquantes Une taxonomie a été mise en place à la fin des années 70 afin d'expliciter les différents profils de données manquantes possibles, leurs caractéristiques et leurs impacts sur les analyses statistiques (Tableau 3.1).

Missing Completely at Random (MCAR)
Données manquantes indépendantes de l'état du patient, des covariables observées, des facteurs socio-démographiques **Données manquantes complètement aléatoires**
Missing at Random (MAR)
Données manquantes qui dépendent de la réponse observée et de certaines covariables **Données manquantes aléatoires**
Missing Not at Random (MNAR)
Données manquantes qui dépendent de la réponse non observée **Données manquantes non aléatoires**

TABLE 3.1 – Taxonomie des données manquantes selon la classification de Rubin [1976] et Little and Rubin [1987]

Le modèle de données manquantes On retrouve différents types de données manquantes dans les études comme des données manquantes ponctuelles, des visites manquantes intermittentes et des sorties d'étude. Dans le premier cas, le patient a été contacté à la visite t mais pour certaines variables, l'information est manquante. Dans le second cas, le patient ne vient pas à la visite t mais peut revenir à des visites ultérieures $t' > t$. Pour les sorties d'étude il ne revient pas aux visites suivantes, il quitte l'étude et nous perdons l'information que nous aurions du mesurer pour toutes les visites $t' \geq t$. Pour les deux derniers cas, il manque les mesures de toutes les variables dépendantes du temps pour ce patient à chacune des visites non effectuées. Dans un contexte où de nombreuses variables sont mesurées, cela veut dire que l'on perd beaucoup d'informations. Il existe d'autres types de données manquantes mais nous ne traitons ici que les typologies que l'on rencontre dans

nos bases de données :

- Patients sortis d'étude
- Visites manquantes intermittentes
- Données manquantes partielles sur les visites effectuées (*ponctuelles*)
- Variables soumises à un seuil de détection

3.2 Patients sortis d'étude

Supposons que l'on observe une réponse $y_{i,t}$ qui dépend d'un vecteur de covariables $\mathbf{x}_{i,t}$ ainsi qu'un vecteur de covariables auxiliaires $\mathbf{z}_{i,t}$ pour tout individu i au temps t. On parle d'attrition lorsque l'on perd certains patients au cours du temps. Le vecteur réponse se note alors $\mathbf{y}_i = (\mathbf{y}_i^o, \mathbf{y}_i^m)$ où $\mathbf{y}_i^o$ représente la partie observée du vecteur et $\mathbf{y}_i^m$ représente la partie manquante du vecteur. Le vecteur $\mathbf{P}_i$ indique le fait d'être observé (i.e. $P_{i,t} = 1$ si $y_{i,t} \in \mathbf{y}_i^o$) ou d'être manquant (i.e. $P_{i,t} = 0$ si $y_{i,t} \in \mathbf{y}_i^m$).

Mécanisme de l'attrition L'intérêt porte alors sur la distribution du vecteur $\mathbf{P}_i = (P_{i,t})_{t=1,\ldots,T_i}$ qui représente la distribution des données manquantes (DOM). La classification donnée dans le tableau 3.1 peut être affinée en quatre classes (Little [1995]) pour obtenir la classification de la figure 3.1. On parlera d'attrition complètement aléatoire lorsque le fait d'être sorti d'étude ne dépend que de covariables auxiliaires comme le fait d'avoir déménagé. Si l'attrition dépend aussi de covariables du modèle de régression de la réponse $\mathbf{y}$ on dira qu'elle dépend de covariables (CD) comme par exemple une attrition plus importante chez les femmes alors que le sexe est un facteur de risque de la maladie. Lorsque les sorties d'étude sont liées aux valeurs observées de la réponse aux temps précédents, on parle d'attrition MAR. Imaginons un cas où les patients "sur la mauvaise pente" - chez qui il a été mesuré une mauvaise réponse aux temps précédents - présentent une plus grande attrition que les autres. Les mesures effectuées jusqu'à la sortie d'étude permettent de reconnaître ces patients. Si par contre, le fait de sortir de l'étude au temps t dépend de la valeur que l'on aurait dû observer au temps t on parle d'attrition MNAR. Les patients

souffrant d'un déclin cognitif déclaré entre deux visites ne sont plus capables de remplir les questionnaires. La donnée est manquante jusqu'à la fin de l'étude et la trajectoire observée ne permet pas de le prédire.

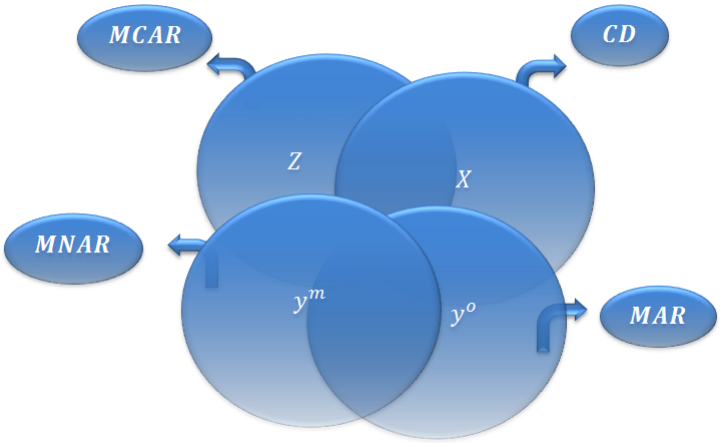


FIGURE 3.1 – Classification de Rubin [1976], Little and Rubin [1987] et Little [1995]

Cette typologie se traduit par les DOM suivant :

$$\begin{aligned} MCAR &: \mathbb{P}(\mathbf{P}_i = \mathbf{p}_i|\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i) = \mathbb{P}(\mathbf{P}_i = \mathbf{p}_i|\mathbf{Z}_i) \\ CD &: \mathbb{P}(\mathbf{P}_i = \mathbf{p}_i|\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i) = \mathbb{P}(\mathbf{P}_i = \mathbf{p}_i|\mathbf{X}_i, \mathbf{Z}_i) \\ MAR &: \mathbb{P}(\mathbf{P}_i = \mathbf{p}_i|\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i) = \mathbb{P}(\mathbf{P}_i = \mathbf{p}_i|\mathbf{y}_i^o, \mathbf{X}_i, \mathbf{Z}_i) \\ MNAR &: \mathbb{P}(\mathbf{P}_i = \mathbf{p}_i|\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i) = \mathbb{P}(\mathbf{P}_i = \mathbf{p}_i|\mathbf{y}_i^m, \mathbf{y}_i^o, \mathbf{X}_i, \mathbf{Z}_i) \end{aligned} \tag{3.1}$$

Données manquantes ignorables Le premier objectif de cette classification est de savoir pour quelles conditions les données manquantes sont *ignorables* ce qui dépend des méthodes statistiques utilisées et des hypothèses sous-jacentes. Les méthodes fréquentistes, comme les GEE, reposent souvent sur l'hypothèse MCAR et ne sont pas robustes aux données de types MAR et MNAR. Liang and Zeger [1986] remarquent que, même si la matrice de corrélation de travail est mal spécifiée, l'estimateur des GEE et sa matrice de variance covariance sont consistants sous l'hypothèse d'attrition MCAR. Cependant, si les sorties d'étude sont MAR, et même si la matrice $\mathbf{R}_i$ est correcte, l'estimateur est consistant mais

l'estimateur de sa variance ne l'est pas toujours (Kenward and Molenberghs [1998]).

Les données manquantes de type MCAR sont ignorables au sens de Rubin [1976], Little and Rubin [1987] et Little [1995] puisque ce type de données manquantes n'entraîne pas de biais dans l'analyse du lien entre covariables et réponse. Si certaines covariables du modèle de régression sont liées au mécanisme des données manquantes, ou si les données sont de type MAR, elles sont non ignorables puisque l'estimateur des GEE sera biaisé si l'on ne prend pas correctement en considération le processus de données manquantes (Liang and Zeger [1986]; Schafer and Graham [2002]). Cependant de nombreuses techniques ont été développées pour prendre en compte les données MAR dans les analyses statistiques. Les données de type MNAR sont informatives et peuvent donc entraîner un biais important. Malheureusement, cette typologie de données manquantes est difficile à mettre en évidence puisqu'elle dépend d'une valeur non observée. Il est possible toutefois de prendre en compte ces données en faisant des hypothèses de distribution et en vérifiant l'impact de ces hypothèses par analyses de sensibilité (Molenberghs et al. [2014]).

Méthodes usuelles Si l'on estime $\hat{\boldsymbol{\beta}}$ à l'aide des GEE sur les données observées on obtient les équations suivantes :

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{K} \mathbf{D}_i^T \mathbf{V}_i^{-1} \boldsymbol{\Delta}_i [\mathbf{y}_i - \boldsymbol{\mu}_i] = 0 \tag{3.2}$$

Où $\boldsymbol{\Delta}_i = \mathbf{diag}(\mathbf{P_{i,1}}, \ldots, \mathbf{P_{i,T_i}})$ est la matrice diagonale des indicateurs de présence. Cette analyse ignore les données manquantes et traite les données observées comme complètes. En général, l'estimateur ne sera consistant que si les données manquantes sont MCAR, dans ce cas là, $\mathbb{E}(\boldsymbol{\Delta}_i[\mathbf{y}_i - \boldsymbol{\mu}_i]) = 0$ puisque $\boldsymbol{\Delta}_i$ et $\mathbf{y}_i$ sont indépendantes. Pour des données MAR en revanche, ces quantités sont corrélées et $\mathbb{E}(\boldsymbol{\Delta}_i[\mathbf{y}_i - \boldsymbol{\mu}_i]) \neq 0$.

Lorsque les données manquantes ne concernent que la réponse on peut utiliser la méthode du Last Observation Carried Forward (LOCF) ou l'on remplace les valeurs man-

quantes de la réponse par la dernière valeur observée. Cette méthode, très utilisée dans l'industrie pharmaceutique, est aujourd'hui fortement déconseillée. Il est possible d'améliorer cette méthode en utilisant une imputation de type "Hot-deck" où les valeurs non observées sont remplacées par des données observées sur une population similaire. Cette méthode peut être répétée M fois afin d'obtenir de multiples imputations (Rubin [1987]). Ces deux méthodes supposent que la valeur est manquante uniquement pour la réponse, mais dans notre contexte où les covariables sont dépendantes du temps ce n'est pas le cas. Les méthodes d'imputation ne seront donc pas utilisées pour combler les sorties d'étude dans ce travail de thèse. Toutefois, il existe de nombreuses méthodes pour prendre en compte les sorties d'étude lorsque les données manquantes ne concernent que la réponse (Carpenter et al. [2006]; Diggle et al. [2007]; Molenberghs et al. [2014]...).

Les méthodes Inverse Probability Weighting (IPW) Une façon intuitive de réduire le biais induit par les visites non observées est de pondérer judicieusement les données observées pour que leur score soit plus proche du score de l'échantillon total ou de la population. Cette idée se retrouve dans de nombreux articles parmi lesquels Robins et al. [1994], Robins et al. [1995] et Rotnitzky et al. [1998]. Dans cette littérature, les lignes (i.e. une visite d'un patient) observées sont pondérées par l'inverse de la probabilité que le patient effectue cette visite afin qu'elles soient plus représentatives de la population globale. De cette manière, les sujets qui représentent une classe de patients à risque de sortir de l'étude auront une probabilité faible de rester dans l'étude et donc un poids important dans l'analyse.

Cette méthode nécessite deux modèles :

- un modèle pour relier la réponse aux variables explicatives : GEE
- un modèle pour estimer les probabilités de sortie d'étude : régression logistique

Les GEE pondérées ou WGEE sont solutions de :

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{K} \mathbf{D}_i{}^T \mathbf{V}_i{}^{-1} \mathbf{W}_i [\mathbf{y}_i - \boldsymbol{\mu}_i] = 0 \tag{3.3}$$

Où $\mathbf{W}_i$ est une matrice diagonale de taille $T_i \times T_i$ composée des poids $P_{i,t}/\omega_{i,t}$. Ces poids représentent l'inverse de la probabilité de rester dans l'étude $\omega_{i,t} = \mathbb{P}(P_{i,t} = 1|obs)$. Selon les hypothèses faites et le modèle de distribution de données manquantes, *obs* peut contenir des covariables fixes au cours du temps, des covariables dépendantes du temps, des covariables mesurées à baseline, le vecteur de réponses observées aux temps précédents... En pratique on estime les probabilités $\lambda_{i,t}$ pour $t \in \{1, \ldots, T_i\}$ où $\lambda_{i,1} = 1$ puisque tous les individus effectuent la visite 1 puis $\lambda_{i,t} = \mathbb{P}(P_{i,t} = 1|P_{i,(t-1)} = 1, obs)$ est estimée grâce à une régression logistique. On obtient alors une estimation pour la matrice de poids où $\hat{\omega}_{i,\tau} = \prod_{t=1}^{\tau} \hat{\lambda}_{i,t}$ puis on calcule la solution à l'équation (3.3), on retrouve $\mathbb{E}(P_{i,t}\omega_{i,t}^{-1}[y_{i,t} - \mu_{i,t}]) = 0$.

Cette méthode est facilement utilisable grâce à la procédure GEE dans SAS (Lin and Rodriguez [2014]) et au package `ipw` dans R (van der Wal et al. [2011]). Toutefois, seule la procédure GEE permet d'estimer correctement la matrice de variance covariance issue d'une telle estimation. Dans le package `geepack` de Halekoh et al. [2006], les poids ne sont pas bien intégrés dans le calcul de la matrice de variance covariance et les intervalles de confiance sont conservateurs.

Les estimateurs issus des WGEE sont consistants mais biaisés sur des échantillons finis. En pratique cela ne pose pas de problème pour les analyses mais cette méthode suppose que le modèle de distribution choisi pour les données manquantes (DOM) soit consistant. Si cette modélisation est mauvaise, l'estimateur peut être sévèrement biaisé (Lipsitz et al. [2000]; Preisser et al. [2002]). Pour améliorer les performances de cet estimateur, une version robuste qui nécessite de spécifier un modèle pour les données observées sachant les données manquantes a été développée (Scharfstein et al. [1999]; Davidian et al. [2005]; Tsiatis [2007]). Afin de restreindre le champ des recherches, nous supposerons que notre

modélisation du DOM est correcte et utiliserons uniquement les WGEE simples dans nos analyses en comparant les résultats obtenus avec différents DOM.

Remarques Nous supposerons dans ce travail que les patients sortis d'étude sont issus des trois premières possibilités. Nous n'excluons pas la possibilité d'une attrition MNAR mais nous supposerons qu'elle n'est pas observée dans nos données afin de nous concentrer sur le développement d'une méthode de sélection capable d'intégrer les données manquantes.

3.3 Les visites manquantes intermittentes

Les méthodes de types IPW peuvent être utilisées pour prendre en compte les visites manquantes intermittentes. Une régression logistique pour estimer la probabilité de manquer une visite peut être ajustée afin de réduire le biais dû à des visites manquantes intermittentes de type MAR. Dans les études cliniques que nous analysons, les visites manquantes intermittentes sont rarement liées à l'état du patient et il est difficile de trouver un modèle de régression logistique qui les explique. Nous supposerons donc que ces visites manquantes - qui représentent souvent un faible pourcentage - sont MCAR ou dépendantes du temps (plus on se rapproche de la fin de l'étude plus le patient est à risque de manquer une visite). Une étude par simulations de l'effet de telles visites manquantes intermittentes sur l'estimateur des GEE a été mise en place.

3.3.1 Le protocole de simulations

Nous avons étudié (Geronimi and Saporta [2015]) l'effet de visites manquantes intermittentes sur une réponse continue $\mathbf{y}_i$ simulée selon le modèle suivant :

$$\mathbf{y}_i = \beta_0 + \sum_{j=1}^{4} \mathbf{x}_{i,j}\beta_j + \boldsymbol{\epsilon}_i \tag{3.4}$$

où $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) = (1, 0.5, -0.2, 1, -1)$ est le vecteur de coefficients associés

aux covariables $\mathbf{x}_j$ gaussiennes $j \in \{1, \ldots, 4\}$, centrées réduites auto-régressives d'ordre 1 de coefficient 0.3. Le vecteur d'erreurs $\boldsymbol{\epsilon}_i$ est une gaussienne centrée multivariée de matrice de variance covariance $\sigma^2 \mathbf{R}_i(\alpha)$ où $\mathbf{R}_i(\alpha)$ est la matrice de corrélation choisie et σ^2 est le paramètre de variance choisi pour avoir un rapport signal-bruit égal à 1. Nous avons fait varier différents paramètres :

- K le nombre de sujets sur $\mathcal{K} = \{50, 100, 200, 300\}$
- T le nombre de visites prévues $\mathcal{T} = \{4, 6, 9\}$
- $\mathbf{R}_i(\alpha)$ la matrice de corrélation : soit auto-régressive, soit échangeable de paramètre $\alpha \in \mathcal{A} = \{0.1, 0.3, 0.5, 0.6\}$

Sur le jeu simulé complet nous avons imposé des données manquantes sur un certain pourcentage po de la population $po \in \{10\%, 20\%, 30\%, 50\%\}$ chez lesquels nous avons supprimé n visites $n \in \{1, 2, 3\}$. Chacun de ces 96 scénarios est répliqué 1000 fois. Le premier schéma de visites manquantes est MCAR où les patients et les visites sont choisis selon une distribution uniforme sur leur domaine. Puis, pour étudier l'effet des visites manquantes qui sont plus importantes au cours du temps, nous avons imposé le schéma suivant : tous les individus sont présents dans l'étude à la première visite puis, pour chaque individu choisi aléatoirement, on tire une visite à supprimer comme indiqué sur la figure 3.2 :

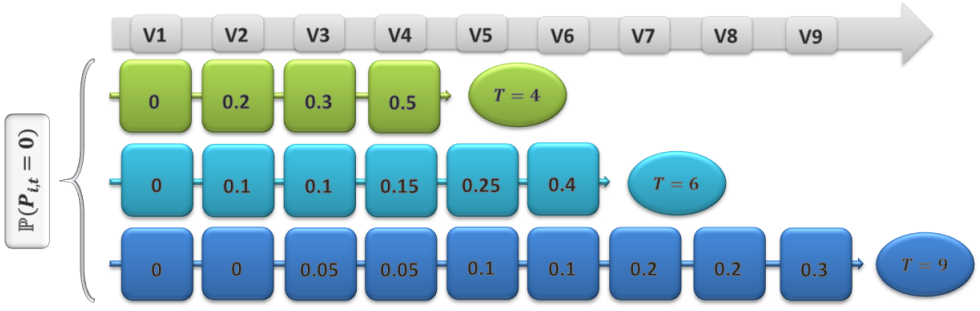


FIGURE 3.2 – Probabilité $\mathbb{P}(P_{i,t} = 0)$ que le patient i manque la visite t selon le nombre de visites prévues

Nous avons testé d'autres scénarios comme par exemple le cas où seulement les trois der-

nières visites sont touchées par des visites manquantes intermittentes et le cas où seulement les premières visites sont concernées. De même, nous avons effectué les mêmes simulations avec une réponse binaire où la structure de corrélation est assurée par la méthode de Qaqish [2003]. Les résultats étant très similaires, nous les présentons en Annexe A.

3.3.2 Les résultats

Le vecteur de coefficients $\boldsymbol{\beta}$ est estimé par GEE comme indiqué dans l'équation (1.13) avec une matrice de corrélation de travail $\mathbf{R}_i(\alpha)$ auto-régressive d'ordre 1 à l'aide du package `geepack` de Halekoh et al. [2006]. Nous comparons les deux schémas de données manquantes à l'aide du Biais Absolu Relatif (ARB) estimé sur 1000 réplications défini par :

$$ARB = \frac{1}{B}\sum_{b}\frac{||\hat{\boldsymbol{\beta}}^{(b)} - \boldsymbol{\beta}||_2}{||\boldsymbol{\beta}||_2} \tag{3.5}$$

où la norme est la norme euclidienne pour un vecteur qui se réduit à la valeur absolue pour un scalaire. Ce critère peut être vu comme un pourcentage, ce qui le rend facilement interprétable. La figure 3.3 représente l'évolution de L'ARB en fonction du taux de données manquantes. Chaque colonne représente la répartition du biais relatif absolu pour les 96 modèles testés. Le déséquilibre uniforme représente les visites manquantes intermittentes MCAR et le déséquilibre croissant représente le cas où la probabilité de manquer une visite est croissante en fonction du temps.

Le critère augmente faiblement avec le taux de données manquantes passant d'un ARB médian de 0.091 sur le jeu complet à 0.101 lorsque l'on supprime 3 visites chez 50% des patients. Les ARB les plus importants sont observés pour les jeux de données composés de peu de patients ou de peu de visites où ce critère peut atteindre plus de 20%. La figure 3.4 représente l'évolution de L'ARB en fonction du taux de données manquantes pour $K = 100$ pour le scénario d'un déséquilibre croissant.

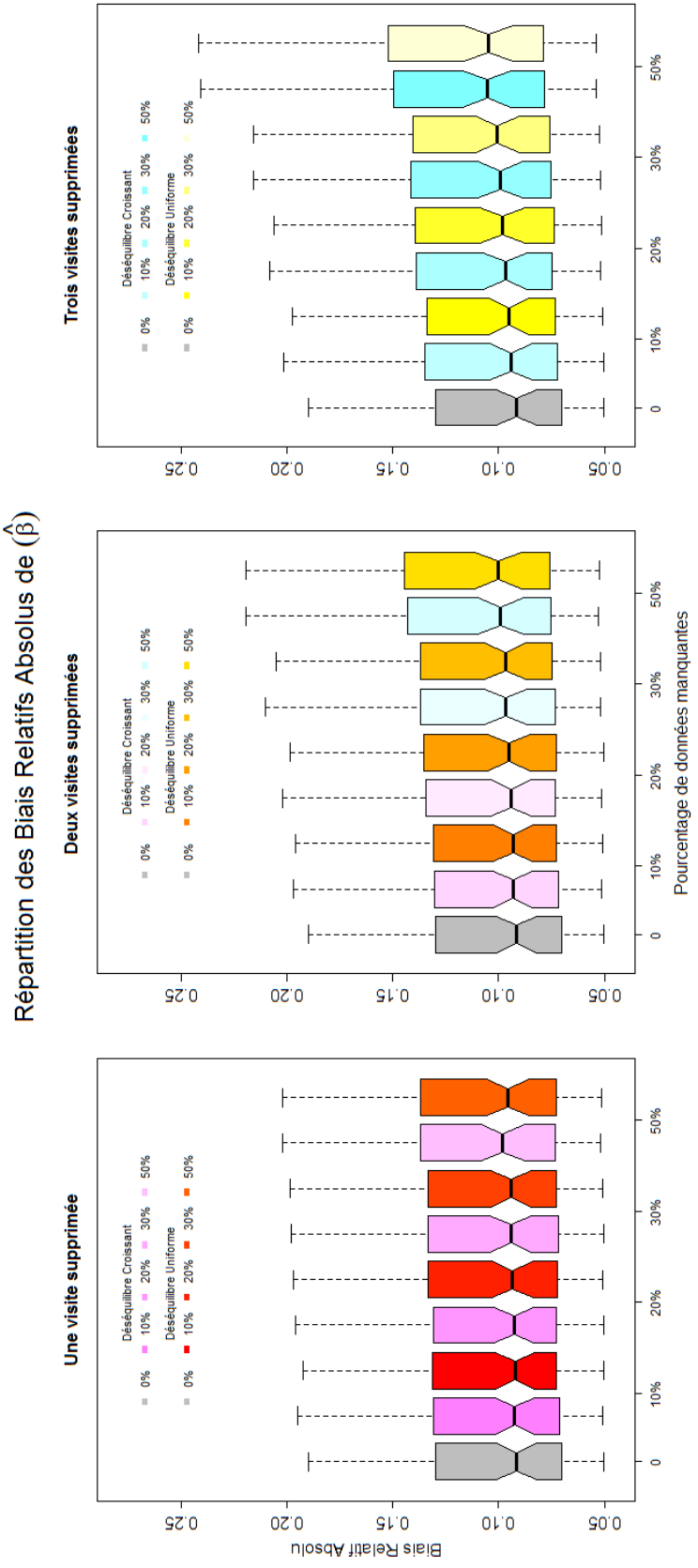


FIGURE 3.3 – Évolution de l'$ARB(\hat{\boldsymbol{\beta}})$ en fonction du pourcentage de données manquantes pour une, deux, ou trois visites supprimées et deux types de données manquantes

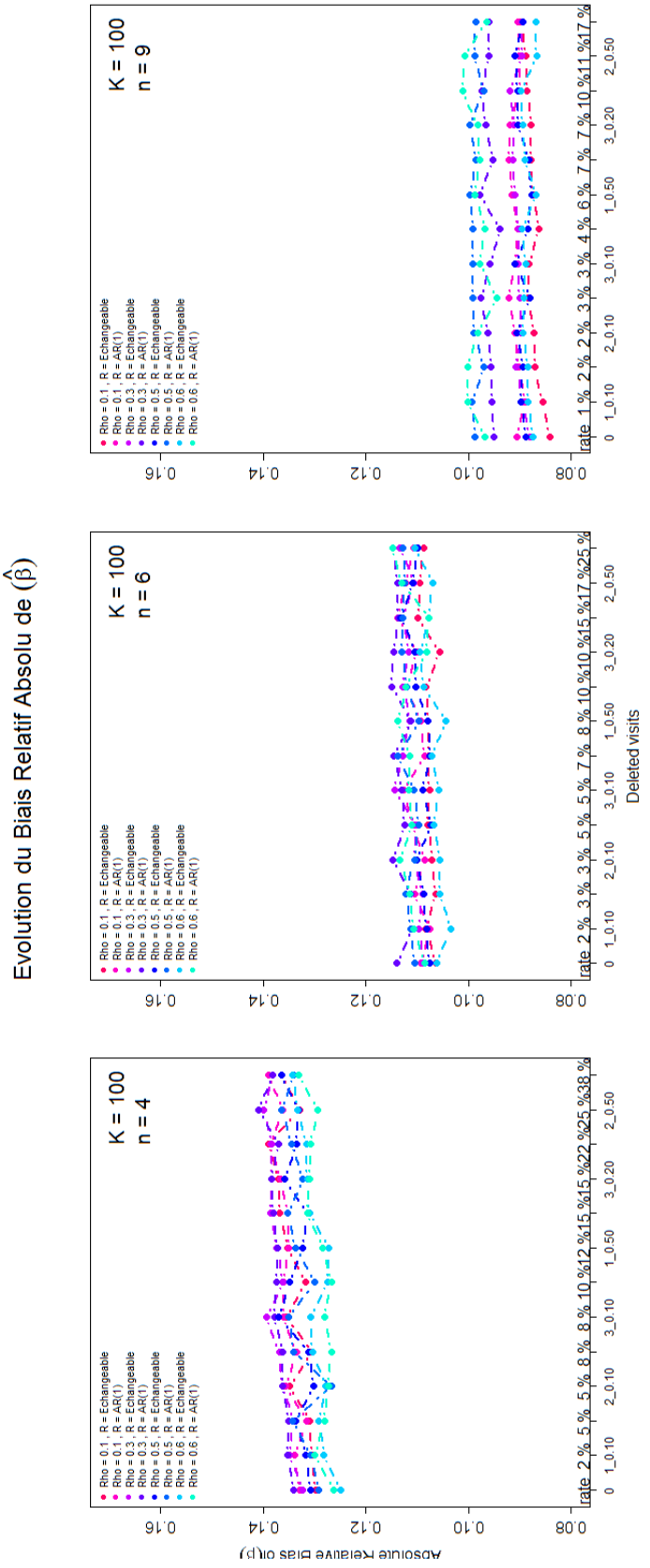


FIGURE 3.4 – Évolution de l'$ARB(\hat{\boldsymbol{\beta}})$ en fonction du pourcentage de données manquantes pour $K = 100$ avec un déséquilibre croissant

On constate que l'ARB n'est que peu modifié par le pourcentage de données manquantes. Il en est de même pour les différents critères que nous avons fait varier. Le seul critère qui semble influer la qualité de l'estimation est le nombre de visites prévues. Ce résultat est attendu puisque qu'il signifie que plus l'échantillon est grand (nombre de visites prévues mais aussi nombre de patients inclus dans l'étude) plus l'estimation est précise.

Sur ces jeux de données simulées, l'estimateur obtenu par GEE est robuste aux données manquantes imposées et les résultats sont très similaires entre les deux schémas de données manquantes imposées. De ce fait, l'estimateur tend à être robuste aux visites manquantes intermittentes qui dépendent du temps. Les résultats plus détaillés figurent en Annexe D.1.

Remarques Les bases de données utilisées pour nos applications sont composées de nombreuses covariables mesurées à différents temps. Nous avons donc choisi de ne pas utiliser de méthodes d'imputation pour ce type de perte d'informations afin de ne pas prendre le risque d'introduire trop de bruit dans nos données complétées. Nous supposerons dans nos analyses que les données manquantes intermittentes sont tout au plus dépendantes du temps et que l'attrition est MCAR ou MAR. Une méthode de type IPW pourra être envisagée afin d'éviter le biais induit par une attrition MAR.

3.4 La problématique des données manquantes ponctuelles

La troisième forme de données manquantes présente dans nos bases de données concerne les données manquantes ponctuelles. Un patient qui a effectué la visite t peut présenter quelques données manquantes sur quelques covariables ou sur la réponse.

3.4.1 Les méthodes usuelles

Plusieurs possibilités s'offrent à nous devant de telles données manquantes :

Cas complet et IPW La première solution est de supprimer les lignes qui présentent des données manquantes et de réaliser l'analyse d'intérêt. Cette méthode est fortement déconseillée puisqu'elle est synonyme de beaucoup de perte d'informations et peut biaiser l'analyse. Il est aussi possible d'utiliser une méthode de type IPW après avoir supprimé ces lignes Robins et al. [1994] afin de réduire le biais que peut induire cette perte. Cependant, comme une visite supprimée revient à supprimer l'information de centaines de variables nous avons préféré utiliser les méthodes d'imputation.

Imputation simple Les méthodes d'imputation simple permettent d'attribuer une valeur à une donnée manquante, puis de réaliser l'analyse d'intérêt sur la base ainsi complétée. Parmi les méthodes utilisées, nous pouvons citer l'imputation par la moyenne ou la médiane qui peuvent être des imputations par groupe puisqu'il est possible de stratifier la population : par exemple faire des imputations par âge, sexe ou en ajustant sur d'autres covariables. Une autre amélioration est d'ajouter du bruit dans l'imputation. Par exemple, un terme d'erreur peut être ajouté à la moyenne afin d'éviter les imputations déterministes qui réduisent la variabilité. Ces méthodes ne prennent malheureusement pas toujours en compte la variabilité induite par l'imputation et créent des écarts types trop faibles (Little and Rubin [1987]). De plus, la structure de la base comme les corrélations ou le lien entre les variables ne sont pas prises en compte dans l'imputation. L'imputation "Hot-deck" qui remplace les données manquantes par des données observées chez les plus proches voisins (à définir) permet de mieux respecter la structure de la base de données mais sous-estime la variance (Rao [1996]).

Pour les méthodes exploratoires d'analyses de données, les méthodes d'analyse factorielles peuvent être adaptées aux données manquantes. Pour ce faire, Josse and Husson [2012] proposent d'utiliser une Analyse en Composantes Principales (ACP) itérative. Cette dernière se comporte comme un algorithme Expectation Minimization (EM) ce qui permet d'estimer les axes et composantes principales en présence de données manquantes. Plus récemment, Audigier et al. [2016] ont proposé le BayesMIPCA. Cet algorithme utilise une ACP bayésienne et permet de réaliser une imputation mutiple en sélectionnant M jeux de données approximativement indépendants simulés grâce aux posteriors. Cette méthode permet de traiter la problématique de la grande dimension où le nombre de covariables est supérieur au nombre d'observations. Cependant deux limites de cette méthode ne nous permettent pas de l'utiliser : elle se restreint aux données continues, et ne permet pas d'imputer les données MNAR.

3.4.2 Différentes méthodes d'Imputation Multiple

Nous avons choisi d'utiliser l'imputation multiple pour sa facilité d'utilisation, de compréhension et sa bonne implémentation sur les logiciels standards. Cette méthode, introduite par Rubin [1987] est aujourd'hui largement utilisée et permet de prendre en compte la variabilité des imputations. L'idée principale est décrite par la figure 3.5. On utilise une règle d'imputation aléatoire afin d'obtenir M jeux complets sur lesquels on réalise l'analyse d'intérêt afin d'estimer notre vecteur de coefficients $\boldsymbol{\beta}$. L'estimateur combiné $\overline{\boldsymbol{\beta}}$ et sa variance sont estimés à l'aide des règles de Rubin :

$$\overline{\boldsymbol{\beta}} = \frac{1}{M} \sum_{m=1}^{M} \hat{\boldsymbol{\beta}}_m \tag{3.6}$$

où $\hat{\boldsymbol{\beta}}_m$ est l'estimateur de $\boldsymbol{\beta}$ sur le m-ème jeu de données imputées $m \in \{1, \ldots, M\}$ estimé par GEE à l'aide de l'équation (1.13). La variance $\boldsymbol{\Sigma}$ de cet estimateur combiné est composée d'un terme de variance intra-imputation $\overline{\mathbf{W}} = \frac{1}{M} \sum_{m=1}^{M} \hat{\mathbf{V}}_{\mathbf{R},m}$ où $\hat{\mathbf{V}}_{\mathbf{R},m}$ est la matrice de variance covariance estimée de $\hat{\boldsymbol{\beta}}_m$ à l'aide de l'équation (1.24) et d'un terme

de variance inter-imputation $\mathbf{B} = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\boldsymbol{\beta}}_m - \overline{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_m - \overline{\boldsymbol{\beta}})^T$. L'estimateur combiné de la variance est donné par :

$$\hat{\boldsymbol{\Sigma}} = \overline{\mathbf{W}} + \frac{M+1}{M}\mathbf{B} \tag{3.7}$$

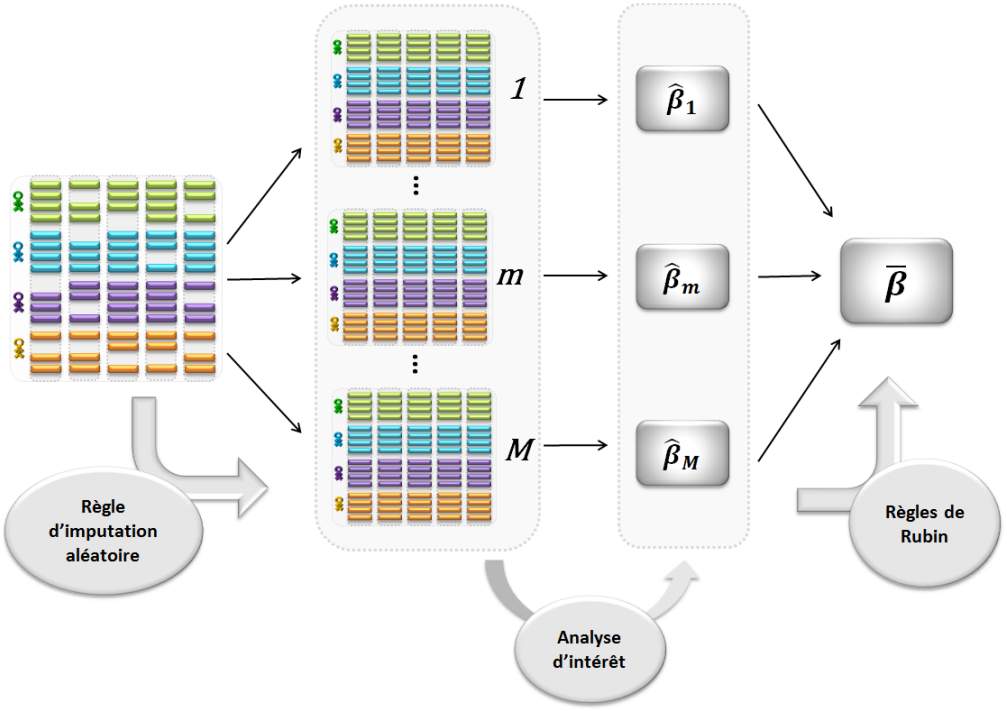


FIGURE 3.5 – Imputation multiple

La littérature sur le sujet est riche (Schafer [1999]; Rubin [1996]; Sterne et al. [2009]; Kenward and Carpenter [2007]) et différentes méthodes aléatoires permettent d'obtenir ces M imputations. Deux méthodes sont très répandues : la méthode de modélisation jointe (Schafer [1997]) et la méthode d'imputation par équations en chaine (Van Buuren [2012]). La première cherche à modéliser les variables conjointement ce qui la rend difficilement applicable en présence de bases de données complexes qui sont composées de différents types de variables (Van Buuren [2007]; Molenberghs et al. [2014]). La deuxième, que l'on retrouve aussi sous le nom de *sequential regression multiple imputation* et *imputation by full conditional specification* utilise des modèles conditionnels séparés pour chaque variable

sachant les autres.

3.4.3 Imputation par équations en chaîne

Nous avons choisi d'utiliser l'imputation par équations en chaîne afin de respecter l'hétérogénéité de nos bases de données qui présentent souvent de nombreuses covariables de différents types (continues, binaires, classes...). L'idée est la suivante : notons $\mathbf{X}$ un échantillon composé de p variables soumises aux données manquantes. Les données manquantes sont d'abord complétées par une première imputation obtenue grâce aux distributions marginales, par exemple imputation par la moyenne. L'algorithme utilise ensuite une concaténation de procédures univariées. On suppose que chaque variable $\mathbf{x}_j$, soumise aux données manquantes, admet pour distribution la fonction $\mathbb{P}(\mathbf{x}_j|\mathbf{X}^{(-j)}, \boldsymbol{\theta}_j)$ où $\mathbf{X}^{(-j)}$ représente la matrice $\mathbf{X}$ privée de la variable $\mathbf{x}_j$ et $\boldsymbol{\theta}_j$ est le paramètre de distribution à estimer. Afin d'ajouter de l'aléa dans l'imputation, on tire un estimateur $\boldsymbol{\theta}_j^*$ grâce à la distribution *a posteriori* $\mathbb{P}(\boldsymbol{\theta}_j|\mathbf{X})$. Ce processus est répété pour chaque variable qui présente des données manquantes en mettant à jour les imputations. L'algorithme ainsi défini utilise à chaque itération l un échantillonneur de Gibbs qui simule successivement :

$$\begin{aligned}
\boldsymbol{\theta}_1^{*(l)} &\sim \mathbb{P}(\boldsymbol{\theta}_1|\mathbf{x}_1^o, \mathbf{x}_2^{(l-1)}, \ldots, \mathbf{x}_p^{(l-1)}) \\
\mathbf{x}_1^{*(l)} &\sim \mathbb{P}(\mathbf{x}_1|\mathbf{x}_1^o, \mathbf{x}_2^{(l-1)}, \ldots, \mathbf{x}_p^{(l-1)}, \boldsymbol{\theta}_1^{*(l)}) \\
&\vdots \\
\boldsymbol{\theta}_p^{*(l)} &\sim \mathbb{P}(\boldsymbol{\theta}_p|\mathbf{x}_p^o, \mathbf{x}_1^{(l)}, \ldots, \mathbf{x}_{p-1}^{(l)}) \\
\mathbf{x}_p^{*(l)} &\sim \mathbb{P}(\mathbf{x}_p|\mathbf{x}_p^o, \mathbf{x}_1^{(l)}, \ldots, \mathbf{x}_{p-1}^{(l)}, \boldsymbol{\theta}_p^{*(l)})
\end{aligned} \tag{3.8}$$

Où $\mathbf{x}_j{}^{(l)} = (\mathbf{x}_j^o,\ \mathbf{x}_j^{*(l)})$ est la j-ème variable imputée à l'itération l en reprenant les notations de Van Buuren and Groothuis-Oudshoorn [2011]. L'utilisation de procédures univariées facilite l'implémentation et permet d'utiliser des procédures différentes selon le type de variables. Cette méthode a fait ses preuves (Horton and Lipsitz [2001]; Moons et al. [2006]; Van Buuren et al. [2006]; Bradburn et al. [2007]) et est largement utilisée aujourd'hui. Nous utilisons le package `mice` de R (Van Buuren and Groothuis-Oudshoorn

[2011]) et suivons leurs recommandations pour ce qui est du choix de la méthode d'imputation et des prédicteurs. Pour chaque variable incomplète, nous imposons la variable *visite* qui représente le numéro de la visite effectuée mais n'imposons pas d'effet aléatoire pour simplifier la méthode et ne pas rendre l'imputation trop longue. Cette méthode s'appelle imputation multiple en classes séparées ou par groupe (Van Buuren et al. [2011]).

3.5 Le cas particulier des variables soumises à un seuil de détection

Un cas particulier de données manquantes que l'on rencontre souvent dans l'analyse de biomarqueurs est la problématique des variables soumises à seuil de détection. La donnée non observée est informative car nous savons qu'elle est en-dessous d'un seuil. En conséquence, le fait d'être manquant dépend de la valeur non observée du marqueur : on se trouve dans le contexte MNAR. Une distinction est faite entre limite de détection (LOD) qui correspond à la plus petite concentration pouvant être détectée et différenciée du bruit de fond, et limite de quantification (LOQ) qui correspond à la plus petite concentration pouvant être quantifiée avec une exactitude et une précision appropriée (figure 3.6).

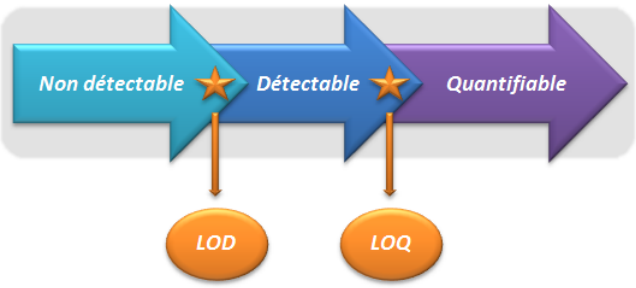


FIGURE 3.6 – Limite de détection et limite de quantification

De nombreux articles traitent de la bonne estimation de ces paramètres, de la gestion du bruit et de l'erreur de mesure (Browne and Whitcomb [2010]; Guo et al. [2010]). En pratique, nous disposons des valeurs observées lorsque cela est possible et de la limite de détection fournie par les laboratoires de mesure. Pour un biomarqueur $\mathbf{b}^*$ nous observons $\mathbf{b}$ où :

$$b_i = \begin{cases} LOD & \text{si } b_i^* \leq LOD \\ b_i^* & \text{si } b_i^* > LOD \end{cases} \tag{3.9}$$

Face à de telles données manquantes plusieurs méthodes sont envisageables. Lorsque l'on souhaite estimer certaines statistiques descriptives pour une réponse soumise à un seuil, Helsel and Cohn [1988] et Shumway et al. [2002] proposent une régression basée sur les rangs, tandis que Klein and Moeschberger [2005] et Meeker and Escobar [2014] proposent d'utiliser la courbe de Kaplan-Meier. Lorsque qu'une seule variable explicative est concernée, Nie et al. [2010]; Richardson and Ciampi [2003] proposent l'imputation par l'espérance attendue sous le seuil. Une deuxième problématique est la distribution de ces biomarqueurs. Certaines méthodes supposent une distribution gaussienne alors que d'autres font l'hypothèse d'une distribution log-normale.

3.5.1 Les méthodes usuelles

La méthode des cas complets où l'on supprime toutes les lignes qui présentent une valeur en dessous du seuil mène à des estimateurs biaisés puisque l'on ne conservera que les valeurs mesurées les plus importantes, d'où une sélection informative.

Imputation simple On retrouve parmi les propositions d'imputations simples l'imputation par 0, LOD, $LOD/2$ ou encore par $LOD/\sqrt{2}$. Si la distribution du biomarqueur est connue, Gleit [1985] et Garland et al. [1993] proposent l'imputation par la moyenne attendue sous le seuil $\mathbb{E}(\mathbf{b}^*|\mathbf{b}^* < LOD)$. Ces imputations sont simples à utiliser et ne font aucune hypothèse de distribution, à part pour l'imputation par la moyenne attendue sous le seuil qui fait intervenir une hypothèse de distribution sur $\mathbf{b}^*$ et l'estimation de ses paramètres. Cependant, l'utilisation d'imputation simple produit des estimateurs et une variance potentiellement biaisés. De nombreuses méthodes d'imputation simple sont utilisées mais critiquées dés lors que le pourcentage de données sous le seuil dépasse les 5%-10% (Lubin et al. [2004]). Pour parer aux limites de ces méthodes, les approches de types "fill-in" proposent de caractériser la distribution du biomarqueur, puis de tirer aléatoirement des valeurs sous le seuil de détection à l'aide de cette distribution (Helsel [1990]; Moschandreas et al. [2001]). En tant qu'imputation simple, la "fill-in" méthode souffre des même

limites que les imputations citées ci-dessus et ne permet pas de modélisation complexe.

Maximum de vraisemblance Supposons que le biomarqueur $\mathbf{b}^*$ de l'équation (3.9) soit gaussien de distribution $\mathbf{X}\boldsymbol{\theta} + \epsilon$ où ϵ et un terme d'erreur de variance σ^2. Dans ce modèle de régression où l'intérêt porte sur l'estimation du vecteur de coefficients $\boldsymbol{\theta}$, il est possible d'utiliser le maximum de vraisemblance sans utiliser d'imputation grâce à la régression Tobit (Tobin [1958]; Gilbert [1987]). Nous cherchons à explorer l'impact de certaines covariables $\mathbf{X}$ sur notre réponse $\mathbf{b}^*$ soumise à un seuil de détection LOD. Alors la log-vraisemblance suivante peut être utilisée :

$$l(\boldsymbol{\theta}, \sigma^2) = \sum_{b_i^* \leq LOD} \log(F(LOD|\mathbf{X}_i\boldsymbol{\theta}, \sigma^2)) + \sum_{b_i^* > LOD} \log(f(b_i^*|\mathbf{X}_i\boldsymbol{\theta}, \sigma^2)) \tag{3.10}$$

où $F(.)$ est la fonction de distribution cumulée et $f(.)$ est la densité de probabilité d'une gaussienne. La première partie de cette équation permet de prendre en compte la probabilité que la mesure soit en-dessous de la limite de détection et la seconde partie prend en compte les valeurs observées au-dessus du seuil. L'estimateur par maximum de vraisemblance peut être obtenu en maximisant l'équation (3.10). Cette méthode est suffisante lorsque l'intérêt porte sur le vecteur $\boldsymbol{\theta}$ (Gilbert [1987]; Lubin et al. [2004]). Dans certains cas, il peut être intéressant d'obtenir des valeurs pour les données sous le seuil, on peut alors obtenir des imputations sous le seuil avec une méthode "fill-in" où $\hat{\boldsymbol{\theta}}$ et $\hat{\sigma}$ sont les estimateur de MLE de l'équation (3.10). Pour ce faire, on calcule :

$$F^{-1}\left\{Unif\left[0, F(LOD|\hat{\boldsymbol{\theta}}, \hat{\sigma})\right]|\hat{\boldsymbol{\theta}}, \hat{\sigma}\right\} \tag{3.11}$$

On tire une uniforme $u_i \sim Unif\left[0, F(LOD|\hat{\boldsymbol{\theta}}, \hat{\sigma})\right]$ entre 0 et la probabilité d'être inférieur au seuil de détection, puis on applique la distribution cumulée inverse sur u_i pour obtenir une valeur sous le seuil de détection. Cette méthode d'imputation simple ne prend pas en compte l'incertitude des imputations mais permet une modélisation plus complexe de la distribution du marqueur. Lorsque les mesures sont répétées dans le temps, il est possible d'utiliser une régression Tobit à effets aléatoires comme dans le package `censReg` de Henningsen [2010].

Imputation multiple Lubin et al. [2004] proposent une imputation multiple composée d'échantillons bootstrap (Efron and Tibshirani [1993]) sur lesquels on applique la régression Tobit. Leur proposition est la suivante :

1. on se donne B échantillons bootstrap en tirant avec remise dans l'échantillon observé
2. sur chaque échantillon bootstrap on estime $\hat{\boldsymbol{\theta}}_b$ et $\hat{\sigma}_b$ à l'aide de l'équation (3.10)
3. on calcule une imputation pour chaque donnée sous le seuil comme pour la méthode "fill-in" décrite dans l'équation (3.11)
4. on utilise les règles de Rubin [1987] pour obtenir des estimateurs combinés, ce qui permet d'intégrer la variabilité des imputations

Dans le cas où ce n'est pas la réponse mais certaines covariables qui sont soumises à un seuil de détection, ce type de méthode peut être envisagé à condition que les variables utilisées pour imputer les autres ne soient pas elles-même soumises à un seuil de détection. Supposons que nous disposions du dosage d'une protéine dans le sang, et du dosage de cette même protéine dans les urines ; lorsque ces deux variables sont soumises à un seuil de détection, il pourrait être intéressant d'utiliser l'information de l'une pour imputer l'autre. Pour ce faire Bernhardt [2013], propose une méthode d'imputation impropre et Lee et al. [2012] proposent une imputation par modélisation jointe. Dans le cas où notre base de données présente d'autres covariables avec des données manquantes d'origines différentes, ces méthodes ne sont plus applicables. Nous proposons donc une imputation multiple à l'aide du package `mice` en faisant appel au package `censReg` pour utiliser la régression Tobit afin d'imputer les variables soumises à un seuil de détection à l'aide d'équations en chaîne.

3.5.2 Une nouvelle fonction d'imputation

L'imputation multiple par équations en chaîne comme définie dans les équations (3.8) peut être utilisée avec différentes formes de fonction d'imputation $\mathbb{P}(.)$. Nous proposons d'utiliser la régression Tobit de la manière suivante. Soit $\mathbf{b}^*$ un biomarqueur gaussien de paramètre $(\boldsymbol{\theta}, \sigma^2)$ soumis à un seuil de détection :

- on estime $\hat{\boldsymbol{\theta}}$, sa matrice de covariances $\hat{\mathbf{V}}_\theta$ (inverse de l'information de Fisher) et $\hat{\sigma}^2$ par à l'aide du modèle Tobit (équation (3.10))
- on tire $\hat{\boldsymbol{\theta}}^*$ sur $\mathcal{N}(\hat{\boldsymbol{\theta}}, \hat{\mathbf{V}}_\theta)$ et $\hat{\sigma}^* = \frac{\hat{\sigma}^2(K-p-1)}{\mathcal{X}^2_{(K-p-1)}}$
- pour chaque valeur sous le seuil, on utilise la méthode définie par l'équation (3.11) à l'aide des paramètres $(\hat{\boldsymbol{\theta}}^*, \hat{\sigma}^*)$

Si la distribution du biomarqueur n'est pas gaussienne, nous la normalisons à l'aide d'une transformation Box-Cox pour pouvoir utiliser les propriétés du modèle Tobit. Pour une base de données qui comporte plusieurs variables soumises à un seuil de détection combinées à des variables qui présentent d'autres formes de données manquantes nous pouvons utiliser cette méthode d'imputation en l'associant à la méthode d'imputation par équations en chaîne de l'équation (3.8). De cette manière, il sera possible d'utiliser l'information des variables soumises à un seuil de détection pour imputer les autres et inversement.

3.5.3 Étude par simulations

Nous proposons une étude par simulations pour évaluer les estimations obtenues par notre méthode et comparer ces résultats avec les méthodes classiques.

Protocole de simulations Nous avons simulé une base de données composée de $K = 200$ patients pour lesquels on mesure 13 variables sur $T = 4$ visites. La variable réponse $\mathbf{y}$ est simulée de la façon suivante :

$$\mathbf{y}_{i,t} = \beta_0 + \sum_{j=1}^{5} x_{i,j,t}\beta_j + \epsilon_{i,t} \tag{3.12}$$

où $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma^2\Sigma)$, et Σ est une matrice auto-régressive de coefficient de corrélation 0.6. La variance σ^2 est choisie pour obtenir un rapport signal bruit égal à 1. La réponse ne dépend que des 5 premières covariables, que l'on appelle actives, à travers le vecteur de paramètre $\boldsymbol{\beta} = (1, 1, 0.2, -0.8, -0.4, 0.6)$ et sont simulées de la façon suivante :

$$\begin{aligned}
\mathbf{x}_1 &= \mathbf{x}_6 - \mathbf{x}_7 + \mathbf{u}_1 \\
\mathbf{x}_2 &= \mathbf{x}_6 - \mathbf{x}_8 + \mathbf{u}_2 \\
\mathbf{x}_3 &= \mathbf{x}_9 - \mathbf{x}_{10} + \mathbf{u}_3 \\
\mathbf{x}_4 &= \mathbf{x}_9 - \mathbf{x}_{11} + \mathbf{u}_4 \\
\mathbf{x}_5 &= \mathbf{x}_{12} - \mathbf{x}_{13} + \mathbf{u}_5
\end{aligned} \tag{3.13}$$

où les variables non actives de $\mathbf{x}_6$ à $\mathbf{x}_{13}$ sont des gaussiennes centrées réduites auto-régressives de coefficient 0.3 et $\mathbf{u}_j$ pour $j \in \{1, \ldots, 5\}$ sont des uniformes sur l'intervalle $[0, 0.1]$. De cette façon, on impose une structure de corrélation spécifique entre les variables actives mais aussi entre les variables actives et non actives comme représenté sur la figure 3.7.

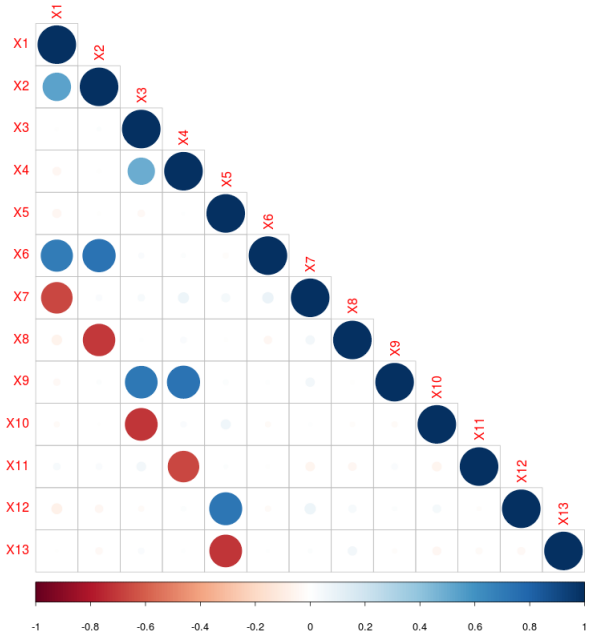


FIGURE 3.7 – Représentation de la structure de corrélation au sein de la base simulée

Données manquantes Nous avons imposé deux types de données manquantes :

1. 10% de données manquantes sur les 8 variables non actives et un pourcentage qui

varie entre 5%, 10%, 15%, 25%, 35%, 50%, 60%, 75%, et 85% pour les 5 variables actives

2. un pourcentage de données manquantes sur les 13 variables qui varie entre 5%, 10%, 15%, 25%, 35%, 50%, 60%, 75%, et 85%

Les données sont supprimées sous le quantile correspondant au pourcentage de données manquantes imposé et ce quantile sera utilisé comme limite de détection. De cette manière, nous pouvons évaluer l'impact de notre méthode d'imputation sur l'estimateur $\hat{\boldsymbol{\beta}}$ en fonction du taux de données manquantes et des variables impactées par les données manquantes. En comparant notre méthode avec les méthode de référence, nous pourrons évaluer l'intérêt d'ajouter l'information de variables externes dans l'imputation.

Les méthodes de référence Nous comparons notre méthode aux imputations simples par LOD, $LOD/2$, $LOD/\sqrt{2}$, par régression Tobit sans ajustement, par régression Tobit en ajustant sur les 12 variables restantes imputées au seuil LOD, par régression Tobit multiple sans ajustement et par régression Tobit multiple en ajustant sur les 12 variables restantes imputées au seuil LOD. Notre méthode est utilisée grâce aux package `mice` et au package `censReg` en spécifiant que les variables à inclure dans le modèle d'imputation doivent avoir un coefficient de corrélation d'au moins 0.2 et 20% de cases *utilisables* (ce qui représente la proportion de cases manquantes pour la variable cible où le prédicteur n'a pas de données manquantes). Nous utilisons 10 imputations obtenues aprés 20 itérations de l'algorithme. La variable *visite*, qui représente le numéro de la visite effectuée, est imposée dans chaque modèle d'imputation, ce que Van Buuren et al. [2011] appelle d'imputation multiple par classe. Après imputation simple, le vecteur de coefficients $\boldsymbol{\beta}$ est estimé par GEE comme indiqué dans l'équation (1.13) avec une matrice de corrélation de travail $\mathbf{R}_i(\alpha)$ auto-régressive d'ordre 1 à l'aide du package `geepack` de Halekoh et al. [2006]. Pour les imputations multiples, les règles de Rubin définies dans les équations (3.6) et (3.7) sont utilisées.

Résultats Nous avons calculé l'$ARB(\hat{\boldsymbol{\beta}})$ selon la méthode indiquée dans l'équation (3.5) sur 500 réplications. La figure 3.8 représente l'évolution de ce critère en fonction du pourcentage de données manquantes pour les deux scénarios de données manquantes imposés. On retrouve le tracé pour chacune des 9 méthodes évaluées : sans données manquantes, imputation simple par LOD, par $LOD/2$ par $LOD/\sqrt{2}$, régression Tobit avec et sans ajustement, régression Tobit et bootstrap avec et sans ajustement ainsi que notre méthode. Le détail des résultats se trouve en annexe B.

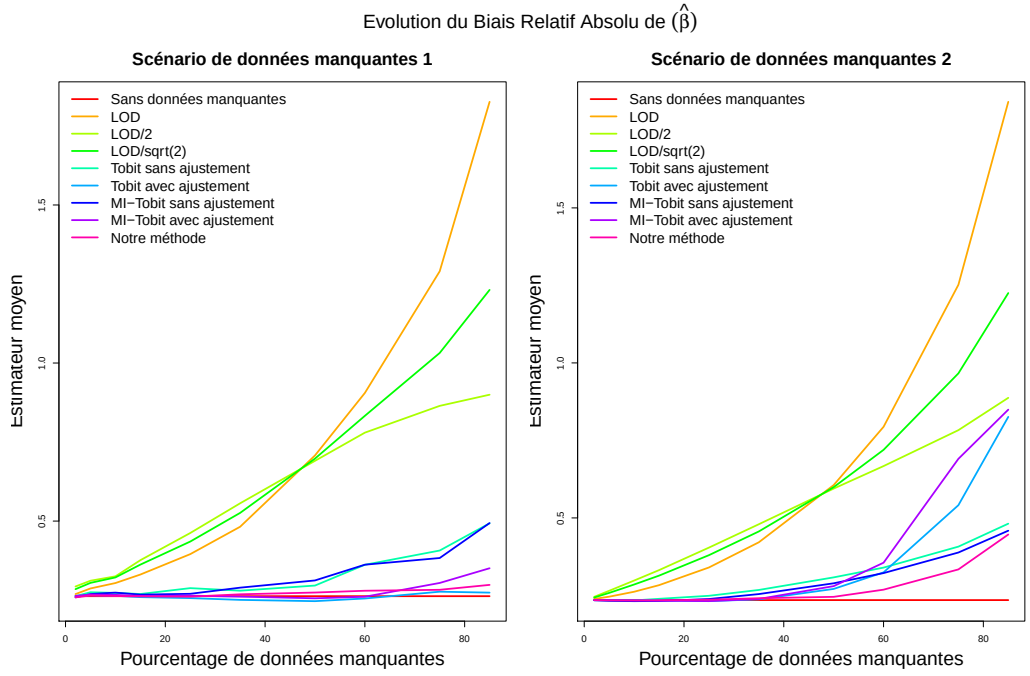


FIGURE 3.8 – Évolution de l'$ARB(\hat{\boldsymbol{\beta}})$ en fonction du pourcentage de données manquantes pour les 2 scénarios de données manquantes

Les trajectoires pour les trois imputations naïves sont similaires entre les deux scénarios de données manquantes. Pour le premier scénario, où 10% de données manquantes sont imposés sur les variables non actives et un taux de données manquantes progressif est imposé aux variables actives, on observe que les trajectoires de la régression Tobit sans ajustement et la régression Tobit sans ajustement avec bootstrap sont très similaires. La régression Tobit avec ajustement et bootstrap est proche de l'ARB sans données manquantes jusqu'au taux de données manquantes de 60% où sa trajectoire croît vers un ARB moins

performant. La régression Tobit avec ajustement sans bootstrap et notre méthode sont les plus performantes avec un ARB très proche de celui obtenu sans données manquantes quel que soit le taux de données manquantes.

Pour le deuxième scénario, la régression Tobit avec ajustement qu'elle soit avec ou sans bootstrap montre un ARB important passé le seuil de 40% de données manquantes. Aprés 60%, la régression Tobit sans ajustement avec et sans bootstrap devient plus performante. Notre méthode admet les meilleurs résultats bien que l'ARB devienne important lorsque le taux de données manquantes dépasse 60%.

La mauvaise précision des méthodes d'imputation simple était attendue et la régression Tobit sans ajustement, qu'elle soit avec ou sans bootstrap, améliore l'estimateur. On remarque un réel apport de l'ajustement pour les deux types de régression Tobit qui donnent des résultats proches de notre méthode. Dans le premier scénario, les méthodes avec ajustement et notre méthode sont similaires. Dans le deuxième scénario notre méthode est la meilleure bien qu'elle soit biaisée après 60% de données manquantes.

Les figures 3.9 et 3.10 représentent l'évolution de l'estimateur moyen pour chaque paramètre β_j estimé. A part pour l'ordonnée à l'origine (β_1), les méthodes d'imputations naïves surestiment toujours l'impact des covariables sur la réponse quel que soit le scénario. Pour le premier scénario, la figure 3.9 nous montre que toutes les méthodes qui font intervenir un ajustement sont meilleures que les autres. Notre méthode est la plus proche de l'estimateur sans données manquantes : elle est la plus précise quel que soit le pourcentage de données manquantes.

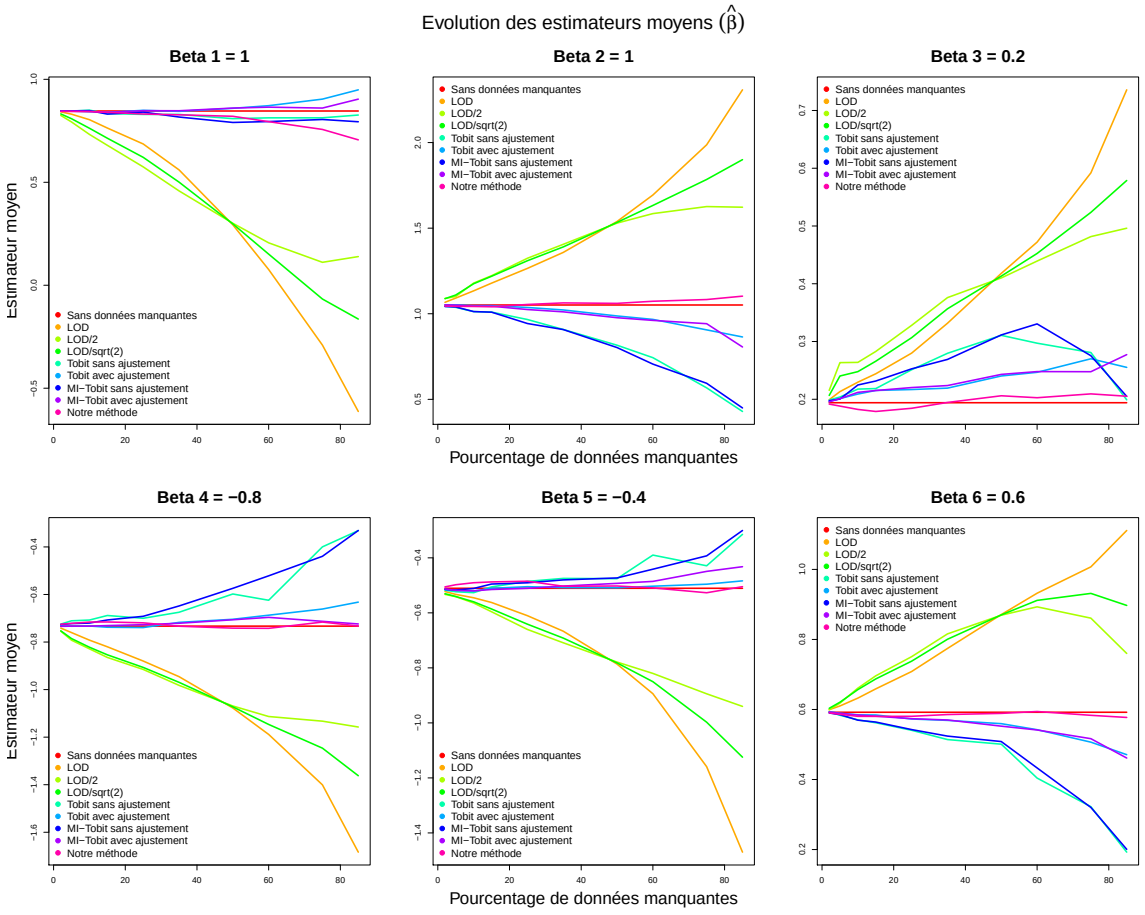


FIGURE 3.9 – Évolution de $\hat{\boldsymbol{\beta}}$ en fonction du pourcentage de données manquantes pour le premier scénario de données manquantes

Pour le deuxième scénario, la figure 3.10 montre que les différences sont moins importantes. Toutes les méthodes montrent un écart important à partir de 60% pour les coefficients les plus importants ($\beta_2 = 1$, $\beta_4 = -0.8$, $\beta_6 = 0.6$).

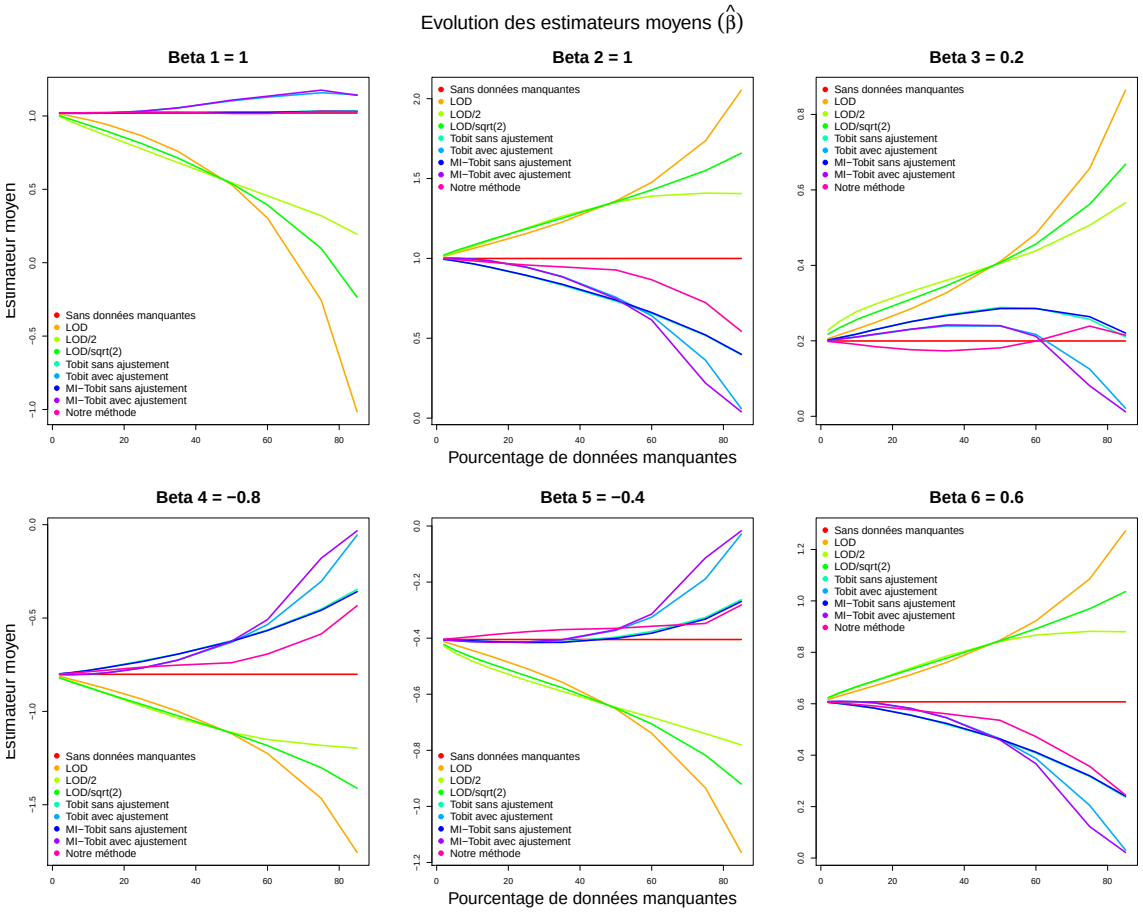


FIGURE 3.10 – Évolution de $\hat{\boldsymbol{\beta}}$ en fonction du pourcentage de données manquantes pour le deuxième scénario de données manquantes

La figure 3.11 représente le temps moyen en secondes pour réaliser l'imputation en fonction du pourcentage de données manquantes pour chaque méthode. Pour le premier scénario, les trajectoires de l'imputation par régression Tobit sans bootstrap avec ou sans ajustement sont confondues. On remarque qu'elles sont stables et que ces imputations sont à peine plus couteûses que les imputations simples naïves. Les trajectoires de l'imputation par régression Tobit et bootstrap avec ou sans ajustement sont confondues et n'augmentent pas au cours du temps bien qu'elles soient supérieures aux méthodes précédentes. Notre méthode représente la méthode la plus couteûses en calcul mais le temps de calcul n'augmente pas avec le pourcentage de données manquantes. Pour le deuxième scénario, on observe les mêmes tendances jusqu'au taux de 60% de données manquantes où la courbe

de temps de la régression Tobit avec bootstrap explose alors que la durée d'imputation de notre méthode décroit à partir de 70%. Cette tendance s'explique par la complexité de calcul que représente l'équation (3.11). En terme de temps de calcul, notre méthode est la plus longue mais reste raisonnable : entre 50 et 100 secondes. Cependant, les bases de données étudiées présentent de nombreuses covariables sujettes aux données manquantes ce qui rend le temps d'imputation plus long.

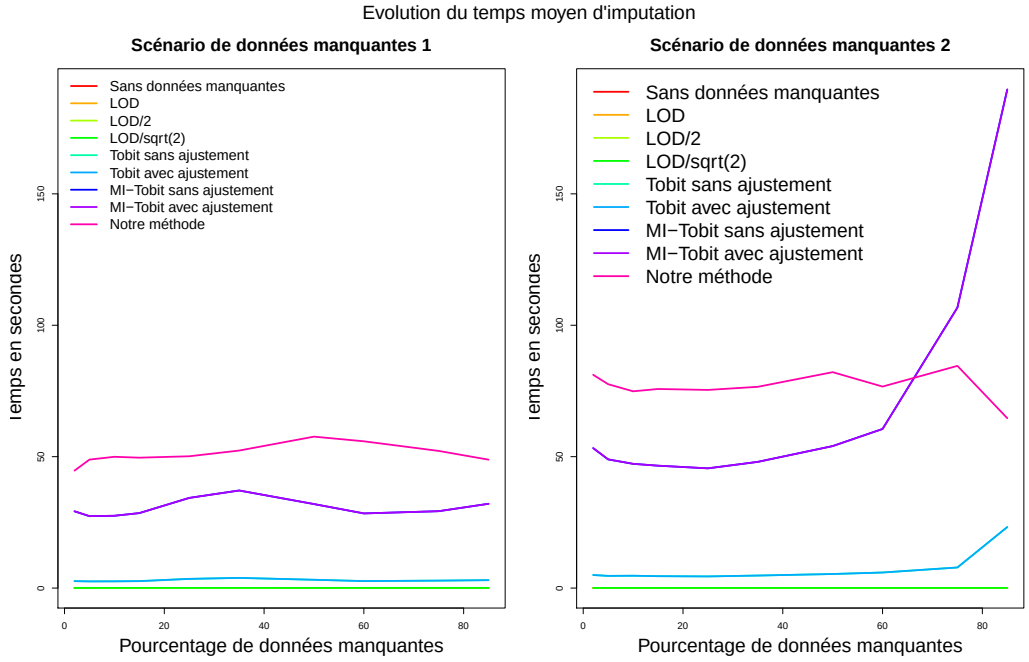


FIGURE 3.11 – Évolution du temps d'imputation en fonction du pourcentage de données manquantes pour les 2 scénarios de données manquantes

Cette étude par simulations nous montre l'intérêt d'intégrer l'information d'autres covariables dans l'imputation de biomarqueurs soumis à un seuil de détection. Cependant, cela nous montre aussi que si l'on intègre de l'information imprécise, comme dans le deuxième scénario où le pourcentage de données manquantes chez les prédicteurs peut atteindre 80%, intégrer cette information ne rajoute que du bruit. En pratique, on ne sait pas quelles variables sont à inclure dans le modèle de régression de la réponse et écarter une variable à cause de son pourcentage de données manquantes peut nous faire perdre une information importante. Inversement, utiliser une variable dont 80% des données ont été mal imputées

n'apporte que du bruit et dessert l'objectif principal. Nous utiliserons dans nos applications notre méthode afin de pouvoir utiliser le package `mice` pour l'imputation de données sous le seuil mais aussi pour l'imputation d'autres types de données manquantes. Lorsqu'une variable présentera un taux de données sous le seuil important, nous réaliserons une étude de sensibilité en comparant les résultats avec d'autres méthodes d'imputation ou en utilisant l'analogue dichotomique qui vaut 1 si $\mathbf{b}^* \geq LOD$ et 0.

Conclusion Nous avons présenté différentes méthodes pour prendre en compte les possibles données manquantes rencontrées dans nos études. Pour les deux derniers cas de données manquantes, nous utiliserons l'imputation multiple par équations en chaîne en utilisant la régression Tobit pour les biomarqueurs soumis à un seuil de détection. Lorsque notre étude sera soumise à une attrition, nous étudierons l'impact de covariables sur l'indicateur de présence, et si besoin est, nous utiliserons la méthode de pondération par la probabilité inverse de réaliser la visite. Seaman et al. [2012] proposent une méthode similaire : ils écartent les sujets qui présentent le plus fort de taux de données manquantes, estiment un modèle logistique pour obtenir des poids de type IPW puis utilisent l'imputation multiple pour les sujets restants qui présentent quelques données manquantes. A l'inverse, nous proposons de garder tous les sujets, d'utiliser la méthode d'imputation multiple pour obtenir M jeux de données imputées et estimer sur chacun de ces jeux les poids correspondants. De cette façon, nous pourrons intégrer dans le modèle logistique des variables qui présentent elles-mêmes des données manquantes. La littérature sur les données manquantes est riche, mais à notre connaissance, peu d'articles traitent d'intégrer ces données manquantes à la sélection de modèles ou de variables.

Chapitre 4

Intégrer les données manquantes dans la sélection de variables pour GEE

Les chapitres 1 et 2 ont proposé un état de l'art sur les méthodes de sélection de modèles et de variables. Cependant, ces méthodes ne permettent pas de prendre en compte les données manquantes présentées dans le chapitre 3. Nous présentons dans ce chapitre quelques méthodes qui permettent de combiner données manquantes et sélection des paramètres de régressions pour les GEE. En présence de données manquantes ponctuelles complétées par imputation multiple, il est crucial d'utiliser une méthode de sélection qui permet d'obtenir une sélection consistante à travers M jeux imputés afin de pouvoir utiliser les règles de Rubin [1987]. Dans la deuxième partie de ce chapitre, nous proposons une nouvelle méthode de sélection de variables : le MI-PGEE. Celui-ci permet de prendre en compte les jeux de données multi-imputées dans la sélection de variables pour GEE.

4.1 Les méthodes de référence

Pour la sélection de modèles Shen and Chen [2012] proposent le Missing Longitudinal Information Criterion (MLIC) pour la sélection du modèle de régression et le MLIC pour la sélection de la structure de corrélation (MLICC). Ces critères prennent en compte les sorties d'étude de type MAR en estimant la perte quadratique attendue à l'aide d'une RSS pondérée par la matrice $\mathbf{W}_i$ des probabilités inverses de rester dans l'étude comme utilisé

pour les WGEE de l'équation (3.3). Le critère s'écrit :

$$MLIC = \sum_{i=1}^{K} (\mathbf{y}_i - \boldsymbol{\mu}_i)^T \mathbf{W}_i (\mathbf{y}_i - \boldsymbol{\mu}_i) + 2\mathbf{tr}(\mathbb{C}ov(\mathbf{y}_i; \boldsymbol{\mu}_i)) \tag{4.1}$$

La première partie de ce critère est une somme des résidus au carré comme définie dans la section 1.2.1, pondérés par l'inverse de la probabilité de sortir de l'étude au temps t. La deuxième partie du critère est une pénalité de covariance entre la réponse et son espérance. Cette quantité fait appel à la matrice de poids $\mathbf{W}_i$ des GEE pondérées de l'équation (3.3) et au score de vraisemblance partielle du modèle d'attrition.

De façon similaire, Gosho [2015] propose deux critères basés sur la quasi-vraisemblance pondérée. Il propose de modifier le QIC défini dans l'équation (1.33) en pondérant la quasi-vraisemblance de l'équation (1.12) par la matrice de poids $\mathbf{W}_i$ des IPW-GEE de la section 3.2. On retrouve cette vraisemblance pondérée $Q_{\mathbf{W}}(\hat{\boldsymbol{\beta}}_I^*)$ comme première partie de ces deux critères :

$$\begin{aligned} QIC_{\mathbf{W}}(\hat{\boldsymbol{\beta}}^*) &= -2Q_{\mathbf{W}}(\hat{\boldsymbol{\beta}}_I^*) + 2\mathbf{tr}(\hat{\Omega}_I \hat{\mathbf{V}}_{\mathbf{R}}) \\ QIC_{\mathbf{W}r}(\hat{\boldsymbol{\beta}}^*) &= -2Q_{\mathbf{W}}(\hat{\boldsymbol{\beta}}_I^*) + 2\mathbf{tr}(\hat{\Omega}_I \hat{\mathbf{V}}_{\mathbf{W}}) \end{aligned} \tag{4.2}$$

Le premier critère proposé par Gosho [2015], le $QIC_{\mathbf{W}}(\hat{\boldsymbol{\beta}}^*)$, ne modifie pas la définition du degré de liberté donnée par Pan [2001a]. En revanche, le degré de liberté du $QIC_{\mathbf{W}r}(\hat{\boldsymbol{\beta}}^*)$ prend en compte les sorties d'études en estimant la matrice de covariance $\hat{\mathbf{V}}_{\mathbf{W}}$ à l'aide du score de vraisemblance partielle du modèle d'attrition (Preisser et al. [2002]).

D'autres critères comme le Generalized Longitudinal Information Criterion (GLIC) de Shen and Chen [2015] basé sur la perte quadratique attendue, ont été proposés. Cependant ces modèles ne permettent pas de prendre en compte les données manquantes ponctuelles. Si une des covariables n'est pas mesurée à un temps d'observation, ces méthodes ne sont plus applicables.

Lorsque les données manquantes ponctuelles sont complétées par imputation multiple, Shen and Chen [2013] propose de calculer le QIC (équation (1.33)) ou le MLIC de l'équation (4.1) sur chaque jeu imputé puis de calculer la moyenne de ces critères sur les M imputations. On obtient le MI-QIC et le MI-MLIC :

$$\begin{aligned} MI\text{-}QIC &= \frac{1}{M}\sum_m QIC_m \\ MI\text{-}MLIC &= \frac{1}{M}\sum_m MLIC_m \end{aligned} \tag{4.3}$$

Pour la sélection de variables En présence de nombreuses covariables, lorsque les méthodes de sélection de modèles ne sont plus suffisantes, Tzeng et al. [2010] proposent l'IPW-PGEE. L'idée est d'estimer des poids de types IPW comme utilisés pour les WGEE de l'équation (3.3) et de les intégrer dans l'équation (2.8) des PGEE. On obtient alors l'opérateur suivant :

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{K} (\mathbf{y}_i - \boldsymbol{\mu}_i)^T \mathbf{V}_i^{-1} \mathbf{W}_i (\mathbf{y}_i - \boldsymbol{\mu}_i) + \lambda \mathcal{P}(\boldsymbol{\beta}) \right\} \tag{4.4}$$

Le choix du paramètre de régularisation λ se fait par minimisation d'un critère de type BIC :

$$\sum_{i=1}^{K} (\mathbf{y}_i - \boldsymbol{\mu}_i)^T \mathbf{V}_i^{-1} \mathbf{W}_i (\mathbf{y}_i - \boldsymbol{\mu}_i) + \log(K) df_\lambda \tag{4.5}$$

où df_λ représente le nombre de covariables associées au modèle. Malheureusement, cette méthode ne permet pas de prendre en compte les données manquantes ponctuelles lorsqu'elles sont complétées par imputation multiple.

Sélection automatique Wood et al. [2008] proposent une généralisation de la méthode de sélection automatique ascendante-descendante dite *stepwise* pour les jeux multi-imputés. La méthode stepwise débute avec le modèle composé de toutes les variables observées puis à chaque itération, combine une étape d'élimination (*backward* : la variable associée à la

plus grande p-value est éliminée du modèle) avec une étape de sélection (*forward* : la variable associée à la plus petite p-value est ajoutée au modèle). De cette manière, chaque variable qui a été éliminée du modèle peut être ré-intégrée si elle est devenue statistiquement significative. Pour adapter cette méthode aux jeux de données imputées, ils proposent d'utiliser les tests de Wald pour imputation multiple de Rubin [1987]. A chaque étape de l'algorithme, le modèle défini doit être estimé sur les M jeux de données imputées, puis l'estimateur combiné doit être calculé afin de pouvoir définir la statistique du test de Wald. Une extension pour GEE est facilement envisageable puisqu'il suffirait d'ajuster la statistique de Wald pour faire apparaître les corrélations intra-sujets. Cependant, bien que Wood et al. [2008] obtiennent de bon résultats sur simulations, ils remarquent que cette méthode n'est pas envisageable en présence de trop nombreuses covariables car elle fait appel à des calculs trop lourds. Cette méthode souffre des mêmes limites que les méthodes de sélection de modèles.

4.2 Le Mutiple Imputation Penalized Generalized Estimating Equations (MI-PGEE)

Introduit par Chen and Wang [2013], le MI-LASSO est un opérateur qui permet de prendre en compte des jeux de données multi-imputées dans la sélection de variables. L'opérateur est défini par :

$$\min_{\boldsymbol{\beta}_{m,j}} \left\{ \sum_{m=1}^{M} \sum_{i=1}^{K} (\mathbf{y}_{m,i} - \boldsymbol{\mu}_{m,i})^T (\mathbf{y}_{m,i} - \boldsymbol{\mu}_{m,i}) + \lambda \sum_{j=1}^{p} \sqrt{\sum_{m=1}^{M} \beta_{m,j}^2} \right\} \tag{4.6}$$

où $\mathbf{y}_{m,i}$ et $\boldsymbol{\mu}_{m,i}$ représentent les vecteurs réponse et moyenne de l'individu i sur le jeu de données imputées m pour $i \in \{1, \ldots, K\}$ et $m \in \{1, \ldots, M\}$. L'astuce est d'utiliser la pénalité Group-LASSO comme définie dans l'équation (2.7) en considérant l'ensemble des M coefficients $\beta_{m,j}$ pour $m \in \{1, \ldots, M\}$ de la même variable $\mathbf{x}^{(j)}$ comme un groupe. De cette façon, $\hat{\beta}_{m,j}$ ne dépend pas uniquement du m-ième jeu de données imputées mais de l'ensemble des M imputations, ce qui permet d'ajuster conjointement les M modèles de régression. Grâce aux propriétés de la pénalité Group-LASSO, l'ensemble des coefficients d'une même variable seront soit tous exactement mis à zéro, soit tous non nuls. La méthode

fournit donc une sélection consistante à travers les jeux de données imputées.

Cet opérateur est défini pour une variable continue et des observations indépendantes. A notre connaissance, aucune méthode ne permet de faire de la sélection de variables en présence de données manquantes pour données longitudinales. Nous proposons une extension du MI-LASSO afin de combler ce vide.

4.2.1 La méthode

Si l'on considère les matrices de variance covariance de travail $\mathbf{V}_{m,i}$ comme fixes, pour l'individu i sur le m-ème jeu de données imputées, notre opérateur peut être défini par :

$$\min_{\boldsymbol{\beta}_{m,j}} \left\{ \sum_{m=1}^{M} \sum_{i=1}^{K} (\mathbf{y}_{m,i} - \boldsymbol{\mu}_{m,i})^T \mathbf{V}_{m,i} (\mathbf{y}_{m,i} - \boldsymbol{\mu}_{m,i}) + \lambda \sum_{j=1}^{p} \sqrt{\sum_{m=1}^{M} \beta_{m,j}^2} \right\} \tag{4.7}$$

On retrouve alors en première partie la somme des WRSS sur les M jeux de données imputées ainsi que la pénalité Group-LASSO sur la deuxième partie. On considère ici p groupes chacun composé de M coefficients estimés. La WRSS permet de prendre en compte les corrélations intra-patients dues au temps à l'aide des matrices de variance covariance de travail $\mathbf{V}_{m,i}$, tandis que la pénalité Group-LASSO permet d'intégrer l'information fournie par chacun des M jeux de données imputées.

L'objectif est de calculer les racines de la dérivée première du problème d'optimisation (4.7). Obtenir la solution de cette équation est complexe puisque la pénalité est non différentiable en zéro. Nous utilisons donc la méthode du LQA comme définie dans la section 2.2.2 (Chen and Wang [2013]). Supposons que nous détenons les estimateurs $\boldsymbol{\beta}_m^{(l)}$ sur chaque jeu de données imputées à l'itération l de l'algorithme. Tant que $\sqrt{\sum_{m=1}^{M} (\beta_{m,j}^{(l)})^2} > 0$ nous pouvons utiliser l'estimation suivante :

$$\sqrt{\sum_{m=1}^{M} \beta_{m,j}^2} \approx \frac{\sum_{m=1}^{M} \beta_{m,j}^2}{\sqrt{\sum_{m=1}^{M} (\beta_{m,j}^{(l)})^2}} \tag{4.8}$$

pour chaque variable $j \in \{1, \ldots, p\}$. En notant $c_j^{(l)} = 1/\sqrt{\sum_{m=1}^{M} (\beta_{m,j}^{(l)})^2}$, on obtient une version simplifiée de la pénalité :

$$\begin{aligned} \mathcal{P}(\boldsymbol{\beta}) &= \lambda \sum_j c_j^{(l)} \sum_m \beta_{m,j}^2 = \lambda \sum_m \boldsymbol{\beta}_m^T \mathbf{C}^{(l)} \boldsymbol{\beta}_m \\ \dot{\mathcal{P}}(\boldsymbol{\beta}) &= 2\lambda \sum_m \mathbf{C}^{(l)} \boldsymbol{\beta}_m \\ \ddot{\mathcal{P}}(\boldsymbol{\beta}) &= 2\lambda \sum_m \mathbf{C}^{(l)} \end{aligned} \tag{4.9}$$

où la matrice $\mathbf{C}^{(l)}$ est la matrice diagonale de taille $p \times p$ composée des poids $c_j^{(l)}$. Comme le coefficient 2 se compense avec la dérivée du score des GEE, nous pouvons transformer le problème d'optimisation (4.7) en un système d'équations comme suit :

$$\sum_{m=1}^{M} \left\{ \sum_{i=1}^{K} D_{m,i}^T \mathbf{V}_{m,i} [\mathbf{y}_{m,i} - \boldsymbol{\mu}_{m,i}] - \lambda \mathbf{C}^{(l)} \boldsymbol{\beta}_m \right\} = 0 \tag{4.10}$$

On obtient une somme de M GEE pénalisées par pénalité Ridge, pondérées par la matrice de poids $\mathbf{C}^{(l)}$, seule quantité commune à ces M régressions. Nous proposons de calculer la solution de cette équation en calculant parallèlement les M solutions de ces PGEE où la matrice de poids $\mathbf{C}^{(l)}$ est mise à jour à chaque itération. La régression Ridge permet d'écraser les coefficients vers zéro mais ne permet pas de les mettre exactement à zéro. Pour surmonter cette limite et permettre à notre opérateur de réduire les coefficients et sélectionner les variables importantes nous imposons $\beta_{m,j} = 0$ pour $m \in \{1, \ldots, M\}$ dès que $\sum_m (\beta_{m,j})^2 < 5^{-10}$.

4.2.2 Algorithme de calcul

Chacune de ces GEE avec pénalité Ridge peut s'écrire :

$$U_P(\boldsymbol{\beta}_m) = U(\boldsymbol{\beta}_m) - \dot{\mathcal{P}}(\boldsymbol{\beta}_m) = \sum_{i=1}^{K} D_{m,i}^T \mathbf{V}_{m,i} [\mathbf{y}_{m,i} - \boldsymbol{\mu}_{m,i}] - \lambda \mathbf{C}^{(l)} \boldsymbol{\beta}_m \tag{4.11}$$

En utilisant le développement de la section 2.2.2 on obtient la formule récursive suivante :

$$\boldsymbol{\beta}_m^{(l+1)} = \boldsymbol{\beta}_m^{(l)} - (\dot{U}(\boldsymbol{\beta}_m^{(\boldsymbol{l})}) - \lambda \mathbf{C}^{(l)})^{-1}(U(\boldsymbol{\beta}_m^{(\boldsymbol{l})}) - \lambda \mathbf{C}^{(l)}\boldsymbol{\beta}_m^{(l)}) \tag{4.12}$$

Si l'on réécrit cette formule pour faire apparaître en bleu les différences avec l'équation (1.21) des GEE on obtient :

$$\hat{\boldsymbol{\beta}}_m^{(l+1)} = \hat{\boldsymbol{\beta}}_m^{(l)} - \left(-\mathbf{D}^T\mathbf{V}^{-1}\mathbf{D} - \lambda \mathbf{C}^{(l)}\right)^{-1} \left(\mathbf{D}^T\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}) - \lambda \mathbf{C}^{(l)}\boldsymbol{\beta}\right)\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(l)}} \tag{4.13}$$

Lorsque le groupe de coefficients associé à la variable $\mathbf{x}^{(j)}$ est réduit à zéro, le processus est non inversible puisque le poids correspondant c_j devient important. Pour parer à cette limite, nous imposons $\beta_{m,j}^{(l)} = \delta$ pour $m \in \{1, \ldots, M\}$ dès que $\sum_m (\hat{\beta}_{m,j}^{l})^2 \leq M\delta^2$ avec $\delta = 10^{-10}$ comme recommandé dans Chen and Wang [2013].

Algorithm 3 Multiple Imputation Penalized Generalized Estimating Equations

Initialize:
$l = 0$
${\boldsymbol{\beta}_m}^{(l)}$, ${\boldsymbol{\alpha}_m}^{(l)}$, ${\phi_m}^{(l)}$ pour $m \in \{1, \ldots, M\}$
$\mathbf{C}^{(l)} = \mathbf{I}_p$
$diff = 10$
while $diff > \epsilon$ and $l \in \{0, \ldots, l_{max}\}$ **do**
  for $m \in \{1, \ldots, M\}$ **do**
    $\mathbf{U}_{P,m}^{(l)} = {\mathbf{D}_m}^T {\mathbf{V}_m}^{(l)-1}(\mathbf{y}_m - {\boldsymbol{\mu}_m}^{(l)}) - \lambda \mathbf{C}^{(l)}\boldsymbol{\beta}$
    $\dot{\mathbf{U}}_{P,m}^{(l)} = -{\mathbf{D}_m}^T {\mathbf{V}_m}_{(l)}^{-1}\mathbf{D}_m - \lambda \mathbf{C}^{(l)}$
    ${\boldsymbol{\beta}_m}^{(l+1)} = {\boldsymbol{\beta}_m}^{(l)} - {\dot{\mathbf{U}}_{P,m}^{(l)}}^{-1}\mathbf{U}_{P,m}^{(l)}$
  end for
  if $\sum_m (\beta_{m,j}^{(l+1)})^2 \leq M\delta^2$ **then**
    $\beta_{m,j}^{(l+1)} = \delta$ pour $m \in \{1, \ldots, M\}$
  end if
  $diff = max\{||{\boldsymbol{\beta}_m}^{(l+1)} - {\boldsymbol{\beta}_m}^{(l)}||_2^2\}$
  $l = l + 1$
  $\mathbf{C}^{(l)} = \mathbf{diag}\left(\left\{1/\sqrt{\sum_m (\beta_{m,j}^{(l)})^2}\right\}_j\right)$
  mise à jour des paramètres ${\boldsymbol{\alpha}_m}^{(l)}$ et ${\phi_m}^{(l)}$ pour $m \in \{1, \ldots, M\}$
end while
if $\sum_m (\beta_{m,j}^{(l)})^2 < 5^{-10}$ **then**
  $\beta_{m,j}^{(l)} = 0$ pour $m \in \{1, \ldots, M\}$
end if

4.2.3 Choix du paramètre λ

Le paramètre de régularisation λ doit être choisi en optimisant un critère de qualité du modèle sur une grille de valeurs données. Les critères définis dans la section 2.3 peuvent être modifiés pour intégrer l'information des M imputations. Les critères de type RSS peuvent être utilisés en calculant la moyenne sur les M imputations comme dans le tableau 4.1 :

Régressions pénalisées	PGEE	MI-PGEE
RSS	$WRSS_{\mathbf{Q}}$	
$(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$	$\sum_i(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^T\hat{\mathbf{Q}}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})$	$\sum_m WRSS_{\mathbf{Q}}^m$

TABLE 4.1 – Critères de type RSS pour la sélection de λ

La WRSS peut faire intervenir le nombre de visites effectuées par chaque sujet afin que chaque individu ait le même poids dans l'analyse même s'ils n'effectuent pas le même nombre de visites. La matrice $\mathbf{Q}_i$ de taille $T_i \times T_i$ peut être la matrice identité (on retrouve la RSS), la matrice des variances marginales $\mathbf{A}_i$, la matrice de corrélation de travail $\mathbf{R}_i$ ou de variance covariance $\mathbf{V}_i$. Pour chacun des critères du tableau 4.1 il est possible d'utiliser les méthodes de validation croisée définies dans la section 1.2. En présence de données longitudinales, la création de groupes se fera sur les sujets et non sur les observations. Ces méthodes étant gourmandes en calculs il est possible d'utiliser la méthode de validation croisée généralisée de l'équation (2.22) :

Régréssions pénalisées	PGEE	MI-PGEE
$RSS/K(1 - p(\lambda,\gamma)/K)^2$	$WRSS_{\mathbf{Q}}/K(1 - p(\lambda,\gamma)/\tilde{n})^2$	$\sum_m GCV^m$

TABLE 4.2 – Critères de type GCV pour la sélection de λ

Le tableau 4.2 définit ce type de critères pour les MI-PGEE où $p(\lambda, \gamma)$ est le nombre effectif de paramètres du modèle composé du nombre de coefficients non nuls multipliés par le taux de réduction comme défini dans l'équation (2.21).

Régréssions pénalisées	PGEE	MI-PGEE
$K\log(RSS/K) + p(\lambda,\gamma)cn$	$N\log(WRSS_{\mathbf{Q}}/N) + dfcn$	$NM\log(\sum_m WRSS_{\mathbf{Q}}^m/NM) + dfscn$

TABLE 4.3 – Critères de type BIC, AIC pour la sélection de λ

Les critères de types AIC et BIC sont nombreux comme le montre le tableau 4.3 où $cn \in \{\log(Nbr), Nbr\}$ et $Nbr \in \{N, K, \tilde{n}\}$. Le degré de liberté df pour les PGEE peut avoir

plusieurs définitions comme détaillé dans la section 2.3.2. Pour chacun des tableaux 4.1, 4.2 et 4.3, la $WRSS$ peut être remplacée par une $Wdev$ comme définie pour le QGCV de Fu [2005].

Pour la sélection de λ pour les MI-PGEE, nous avons choisi de définir un unique critère de type BIC pour la sélection du paramètre de régularisation :

$$BIC = NM \log(\sum_m WRSS^m_{\mathbf{R}}/NM) + dfs \log(\sum_m \tilde{n}^m) \tag{4.14}$$

La RSS est pondérée par la matrice de corrélation de travail et le nombre effectif d'observations $\tilde{n}^m$, estimé sur chaque jeu de données imputées, est utilisé afin de faire un compromis entre un BIC faible qui utilise le nombre de patients K et un BIC lourd qui utilise le nombre d'observations N. La première partie de ce BIC fait intervenir le logarithme de la $WRSS$ qui peut être remplacé par une $Wdev$ pour une réponse discrète. Le degré de liberté du modèle est celui proposé par Yuan and Lin [2006] pour le Group-LASSO et Chen and Wang [2013] pour le MI-LASSO adapté aux GEE :

$$dfs = \sum_{j=1}^{p} I(\sqrt{\sum_{m=1}^{M} \hat{\beta}^2_{m,j}} > 0) + \sum_{j=1}^{p} \frac{\sqrt{\sum_{m=1}^{M} \hat{\beta}^2_{m,j}}}{\sqrt{\sum_{m=1}^{M} \tilde{\beta}^2_{m,j}}}(M-1) \tag{4.15}$$

où $\tilde{\beta}_{m,j}$ est le j-ème coefficient estimé sur le m-ème jeu de données imputées avec le modèle complet (i.e. en utilisant les GEE non pénalisées). On retrouve le nombre de groupes non nuls et le taux de *shrinkage* associé au modèle.

4.3 Comparaisons sur simulations

Nous avons comparé notre méthode à la méthode classique : imputation simple des données manquantes par la moyenne et PGEE. Pour ce faire, nous avons simulé des données complètes pour lesquelles nous avons imposé différents types de données manquantes et comparé les estimateurs et sélections obtenus.

4.3.1 Protocole de simulations

La base simulée est composée de $K = 100$ patients pour $T = 4$ visites effectuées. La réponse $y_{i,t}$ est simulée selon un modèle linéaire simple $y_{i,t} = 1 + \mathbf{x}_{i,t}^T \boldsymbol{\beta} + \epsilon_{i,t}$ où le vecteur d'erreur $\boldsymbol{\epsilon}$ est centré, auto-régressif d'ordre un avec un coefficient d'auto-corrélation $\rho_y = 0.7$ et de variance σ_y^2 choisie pour avoir un rapport signal bruit égal à 1. La matrice de covariables $\mathbf{X}$ est composée de $p = 30$ covariables centrées réduites associées à la réponse grâce au vecteur $\boldsymbol{\beta} = (1, 0.5, -0.2, 1, -1, 0, \ldots, 0)$. Les cinq premières variables sont dites actives puisqu'elles interviennent dans la création de la réponse $\mathbf{y}$, les autres variables dites non-actives sont des variables de bruit. L'objectif est donc de réussir à dissocier le signal porté par les variables actives, du bruit. Trois types de corrélation différents sont imposés entre covariables :

Indépendantes
$cor(\mathbf{x}_l, \mathbf{x}_k) = 0 \forall k \neq l \in \{1, \ldots, 30\}$
Uniformes
$cor(\mathbf{x}_l, \mathbf{x}_k) = 0.5 \forall k \neq l \in \{1, \ldots, 30\}$
Fortes corrélations
$cor(\mathbf{x}_1, \mathbf{x}_3) = 0.9$; $cor(\mathbf{x}_2, \mathbf{x}_5) = -0.8$; $\mathbf{x}_4 = \mathbf{x}_1 - \mathbf{x}_2 + U(0, 0.005)$ et $cor(\mathbf{x}_l, \mathbf{x}_k) = 0.1$ sinon

TABLE 4.4 – Structure de corrélation entre covariables simulées

Le dernier scénario est utilisé pour mettre en difficulté les algorithmes de sélection de variables (Fu [2003]). Les covariables répétées dans le temps présentent généralement une corrélation due au caractère longitudinal de l'étude. Nous avons donc imposé une structure auto-régressive d'ordre un de coefficient $\rho_x = 0.3$ pour chaque variable.

Données manquantes Nos bases de données ne présentent généralement pas plus de 10% de données manquantes au total. Nous avons imposé des données MCAR et MAR pour 5% ou 10% de données manquantes. Les données manquantes MCAR sont imposées en supprimant aléatoirement 5% à 10% de données y compris sur la réponse. Les données MAR sont simulées de la façon suivante :

$$logit(\mathbb{P}(R_{i,j,t} = 0 | x_{i,j,(t-1)}, y_{i,(t-1)})) = \alpha_0 + \frac{1}{2} x_{i,j,(t-1)} + \frac{1}{2} y_{i,(t-1)} \tag{4.16}$$

où $R_{i,j,t}$ vaut 1 si la variable j pour l'individu i au temps t est observée et α_0 est choisi pour avoir 60% de lignes complètes. Des données manquantes pour la réponse peuvent être simulées de la même manière : $logit(\mathbb{P}(R_{i,t} = 0|y_{i,(t-1)})) = \alpha_0 + \frac{1}{2}y_{i,(t-1)}$. Le schéma de données manquantes dépend ainsi de la réponse et de la covariable au temps précédent afin qu'aucune donnée manquante ne soit liée à ce que l'on aurait dû observer.

4.3.2 Résultats

Nos analyses sont basées sur 10 jeux imputés à l'aide du package `mice` de R Van Buuren and Groothuis-Oudshoorn [2011]. Chaque variable qui présente des données manquantes est imputée par *predictive mean matching* en imposant la réponse, la visite et l'ensemble des autres variables comme prédicteurs. Cette méthode propose de calculer une imputation pour chaque données manquantes de la variable $\mathbf{x}_j$ à l'aide d'un modèle linéaire gaussien classique. Dans un deuxième temps, on choisit aléatoirement une observation parmis les 5 observations non manquantes les pus proches de l'imputation proposée.

Le paramètre de régularisation λ est choisi en minimisant un critère de type BIC comme défini dans l'équation (4.14) pour le MI-PGEE. Nous comparons les résultats obtenus en utilisant notre méthode, les PGEE avec imputation simple et les PGEE sur le jeu de données complet. La performance des méthodes de sélection est évaluée à l'aide du MSE ainsi que de la sensibilité et spécificité de la sélection estimés sur 200 réplications.

$$MSE = \frac{1}{M}\sum_{m}(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})^T X_m^T X_m(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}) \tag{4.17}$$

Pour l'imputation multiple, le MSE est calculé comme étant la moyenne des MSE obtenus sur chaque jeu de données imputées. Pour l'imputation simple ou sur jeu complet, on retrouve la forme classique où $M = 1$ dans l'équation (4.17). La méthode la plus performante sera celle associée au MSE le plus faible et au couple (sensibilité, spécificité) le plus proche de $(1, 1)$. Ces deux quantités sont définies par :

$$SEN = \frac{\text{Nombre de variables actives sélectionnées}}{\text{Nombre de variables actives}}$$
$$SPE = \frac{\text{Nombre de variables non-actives non-sélectionnées}}{\text{Nombre de variables non-actives}} \tag{4.18}$$

		5% MCAR		10% MCAR	
	PGEE & Comp	PGEE & IS	MIPGEE	PGEE & IS	MIPGEE
1st Pattern					
MSE	0.440	**0.610**	0.633	0.774	**0.756**
SEN	0.922	**0.895**	0.852	**0.874**	0.851
SPE	0.632	0.686	**0.834**	0.704	**0.784**
2nd Pattern					
MSE	0.587	0.612	**0.252**	0.718	**0.339**
SEN	0.790	0.720	**0.839**	0.723	**0.827**
SPE	0.686	0.719	**0.795**	0.757	**0.766**
3rd Pattern					
MSE	0.890	0.987	**0.495**	1.097	**0.590**
SEN	0.726	**0.677**	0.609	**0.670**	0.610
SPE	0.643	0.697	**0.824**	0.754	**0.798**

TABLE 4.5 – Mean square error (MSE), sensibilité (SEN) et spécificité (SPE) de la sélection pour des données manquantes MCAR avec PGEE sur données complètes (PGEE & Comp), PGEE sur imputation par la moyenne (PGEE & IS) et MI-PGEE sur 10 imputations. Les caractères gras représentent les meilleurs résultats entre imputation simple et MI-PGEE

La table 4.5 présentent les résultats obtenus pour des données manquantes MCAR. Le MI-PGEE obtient la meilleure spécificité quelque soit le scénario envisagé et le plus petit MSE dès que les covariables sont corrélées entres elles. Les sensibilités obtenues sont similaires pour le premier scénario et meilleures pour le deuxième. Malheureusement, toutes les méthodes sont en difficulté sur le troisième scénario. La table 4.6 présentent les résultats obtenus pour des données manquantes MAR. Les résultats sont similaires à ceux de la table 4.5 : on constate de meilleurs spécificités pour le MI-PGEE quelque soit le schéma de corrélation et de plus petites erreurs.

Pour toutes nos simulations, notre méthode obtient les meilleurs taux de spécificité ce qui veut dire que le MI-PGEE est une bonne méthode pour supprimer les covariables *non importantes*. Pour le troisième scénario, le taux de sensibilité de notre méthode est faible ce qui est essentiellement dû à la mauvaise sélection de la variable $\mathbf{x}_3$ associée au coefficient $\beta_j = -0.2$ qui est le plus petit coefficient que nous ayons utilisé. Notre méthode sélectionne

avec un faible taux les variables associées à de petits coefficients. Le troisième schéma de corrélation affecte toutes les sélections en terme de sensibilité et les méthodes avec imputation simple en termes de spécificité. Lorsque les corrélations sont fortes entre variables actives, d'importantes variables ne sont pas sélectionnées quelle que soit la méthode utilisée.

		5% MAR		10% MAR	
	PGEE & Comp	PGEE & IS	MIPGEE	PGEE & IS	MIPGEE
1st Pattern					
MSE	0.412	0.592	**0.580**	**0.687**	0.700
SEN	0.919	**0.882**	0.852	**0.854**	0.840
SPE	0.635	0.735	**0.825**	0.756	0.795
2nd Pattern					
MSE	0.569	0.572	**0.235**	0.608	**0.289**
SEN	0.801	0.705	**0.842**	0.709	**0.842**
SPE	0.686	0.689	**0.809**	0.701	**0.786**
3rd Pattern					
MSE	0.890	1.180	**0.480**	1.062	**0.599**
SEN	0.726	**0.678**	0.622	**0.645**	0.621
SPE	0.643	0.641	**0.816**	0.626	**0.783**

TABLE 4.6 – Mean square error (MSE), sensibilité (SEN) et spécificité (SPE) de la sélection pour des données manquantes MAR avec PGEE sur données complètes (PGEE & Comp), PGEE sur imputation par la moyenne (PGEE & IS) et MI-PGEE sur 10 imputations. Les caractères gras représentent les meilleurs résultats entre imputation simple et MI-PGEE

Peu importe le contexte, passer de 5% à 10% de données manquantes impacte peu le taux de sensibilité mais diminue les taux de spécificité et augmente le MSE. Lorsque l'on compare les deux tables, on remarque que passer de données manquantes MCAR à MAR n'impacte pas les sélections du MI-PGEE alors que les PGEE avec imputation simple présentent de moins bons résultats.

Les MSE obtenus sur ces simulations sont importants, ce qui peut être expliqué par le fait qu'utiliser une approximation par régressions Ridge écrase beaucoup les coefficients. Nous proposons donc d'utiliser le MI-PGEE pour sélectionner le sous-groupe de variables qui impacte le plus la réponse dans un premier temps, puis les règles de Rubin associées aux GEE classiques sur le modèle sélectionné afin d'obtenir un estimateur final. Utiliser une méthode de type LASSO pour la sélection de variables, puis estimer les coefficients associés aux variables sélectionnées est une méthode utilisée par Efron et al. [2004] et discutée par Bühlmann and Van De Geer [2011].

4.4 Robustesse de la méthode

Pour évaluer la robustesse de notre méthode au pourcentage de données manquantes, nous avons conduit une deuxième étude par simulations.

Protocole de simulations Le protocole de simulations est le même que pour la première étude de comparaison avec des corrélations uniformes entre covariables et des données manquantes de type MAR. Nous avons imposé un taux de données manquantes qui varie entre 20%, 30%, 50% et 70% afin d'évaluer les capacités de notre méthode dans un contexte extrême.

Résultats Nous avons comparé les spécificités et sensibilités de la sélection obtenue par imputation simple et PGEE (IS-PGEE) et imputation multiple et PGEE (MI-PGEE). Les résultats sont plus explicites en observant le taux de sélection par variable, aussi nous présentons le taux de sélection de chaque variable en fonction de la méthode de sélection utilisée et du pourcentage de données manquantes imposé (figure 4.1).

La sélection obtenue par imputation simple et PGEE présente une spécificité moyenne qui décroit de façon conséquente avec le pourcentage de données manquantes passant de 0.74 pour 20% de données manquantes à 0.03 pour 70% de données manquantes. Ce résultat est illustré sur la figure 4.1 par les forts de taux de sélection de toutes les variables confondues. En conséquence, cette méthode admet de bonnes sensibilités puisqu'elle sélectionne aussi les variables actives (sensibilité moyenne qui passe de 0.82 pour 20% de données manquantes à 0.97 pour 70%).

Notre méthode présente une sensibilité moyenne qui décroit de 0.727 pour 20% de données manquantes à 0.567 pour 70% de données manquantes. En effet, plus le pourcentage de données manquantes augmente, plus la méthode a du mal à sélectionner les variables actives. Pour la spécificité, on observe une décroissance de 0.8644 pour 10% de données manquantes à 0.7578 pour 70% de données manquantes. Cette quantité décroit faiblement. On observe sur la figure que les taux de sélection des variables non actives augmentent

sans toutefois prendre le pas sur las variables actives.

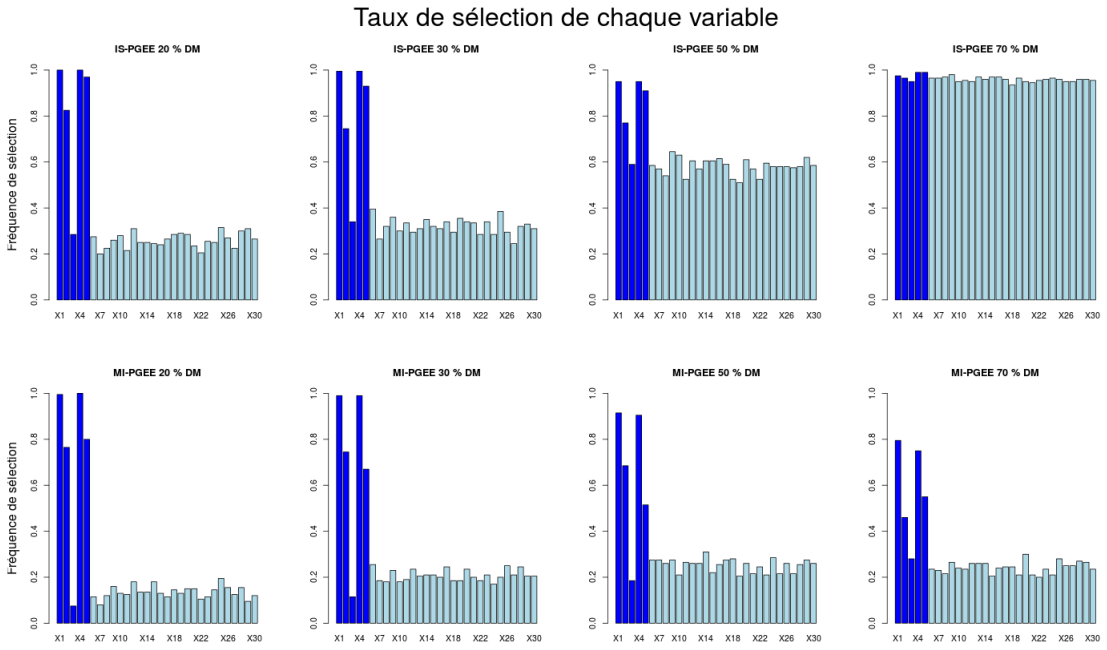


FIGURE 4.1 – Taux de sélection de chaque variable selon la méthode de sélection et le pourcentage de données manquantes imposé. Les variables actives sont en bleu soutenu et les non actives en bleu clair

Ces résultats montrent que notre méthode est robuste aux données manquantes même dans un contexte extrême. La qualité de la sélection est nettement meilleure avec le MI-PGEE qu'avec le IS-PGEE. Cependant, lorsque qu'une grande proportion de données est non observée, nous devons nous poser la question de la fiabilité de ces résultats et mettre en place des analyses de sensibilité afin de vérifier la reproductibilité de notre sélection.

Conclusion Les données manquantes sont inévitables dans les études cliniques. La méthode d'imputation multiple par MICE est flexible, largement utilisée et implémentée dans de nombreux logiciels. Nous proposons le MI-PGEE, une extension du MI-LASSO pour la sélection de variables dans un contexte longitudinal avec des jeux de données multi-imputées. Notre méthode applique parallèlement M GEE pénalisées par Ridge pondérée. Les poids adaptés à chaque variable sont communs à l'ensemble du groupe de coefficients estimés d'une même variable à travers les jeux de données imputées. En conséquence, le MI-

PGEE produit une sélection consistante sur les jeux multi-imputés, ce qui permet d'utiliser les règles de Rubin. Les corrélations intra-sujet sont prises en compte grâce à la matrice de corrélation de travail utilisée pour les GEE.

Notre étude par simulations montre que le MI-PGEE est capable de sélectionner les variables importantes et en même temps d'intégrer les données manquantes et les corrélations intra-sujets. Utiliser les PGEE avec imputation simple conduit à de plus faibles taux de spécificité et de plus grandes erreurs lorsque les covariables sont corrélées. De plus, les données MAR impactent de façon plus importante l'imputation simple que notre méthode, ce qui se traduit par de moins bonnes spécificités et sensibilités que celles obtenues par MI-PGEE. L'utilisation du MI-PGEE est donc recommandée pour les données MAR. Lorsque le MI-PGEE est appliqué à des taux de données manquantes important voire extrême, il reste robuste et montre de meilleures capacités de sélection que le IS-PGEE.

En présence de sorties d'étude, il est possible d'intégrer des poids de type IPW comme dans l'IPW-PGEE. Pour chaque jeu de données imputées nous pouvons ajuster le même modèle de régression logistique afin d'estimer M matrice de poids à intégrer dans l'algorithme. De cette façon, nous pourrons intégrer dans le modèle logistique des variables qui présentent elles mêmes des données manquantes.

Nous avons fait le choix de simuler des bases de données composées de 30 variables afin d'évaluer notre méthode sur des protocoles de simulations similaires à ceux utilisés dans les articles de référence Fu [2003]; Dziak and Li [2006]. Cette dimension est du même ordre que celle rencontrée dans nos bases de données du Chapitre 5 où l'on observe de nombreuses variables sans être dans un contexte de grande dimension. Cependant, il serait intéressant d'évaluer les capacités de la méthode présentée dans un contexte de grande dimension où le nombre de variables dépasse le nombre d'observations ($p >> n$). Théoriquement, la méthode peut s'appliquer dans ce contexte puisque qu'il existe un λ à partir duquel la régression ridge est définie. En pratique, les fonctions implémentées sous `R` doivent être optimisées afin de réduire le temps de calcul. Une version adaptée faisant appel à plus de parallélisation est en cours de developpement et pourra faire l'objet d'un package déposé au Cran.

Chapitre 5

Sélection de marqueurs associés à la sévérité de l'arthrose du genou

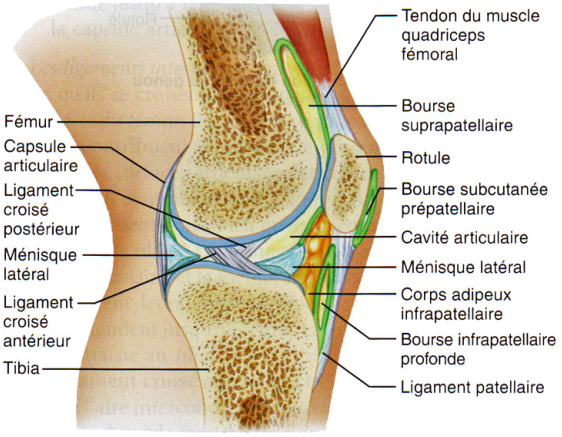


FIGURE 5.1 – L'articulation du genou

L'articulation du genou Le genou est une articulation complexe, qui relie le fémur au tibia. A ce duo, s'ajoute la rotule ou patella. Ces surfaces osseuses sont recouvertes de cartilage dont le rôle est de faciliter le glissement entre les os et leur mouvement tout en protégeant les surfaces osseuses. Dans le cas du genou, cette protection est renforcée par le ménisque qui joue le rôle d'amortisseur de chocs. Les os formant l'articulation sont reliés entre eux par les ligaments comme détaillé sur la figure 5.1. La rotule, quant à elle, est maintenue par le haut grâce au tendon du muscle de la cuisse, et en bas par le ligament rotulien.

5.1 L'arthrose du genou

Épidémiologie de l'arthrose D'après la Société Française de Rhumatologie et l'Association Française de Lutte Anti-Rhumatismale, l'arthrose du genou, ou gonarthrose, touche environ 17% de la population française. Cela signifie qu'entre 9 et 10 millions de Français souffrent de gonarthrose : il s'agit de la maladie rhumatismale la plus fréquente. La prévalence est de 3% chez les personnes de moins de 45 ans, de 65% chez les personnes âgées de plus de 65 ans et atteint 80% chez les personnes âgées de plus de 80 ans.

Mécanisme de la maladie Elle se caractérise par la destruction du cartilage qui s'étend à toutes les structures de l'articulation, notamment l'os et le tissu synovial qui sécrète le liquide synovial qui lubrifie et nourrit le cartilage. Le cartilage, qui tapisse les extrémités osseuses d'une articulation et leur permet de glisser l'une sur l'autre, perd en épaisseur, se fissure et finit par disparaître, entraînant des douleurs et un handicap avec perte de mobilité.

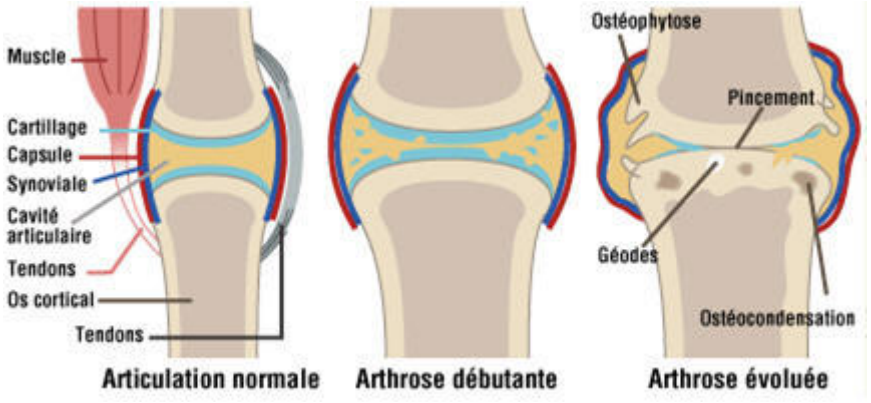


FIGURE 5.2 – Dégradation de l'articulation du genou par l'arthrose

Critère principal Le critère principal retenu comme marqueur d'arthrose est le Joint Space Width (JSW). En ce qui concerne la gonarthrose fémoro-tibiale, il s'agit de l'espace entre le fémur et le tibia. Cet espace est mesuré sur radiographie à différents points stratégiques de contact entre les os de l'articulation comme représenté dans la figure 5.3.

Notre critère principal est le minimum de ces mesures, critère qui représente la sévérité de l'arthrose.

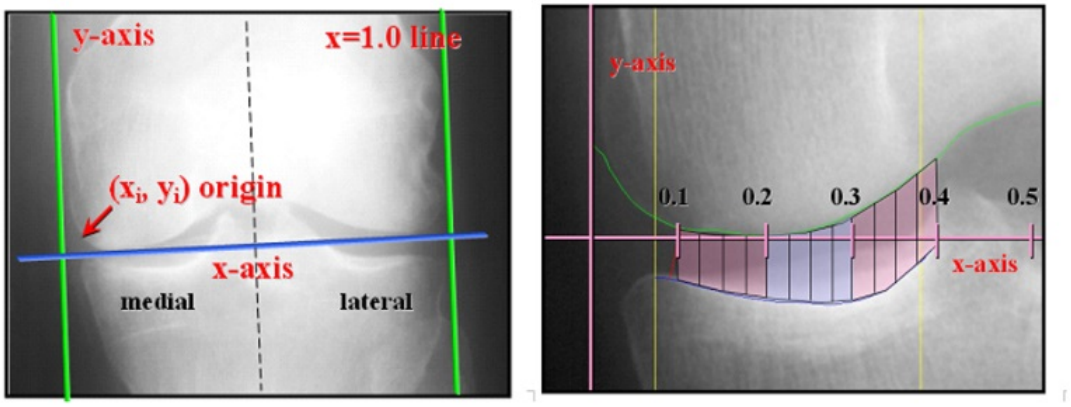


FIGURE 5.3 – Système de mesure de la largeur de l'espace articulaire

Les facteurs de risque La gonarthrose est plus fréquente chez les femmes, avec une augmentation de la fréquence de la maladie après la ménopause. L'obésité ou la surcharge pondérale augmente le risque de développer une arthrose du genou, et accélère la dégradation de l'articulation si la maladie est déjà identifiée. D'autres facteurs de risque sont connus comme l'activité professionnelle (port de charges lourdes), les activités sportives (traumatismes mais aussi surmenage) et les traumatismes articulaires comme les fractures, luxations...

Objectif de l'analyse L'objectif de l'industrie pharmaceutique est de développer des médicaments permettant de stopper l'évolution de la maladie voire de permettre sa régression. Pour ce faire, une meilleure connaissance du mécanisme de la maladie est nécessaire et connaître les biomarqueurs qui sont en lien avec la maladie peut permettre de trouver de nouvelles pistes, voire de nouvelles cibles. La communauté scientifique a donc proposé un certain nombre d'études avec des patients atteints d'arthrose à différents stades où l'on mesure des biomarqueurs sanguins et urinaires, des critères issus d'IRM, des scores de douleur, de rigidité... Tout ce qui pourrait expliquer la maladie, son processus, et/ou son évolution est d'intérêt. L'objectif de notre analyse est d'identifier le sous-groupe de marqueurs qui explique le mieux les différences de JSW entre patients au cours du temps. Nous présentons

deux applications, la première sur le sous-groupe placebo de la base Strontium ranelate Efficacy in Knee OsteoarthItis triAL (SEKOIA), la deuxième sur le projet Foundation for the National Institutes of Health (FNIH) de l'étude publique OsteoArthritis Initiative (OAI).

5.2 L'étude SEKOIA

L'étude SEKOIA (numéro ISRCTN41323372) est une étude clinique traitement contre placebo. Nous nous intéressons à l'histoire naturelle de la maladie. Pour ce faire, nous avons restreint le champ d'étude au bras placebo. Le design de l'étude et les résultats cliniques sont détaillés par Reginster et al. [2012] et Cooper et al. [2012].

5.2.1 La base de données

La base de données est composée de 166 patients ayant chacun 4 visites annuelles prévues. Les écarts de temps entre chaque visite sont quasi-constants (un an) c'est pourquoi nous les avons considéré comme égaux. La figure 5.4 représente la composition de la base de données SEKOIA :

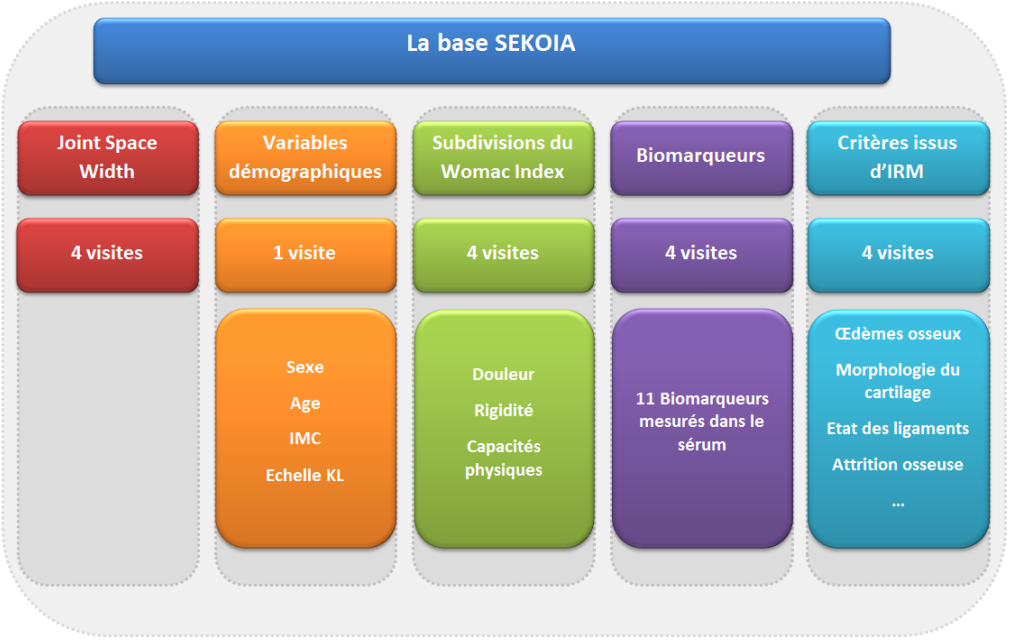


FIGURE 5.4 – Présentation de la base SEKOIA

Le JSW minimal Notre critère principal est un critère continu. La figure 5.5 représente la distribution des quantiles de notre réponse en fonction des quantiles d'une gaussienne pour chaque visite. La distribution est proche d'une gaussienne, nous utiliserons donc un lien identité.

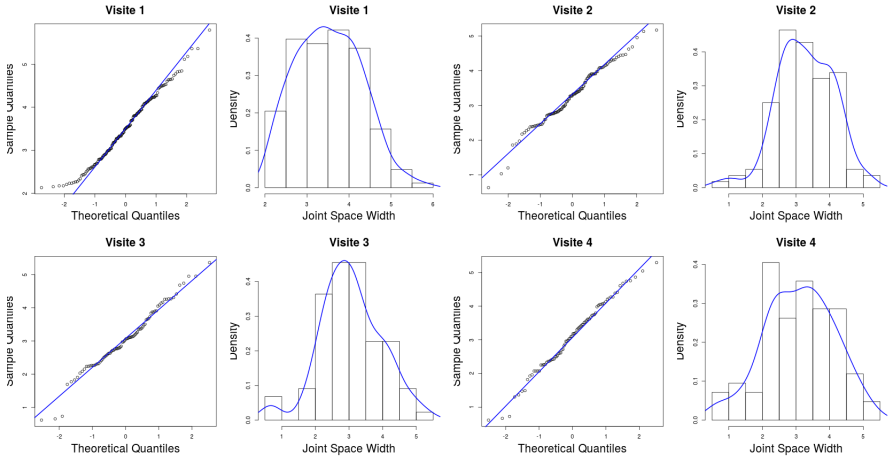


FIGURE 5.5 – Q-Q plots et histogrammes du minimum de la largeur de l'espace articulaire de la base SEKOIA

Le score du Womac Le Western Ontario and McMaster Universities Osteoarthritis Index ou score du WOMAC (McConnell et al. [2001]) est un index qui permet de mesurer les symptômes et handicaps physiques dont souffrent les patients atteints d'arthrose. Ce score est composé de trois dimensions : la douleur, la rigidité, et les capacités physiques divisées en 5, 2 et 17 questions. Chaque question est évaluée sur une échelle de 0 à 100, un score faible correspond à un faible niveau de symptôme, de douleur ou de rigidité. Chaque dimension est sommée sur 500, 200 et 1700 points respectivement. Le score global (compris entre 0 et 2400) représente la somme des trois composantes. Ce score est rempli par le patient en quelques minutes à chacune des visites effectuées.

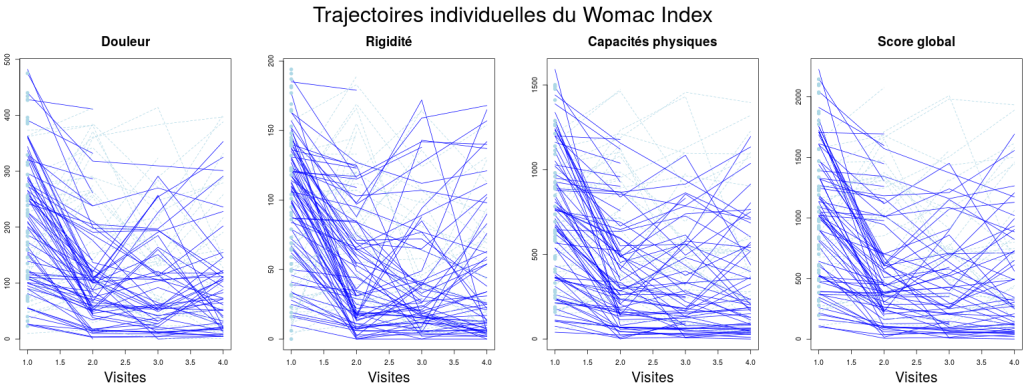


FIGURE 5.6 – Trajectoires individuelles du WOMAC global et de ses composantes : douleur, rigidité et capacités physiques

La figure 5.6 représente les trajectoires individuelles du score du WOMAC en fonction du temps. On remarque que de nombreux patients (en bleu foncé) présentent un effet *prise en charge* avec un score qui décroit de façon conséquente entre la première et la deuxième visite, ce qui montre que ce score souffre de subjectivité. Pour éviter cette confusion, nous avons choisi d'intégrer uniquement la mesure du score à la première visite dans nos covariables.

Les variables démographiques Parmi les covariables mesurées, le sexe, l'Indice de Masse Corporelle (IMC) et l'âge - toutes mesurées à la première visite - sont des facteurs de risque connus de la maladie. A ces facteurs connus s'ajoute l'échelle de Kellgren Lawrence (KL) qui détermine le stade d'avancement de la maladie d'après les radiographies du patient sur une échelle de 0 à 4. Au sein de notre sous-groupe SEKOIA, 120 femmes et 46 hommes souffrent de gonarthrose à des stades II (113 individus) et III (53 individus) répartis de la façon suivante :

	Stade II	Stade III	Total
Femme	81 (48.8%)	39 (23.5%)	120 (72.3%)
Homme	32 (19.3%)	14 (8.4%)	46(27.7%)
Total	113 (68.1%)	53 (31.9%)	166 (100%)

TABLE 5.1 – Répartition du sexe et de l'échelle de Kellgren-Lawrence dans la base SEKOIA

Pour les facteurs continus, l'âge et l'IMC, les statistiques descriptives sont résumées dans le tableau 5.2 :

	Age	IMC
Moyenne (sd)	62,8 (7,5)	29,7 (5)
Médiane (Q_1, Q_3)	61 (58, 68)	29 (25.8, 32.5)

TABLE 5.2 – Moyenne et médiane de l'âge et de l'IMC dans la base SEKOIA

Nous proposons dans nos analyses de sensibilité de la section 5.2.4, d'imposer ces variables dans notre modèle, ce qui équivaut à ne pas les pénaliser. De cette manière, nous pourrons vérifier que l'information portée par ces facteurs de risque reconnus n'est pas noyée dans l'ensemble des variables mesurées ni écrasée par la pénalisation du modèle.

Les critères issus d'IRM Divers critères issus d'IRM du genou sont mesurés dans l'étude SEKOIA. On retrouve dans le tableau 5.3 la liste des critères mesurés selon les recommandations du Whole-Organ Magnetic Resonance Imaging Score (WORMS) (Peterfy et al. [2004]).

Critère	Définition
Morphologie du cartilage	Score sur l'intégrité du cartilage codé de 0 à 6
Oedème de la moelle osseuse	Zone anormalement moins dense sur l'IRM codée de 0 à 3 selon le pourcentage d'os touché
Kyste sous-articulaire	Foyer de signal exempt d'eau avec marges arrondies bien définies codé de 0 à 3 selon le pourcentage d'os touché
Attrition de l'os sous-articulaire	Aplatissement et dépression des surfaces articulaires des os codé 0 pour normal et peut aller jusqu'à 3 pour sévère ($> 50\%$ de l'os)
Ostéophyte marginal	Excroissance du tissu osseux codée de 0 à 7 selon la taille de l'ostéophyte
Ligaments	Variable binaire pour indiquer si le ligament est en bon état (0) ou déchiré (1)
Dommage méniscal	Déchirure du ménisque codée de 0 à 4 selon l'intensité
Corps étrangers	Présence de corps étrangers (morceaux de cartilages détachés) codée de 0 à 3 selon la quantité de corps trouvée.

TABLE 5.3 – Liste des critères issus d'IRM de la base SEKOIA

Chacun de ces critères est mesuré dans divers zones du genou d'après le découpage représenté dans la figure 5.7. Pour une meilleure interprétation et dans l'objectif de tra-

vailler avec un maximum de variables continues nous avons choisi de travailler avec les scores cumulés dans les zones médiale, latérale et sub-spinous du tibia et du fémur et du ménisque. De cette façon, nous évitons le cas complexe des variables codées de 0 à 3 avec de nombreux zéros.

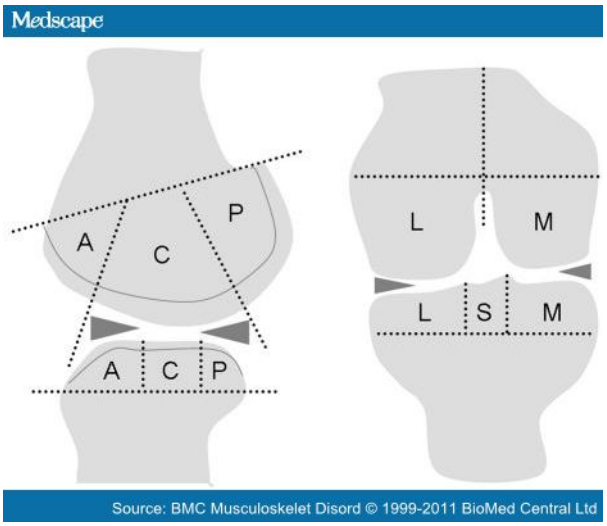


FIGURE 5.7 – Subdivisions par région de la surface articulaire. Le fémur (os supérieur) et le tibia sont séparés entre la partie médiale (M), latérale (L) et sub-spinous (S) pour le tibia. Chacune de ces régions est subdivisée en trois : antérieure (A), centrale (C) et postérieure (P)

Biomarqueurs	Transformations
MMP3	$1/\sqrt{x}$
CSEPI846	$1/x$
BALK2	$\log(x)$
HYALACID	$\log(x)$
SCTX	$\log(x)$
P2CP	$\log(x)$
C12C	$\log(x)$
C2C	$\sqrt{x}$
HCOMP	$\sqrt{x}$
YKL40	$\log(x)$
UCTXC	$\log(x)$

TABLE 5.4 – Transformation choisie pour chaque biomarqueur de l'étude SEKOIA

Les biomarqueurs 11 biomarqueurs en lien avec la dégradation du cartilage, la synthèse du collagène, le remodelage osseux ou cartilagineux et la résorption osseuse ont été mesurés. Les biomarqueurs présents dans l'étude sont souvent de type log-normal. Afin de pouvoir utiliser la régression Tobit présentée dans le chapitre 3, ces biomarqueurs doivent subir une transformation Box-Cox. Ils sont transformés à l'aide de la fonction `boxcoxnc` du package `AID` de R afin de choisir la transformation simple la plus adaptée à la distribution observée. Le tableau 5.4 liste les transformations effectuées. Une analyse de sensibilité est réalisée en section 5.2.4 afin d'évaluer l'impact de ces transformations sur la sélection du sous-groupe de variables le plus en lien avec le JSW.

Des corrélations non-négligeables sont présentes dans notre base de données. On observe sur la figure 5.8 que les critères IRM mesurés dans la même zone ont tendance à être fortement corrélés positivement. Les sous scores Womac présentent également de fortes corrélations positives, résultat attendu au vu des trajectoires similaires de la figure 5.6.

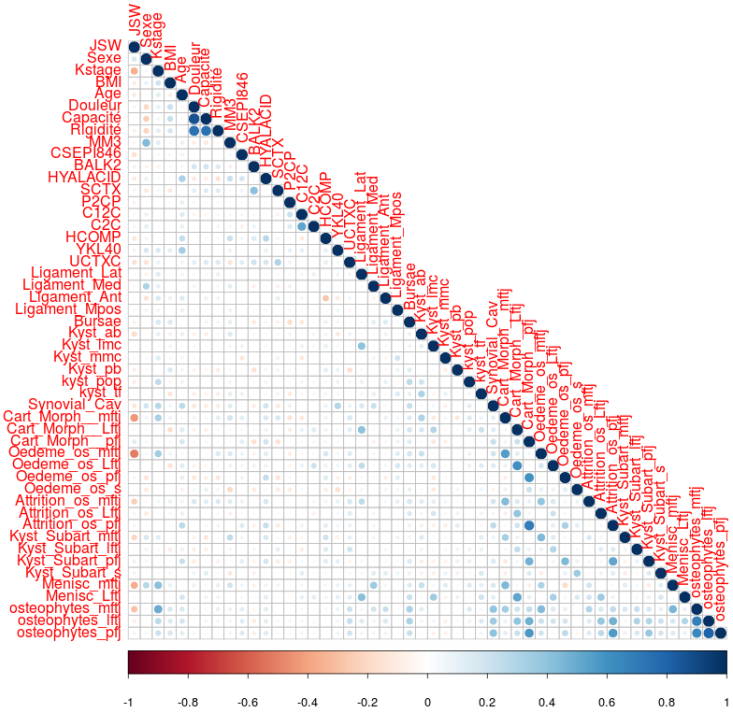


FIGURE 5.8 – Représentation de la structure de corrélation au sein de la base de données SEKOIA

5.2.2 Gestion des données manquantes

La base de données SEKOIA souffre de deux types de données manquantes : des sorties d'étude et des données manquantes ponctuelles. L'attrition est importante dans cette étude puisque 82 patients, soit 49.4% de la population étudiée, présentent une trajectoire incomplète (table 5.5). Pour intégrer cette composante dans nos analyses, nous proposons une étude de sensibilité avec une pondération de type IPW (section 5.2.4).

	Nombre de patients	Nombre de sorties d'étude
Visite $t = 1$	166 (100%)	0 (0%)
Visite $t = 2$	112 (67.5%)	54 (32.5%)
Visite $t = 3$	88 (53%)	24 (14.5%)
Visite $t = 4$	84 (50.6%)	4 (2.4%)

TABLE 5.5 – Nombre (pourcentage) de patients sortis d'étude dans le sous-groupe du bras placebo de l'étude SEKOIA

Les données manquantes ponctuelles représentent 2.39% sur la base totale, ce qui représente 20% de lignes incomplètes et 41% de patients qui présentent des données manquantes. Elles sont réparties de la façon suivante :

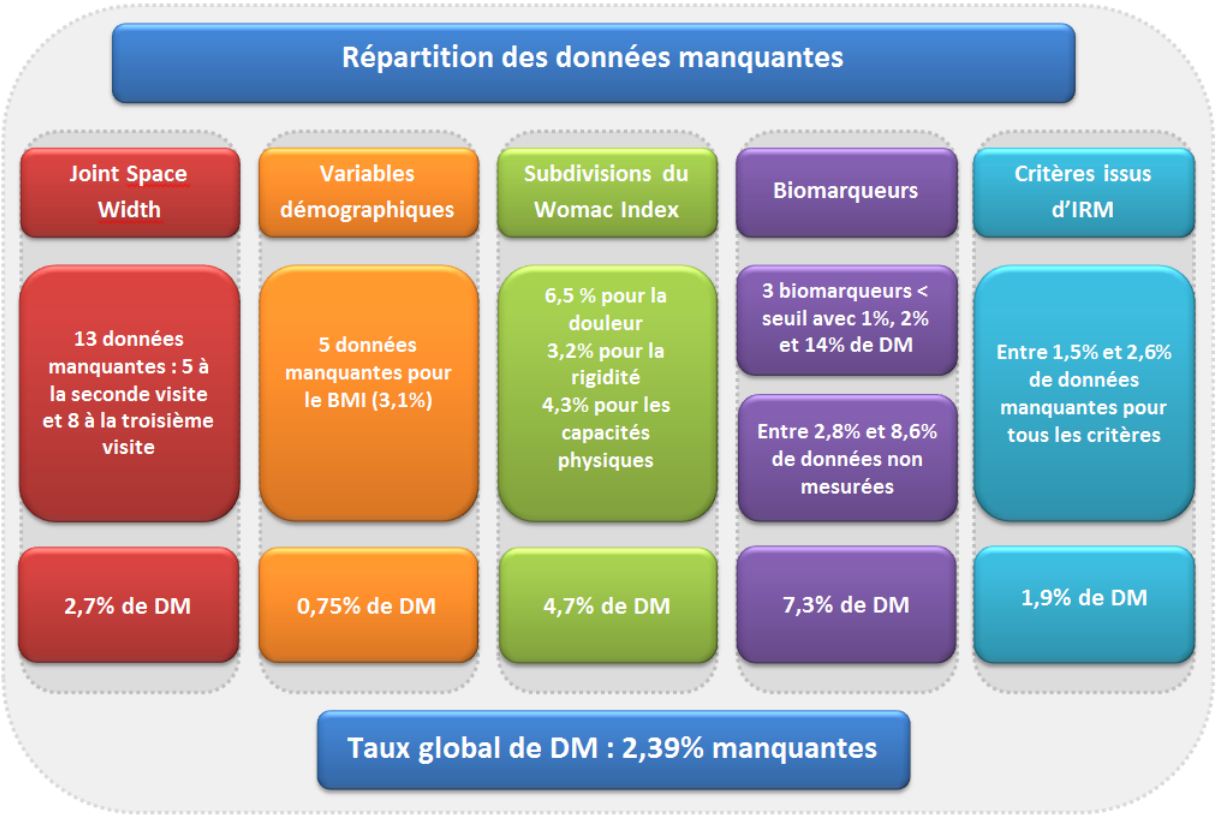


FIGURE 5.9 – Répartition des données manquantes (DM) dans la base SEKOIA

Imputation Multiple Nos analyses sont basées sur 10 imputations effectuées à l'aide du package `mice` de Van Buuren and Groothuis-Oudshoorn [2011]. Les fonctions d'imputa-

tions utilisées sont le *predictive mean matching* pour les variables continues, la régression logistique pour les variables binaires et la régression Tobit pour les variables soumises à un seuil de détection. La réponse et la variable *visite* qui représente le numéro de la visite effectuée sont imposées comme prédicteurs pour chaque variable qui présente des données manquantes. Nous avons utilisé la fonction `quickpred` afin de n'utiliser que des prédicteurs ayant au moins 30% de cases utilisables et un coefficient de corrélation supérieur à 0.1.

Parmi les 11 biomarqueurs, le MMP3, le SCTX et l'HYALACID sont soumis à un seuil de détection avec un taux respectif de 1%, 2% et 14% de données manquantes. Au vu des faibles pourcentages de données manquantes des biomarqueurs SCTX et MMP3, nous présentons uniquement les imputations du biomarqueur HYALACID. Nous pouvons représenter les imputations toutes visites confondues de la manière suivante :

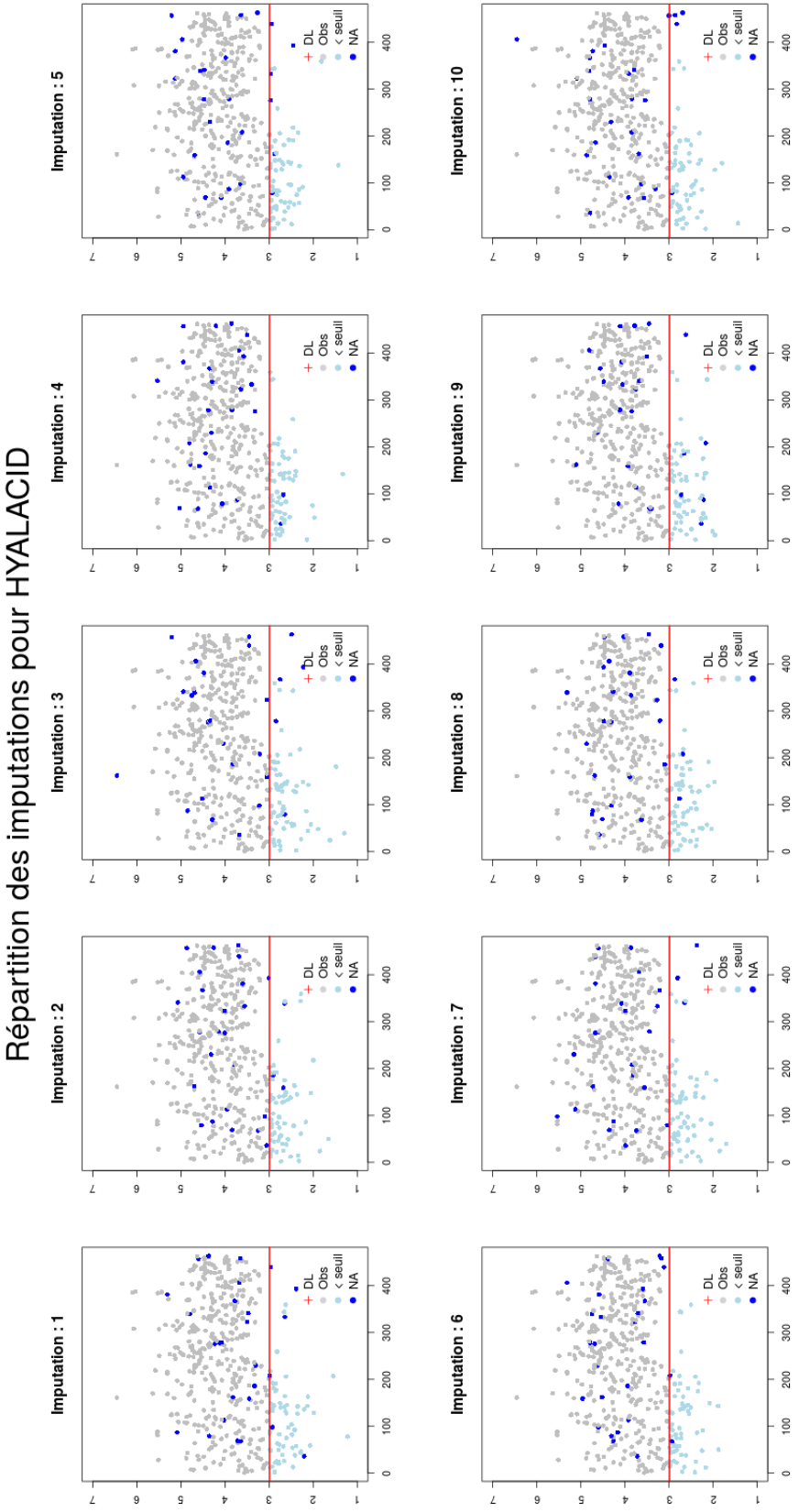


FIGURE 5.10 – Répartition du biomarqueur HYALACID sur 10 jeux de données

5.2.3 Sélection de variables par MI-PGEE

La base de données SEKOIA est composée de 52 variables parmi lesquelles : la largeur de l'espace articulaire, les 4 variables démographiques, les 3 scores du Womac, les 11 biomarqueurs et 33 critères issus d'IRM. Nous nous sommes donnés une grille fine de valeurs de λ possibles pour lesquelles nous avons calculé l'estimateur $\hat{\boldsymbol{\beta}}_\lambda$ et le BIC associés à l'aide du MI-PGEE avec une structure de corrélations auto-régressive d'ordre un.

La figure 5.11 représente l'évolution des coefficients estimés et du BIC. Pour plus de visibilité, nous avons tracé cette évolution en fonction du taux de réduction s :

$$s = \frac{\sum_{j=1}^{52} \sqrt{\sum_m \hat{\beta}_{m,j}^2}}{\sum_{j=1}^{52} \sqrt{\sum_m \tilde{\beta}_{m,j}^2}} \tag{5.1}$$

Où $\hat{\beta}_{m,j}$ et $\tilde{\beta}_{m,j}$ sont les estimateurs de β_j sur le m-ième jeu imputé, par GEE pénalisées et GEE sans pénalisation respectivement. Dessiner l'évolution en fonction de s permet une meilleure visualisation et interprétation des résultats. Pour $s = 1$ aucune pénalisation n'est appliquée, $\lambda = 0$, toutes les variables sont incluses dans le modèle et $\hat{\beta}_{m,j} = \tilde{\beta}_{m,j}$. On remarque qu'on obtient un BIC optimal pour $s = 0.006$ soit un taux de réduction très fort, ce qui peut s'expliquer par le grand nombre de variables non sélectionnées et les fortes corrélations entre variables.

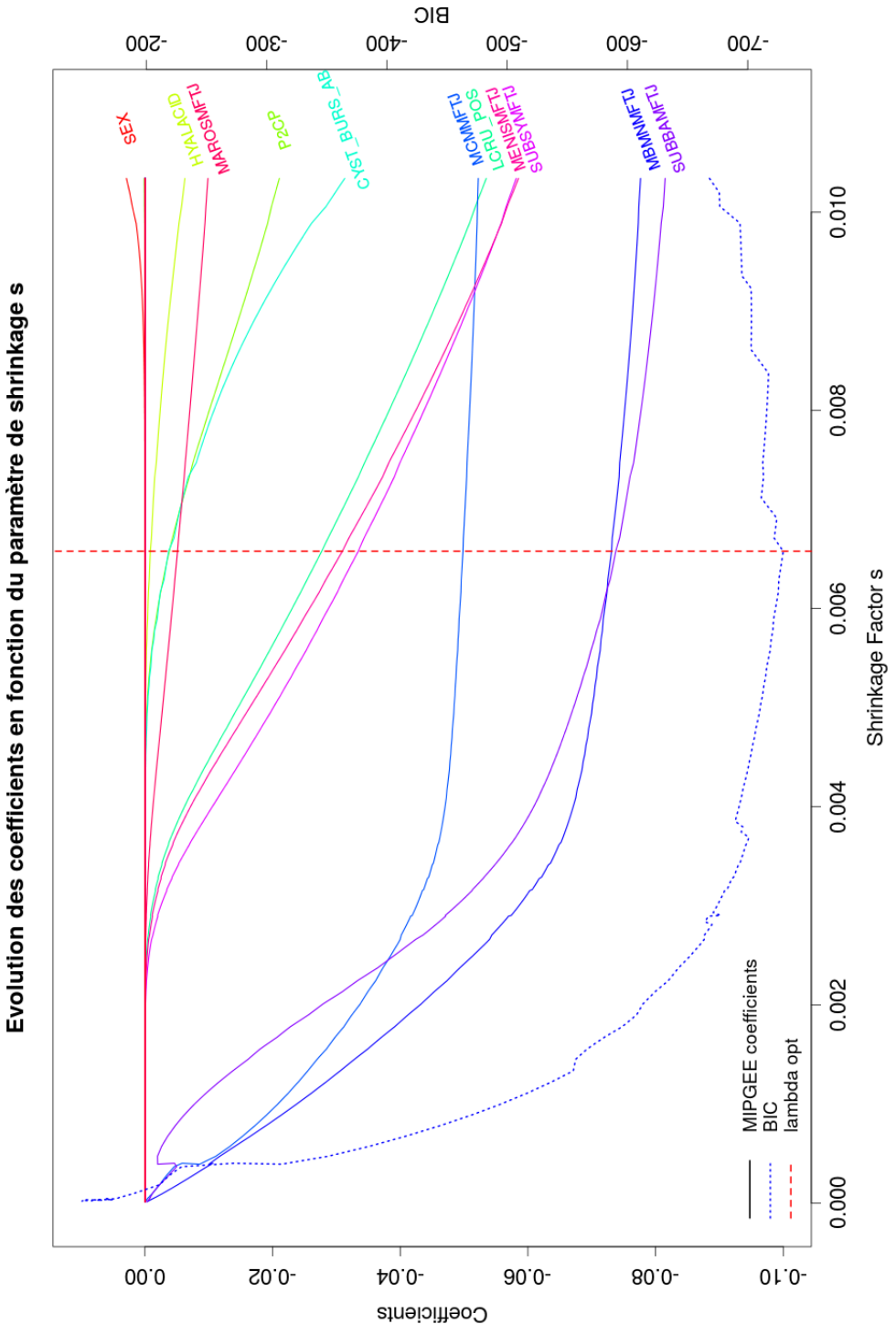


FIGURE 5.11 – Chemin des coefficients estimés par MI-PGEE sur la base SEKOIA

Le modèle final est composé de 10 variables. En estimant les coefficients associés grâce aux règles de Rubin, on obtient les estimateurs suivants :

Variables	$\hat{\beta}_j(et)$	$\lvert\hat{\beta}_j\rvert/et$
HYALACID	−0.038 (0.046)	0.826
P2CP	−0.070 (0.054)	1.296
Ligament Postérieur	−0.146 (0.428)	0.341
Kyste bursine ansérine	−0.275 (0.214)	1.285
Cartilage Morphologie (région MFT)*	−0.066 (0.025)	2.640
Oedème de la moelle osseuse (région MFT)	−0.069 (0.074)	0.932
Attrition de l'os sous-articulaire (région MFT)	−0.098 (0.171)	0.573
Kyste sous-articulaire(région MFT)	−0.136 (0.099)	1.374
Dommmage méniscal (région MFT)	−0.124 (0.099)	1.253
Ostéophyte marginale (région MFT)	−0.029 (0.020)	1.450

TABLE 5.6 – Sélection obtenue par MI-PGEE sur la base SEKOIA, estimation (écart type) par MI-GEE. Le symbole * représente les variables pour lesquelles l'intervalle de confiance ne comprend pas zéro. La région MFT représente la zone médiale fémoro-tibiale

On retrouve 6 scores cumulés différents dans la zone médiale fémoro-tibiale. Ce résultat n'est pas étonnant puisque les patients de l'étude SEKOIA ont été choisis pour avoir une importante arthrose du genou dans cette zone. Ces critères d'inclusion spécifiques sont cependant un biais pour la généralisation de nos résultats.

Les scores cumulés sélectionnés représentent l'ensemble des critères IRM mesurés, ce qui met en évidence que la gonarthose n'est pas uniquement une maladie du cartilage mais de toutes les structures de l'articulation. Les coefficients associés sont tous négatifs. Ce résultat signifie que plus les scores cumulés sont importants (i.e. plus l'articulation est dégradée), plus le JSW est faible, ce qui est conforme à l'intuition. De même, la présence de kyste bursine ansérine et le fait d'avoir le ligament postérieur déchiré sont en lien avec une largeur de l'espace articulaire plus faible.

Parmi les biomarqueurs mesurés, on trouve dans le modèle final le C-Terminal Propeptide of type II procollagen (P2CP) qui est responsable de la synthèse de collagène et l'hyaluronic acid (HYALACID), qui permet le bon fonctionnement du liquide synovial. En d'autres termes, une personne avec une concentration de P2CP ou d'HYALACID qui augmente aura tendance à avoir un JSW qui diminue. Pour l'HYALACID, ce résultat peut s'expliquer par la présence d'un épanchement de synovie (augmentation du liquide syno-

vial, riche en acide hyaluronique). Pour le P2CP, cela peut être une réponse du patient qui produit plus de collagène pour compenser la dégradation de son cartilage.

La seule variable à obtenir un intervalle de confiance ne comprenant pas zéro est le score cumulé de morphologie du cartilage : $-0.066 \in [-0.114; -0.016]$. Pour toutes les autres variables, les écarts types sont importants et souvent du même ordre que l'estimateur lui-même, ce qui peut s'expliquer par la multicolinéraité présente dans l'étude (figure 5.8). De plus, l'étude manque de puissance puisque nous ne disposons que de 166 patients qui souffrent d'une attrition d'environ 50%. Certains coefficients sont donc mal estimés malgré leur sélection. Dans ce contexte, une base de données plus importante est nécessaire pour une analyse plus fine.

5.2.4 Analyses de sensibilité

Inverse Probability Weighting Afin de prendre en compte l'importante attrition de l'étude, nous avons utilisé la méthode IPW. Différents modèles ont été ajustés - comprenant le sexe, l'age, l'IMC, les sous-parties du score Womac, l'échelle de Kellgren Lawrence et le JSW aux temps précédents - maisa ces variables ne sont pas statistiquement significatives et les résultats de sélection sont inchangés par le modèle de régression logistique choisi. Nous présentons ici les résultats pour le modèle de régression suivant :

$$\log(\mathbb{P}(R_{i,t} = 1 | R_{i,(t\text{-}1)} = 1) = \alpha_0 + \alpha_1 D_i + \alpha_2 C_i + \alpha_3 R_i + \alpha_4 JSW_{i,(t\text{-}1)} \tag{5.2}$$

Où D_i, C_i et R_i représentent les scores de douleur, capacités et rigidité du Womac de l'individu i pour $i \in \{1, \ldots, K\}$. Les estimations obtenues sont faibles avec des coefficients proches de zéro et aucune variable n'est statistiquement significative. Dans le but de vérifier que la sélection reste inchangée, nous présentons les résultats de cette méthode. L'estimateur $\hat{\boldsymbol{\alpha}}$ ainsi obtenu nous permet d'estimer les probabilités $\mathbb{P}(R_{i,t} = 1 | R_{i,(t\text{-}1)} = 1)$ et donc les poids $\omega_{i,t}^{-1}$ avec :

$$\hat{\omega}_{i,\tau} = \prod_{t=1}^{\tau} \hat{\lambda}_{i,t} \tag{5.3}$$

Où $\lambda_{i,t} = \mathbb{P}(R_{i,t} = 1 | R_{i,(t\text{-}1)} = 1)$. Nous pouvons estimer ces poids sur chaque jeu de

données imputées et tracer leur évolution comme sur la figure 5.12 :

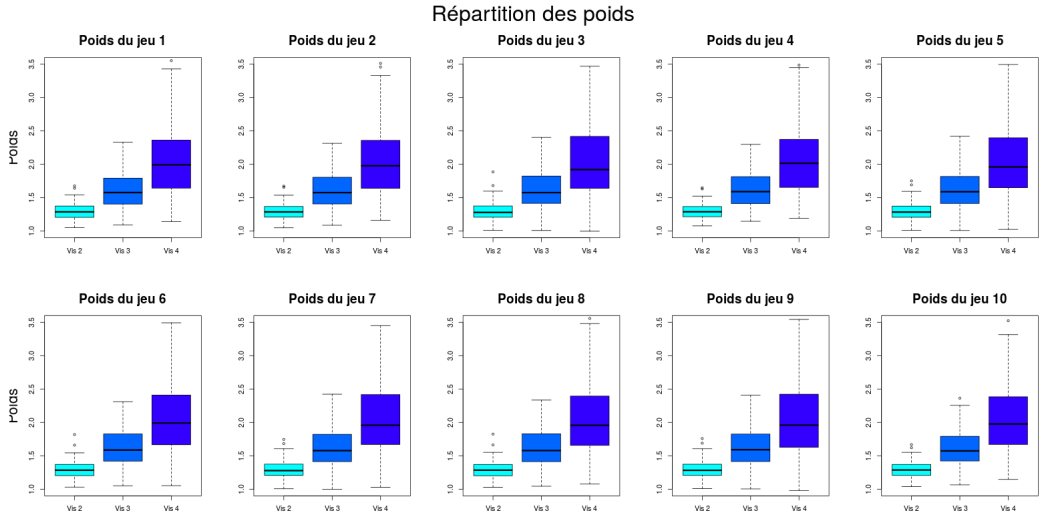


FIGURE 5.12 – Répartition des poids IPW sur chaque jeu imputé

Ce graphique nous permet de vérifier que les poids sont stables (ici compris entre 1 et 3.5) et trés similaires d'un jeu imputé à l'autre. Les poids sont plus importants pour les dernières visites. En avancant dans le temps, la probabilité d'effectuer la visite diminue ce qui conduit à des poids plus importants. Nous pouvons les utiliser dans l'algorithme du MI-PGEE de la même façon que pour l'IPW-PGEE de Tzeng et al. [2010] défini par l'équation (4.4), avec une matrice de poids spécifique à chaque jeu de données imputées. Que l'on applique le MI-PGEE avec ou sans pondération de type IPW ne modifie pas la sélection et très peu les estimations des paramètres. Associé au fait qu'aucun des facteurs de risque connus ne semble en lien avec l'indicatrice de présence, cela semble indiquer une attrition MCAR. Une autre hypothèse est que les variables *responsables* de l'attrition ne sont pas mesurées ici, hypothèse difficilement vérifiable.

Biomarqueurs imputés au seuil Lorsque l'on impute au seuil les biomarqueurs soumis à une limite de détection, la sélection est inchangée. Les estimateurs sont très similaires sauf pour le biomarqueur HYALACID. Cette variable qui présente 14% de données sous le

seuil est associée à un écart type plus faible lorsque l'on utilise l'imputation simple puisque la variabilité des imputations n'est pas intégrée.

Transformations Box-Cox L'ensemble des biomarqueurs de la base a été *normalisé* par transformation Box-Cox. Nous avons utilisé le MI-PGEE sur la base SEKOIA après imputation sans transformation. Dans ce contexte, les biomarqueurs ne sont plus sélectionnés. Ce changement est explicable par la forme de la distribution de ces biomarqueurs qui est souvent de type logarithme.

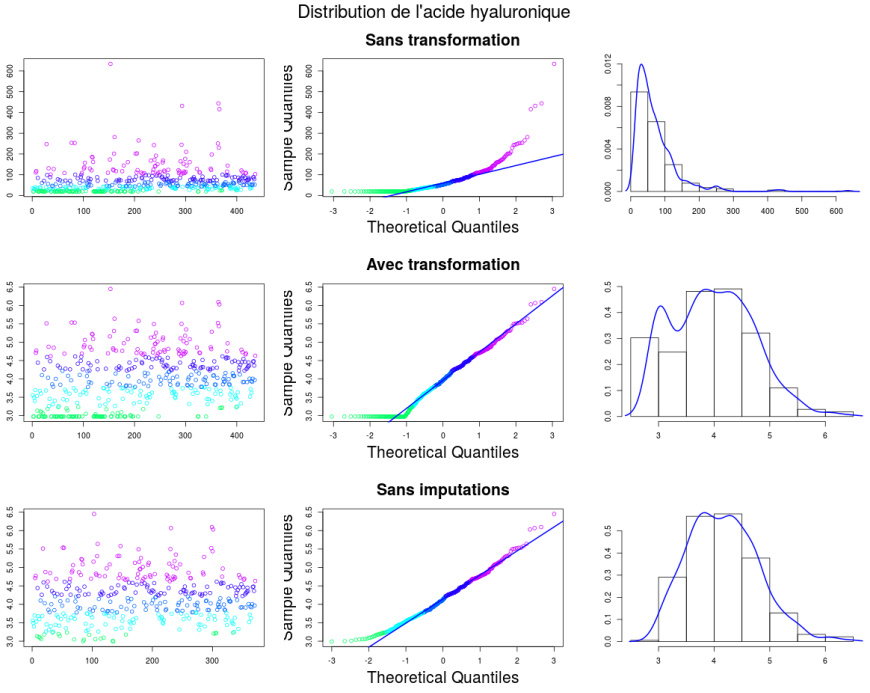


FIGURE 5.13 – Distribution de l'acide hyaluronique avec observations sous le seuil imputées au seuil, avec transformation logarithme et avec transformation logarithme sans imputation. Les points verts, bleus clairs, bleus soutenus, violets et roses représentent les observations sous le quantile 0.2, 0.4, 0.6, 0.8 et 1

La figure 5.13 représente la répartition, les quantiles et la distribution de l'acide hyaluronique. Sur la première ligne, on observe que la distribution du biomarqueur est lognormale avec de nombreuses observations sur les faibles valeurs. En conséquence, une différence de faible importance aura plus d'impact pour les concentrations faibles que pour les concen-

trations importantes. Les deux dernières lignes nous montrent qu'une fois transformé, le nuage de points perd son hétéroscédasticité (la queue de distribution est alourdie sur la deuxième ligne de a figure 5.13 par la sur-représentation du seuil). Lorsque la distribution est log-normale, il est difficile de dissocier les mesures : les valeurs sous le quantile 0.2, 0.4 et 0.6 sont superposées. Lorsque la distribution est transformée, les quantiles sont mieux visibles.

Avec facteurs de risque imposés Lorsque de nombreuses variables sont mesurées, il est possible que des facteurs de risque, c'est à dire des variables cliniques connues pour être en lien avec la maladie, ne soient pas sélectionnées. L'information que portent ces variables peut être noyée dans la masse d'informations collectées (Boulesteix and Sauerbrei [2011]). De plus, nos variables présentent des corrélations non négligeables qui peuvent jouer un rôle sur le taux de réduction appliqué aux coefficients. De ce fait, il est possible que notre sélection passe à côté de variables importantes. Afin de le vérifier, nous avons utilisé le MI-PGEE en imposant les facteurs de risque que sont l'age, le sexe, l'IMC et l'échelle de Kellgren-Lawrence. De cette manière les coefficients qui leur sont associés ne peuvent pas être mis à zéro.

La sélection obtenue est inchangée si ce n'est que ces variables sont ajoutées au modèle final avec les estimations suivantes :

Variables	$\hat{\beta}_j$ (et)
Sexe	0.590 (0.132)
Échelle KL	−0.041 (0.143)
IMC	0.0002 (0.0119)
Age	−0.008 (0.0082)

TABLE 5.7 – Sélection obtenue par MI-PGEE sur la base SEKOIA en fixant les facteurs de risque, estimation (écart type) par MI-GEE

Les estimations des autres paramètres sont presque identiques à celles du tableau 5.6. Seul le sexe, qui vaut un pour les hommes et zéro pour les femmes, obtient un intervalle de confiance qui ne comprend pas zéro ($0.590 \in [0.331; 0.849]$). Pour les autres facteurs de risque, les écarts types sont similaires, voir plus grands que l'estimateur lui-même.

On retrouve bien que le sexe est un facteur de risque connu de la maladie : la prévalence et la sévérité de la maladie sont plus grandes chez les femmes. Par contre, les autres facteurs de risque ne sont pas significatifs. Ceci peut s'expliquer par les biais de sélection : le sous-groupe étudié est composé uniquement de deux classes de l'échelle KL sur cinq, d'une majorité de patients en surpoids ou obésité modérée, et de patients dans la tranche d'âge 55-70 ans. Ces caractéristiques particulières de l'étude peuvent expliquer que ces facteurs de risque ne soient pas mis en évidence par nos analyses, elles ne sont pas représentatives de la population.

Conclusion Notre sélection met en avant une caractéristique connue de l'étude SEKOIA : les patients sont sélectionnés pour avoir une arthrose importante dans la zone fémoro-tibiale médiale (MFT). On retrouve ainsi 6 scores cumulés de critères IRM dans cette zone. Le modèle final montre à quel point l'arthrose impacte l'ensemble de l'articulation avec des scores portant sur le cartilage, les oedèmes et l'attrition osseuse, les kystes, les dommages du ménisque, les ostéophytes marginaux et l'état des ligaments. Deux biomarqueurs sanguins et urinaires ont été sélectionnés : le C-Terminal Propeptide of type II procollagen (P2CP) et l'hyaluronic acid (HYALACID). Nous pouvons supposer que ces biomarqueurs présentent une concentration qui augmente chez les patients dont l'articulation se dégrade. Le P2CP sera produit pour fabriquer du collagène et l'HYALACID pour lubrifier l'articulation. Nos analyses de sensibilité montrent que la sélection est robuste. Bien que le jeu de données ne soit pas suffisant pour avoir des intervalles de confiance satisfaisants, le modèle a permis de trouver des facteurs explicatifs des différences de JSW entre patients au cours du temps.

5.3 Le projet FNIH de l'étude OAI

Lancé en 2002 et achevé en 2015, l'Initiative Osteoarthritis (OAI) représente une collaboration du National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS), du National Institute on Aging (NIA) et du secteur privé, travaillant ensemble pour améliorer l'efficacité du développement des médicaments et des essais cliniques pour le traitement de l'arthrose. L'étude a créé une ressource de recherche d'accès public qui permet à la communauté scientifique dans le monde entier d'examiner et d'analyser les radiographies du genou, les rayons X, IRM et une variété d'autres données cliniques afin de mieux prédire la progression de l'arthrose et de développer de nouveaux traitements pour ralentir la progression ou modifier cette maladie invalidante, qui touche plus de 15 millions de personnes aux États-Unis. Parmi les nombreux efforts de recherche rendus possibles par cette ressource, nous étudions le projet FNIH : Biomarkers Consortium Osteoarthritis Biomarkers Project.

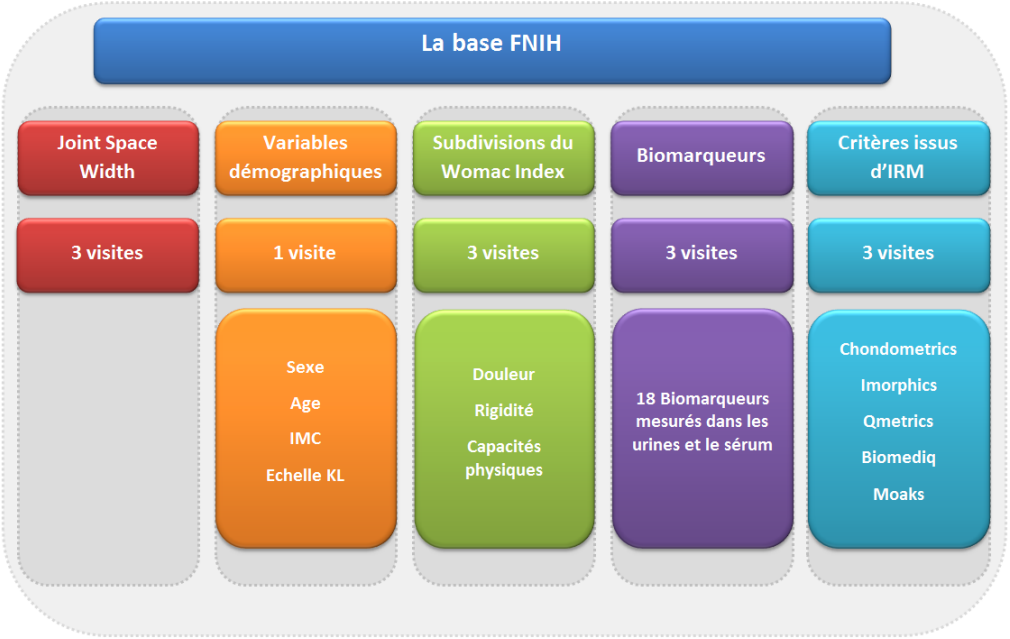


FIGURE 5.14 – Présentation de la base FNIH

5.3.1 La base de données

La base de données est composée de 600 patients pour lesquels de nombreuses variables ont été mesurées sur 3 visites annuelles. Tout comme pour l'étude SEKOIA (figure 5.4), la base de données FNIH est composée de notre variable réponse, la largeur de l'espace articulaire, de variables démographiques, des scores du Womac, de biomarqueurs et de mesures issues d'IRM (figure 5.14). Ces critères sont mesurés selon différentes méthodes : Chondometrics, Imorphics, Biomediq, Qmetrics et le MRI OsteoArthritis Knee Score (MOAKS Hunter et al. [2011]) similaire au WORMS. La base de données est restreinte aux mêmes sources d'informations que celles disponibles pour SEKOIA : la réponse, les variables démographiques, les critères du MOAKS et les biomarquers sanguins et urinaires. Les données issues des autres bases sont étudiées en analyse supplémentaire.

La largeur de l'espace articulaire mesurée sur une plus grande base de données présente une distribution plus proche d'une gaussienne comme le montre le graphique 5.15. Cette base présente des largeurs de l'espace articulaire plus grandes qui correspondent aux patients pour lesquels la maladie n'est pas encore à un stade avancé.

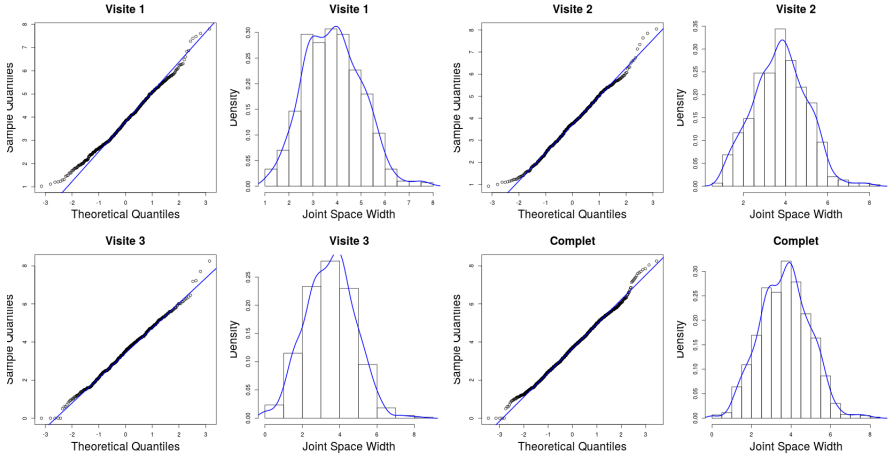


FIGURE 5.15 – Q-Q plots et histogrammes du minimum de la largeur de l'espace articulaire de la base FNIH

Pour les mêmes raisons que pour l'étude SEKOIA, uniquement les sous-scores du Womac mesurés à la première visite seront utilisés.

Les variables démographiques Comme pour la base SEKOIA, on retrouve l'âge, le sexe, l'échelle de Kellgren-Lawrence et l'IMC répartis de la manière suivante :

	Stade I	Stade II	Stade III	Total
Femme	46 (7.7%)	202 (33.7%)	105 (17.5%)	353 (58.8%)
Homme	29 (4.8%)	104 (17.3%)	114 (19%)	247 (41.2%)
Total	75 (12.5%)	306 (51%)	219 (36.5%)	600 (100%)

TABLE 5.8 – Répartition du sexe et de l'échelle de Kellgren-Lawrence dans la base FNIH

Une partie des patients inclus dans l'étude FNIH présentent une arthrose peu avancée, ce qui explique qu'un stade de plus de l'échelle de Kellgren-Lawrence soit représenté. On remarque que le sexe est une variable plus équilibrée dans la base FNIH avec 58.8% de femmes (72.3% pour la base SEKOIA). En ce qui concerne l'âge et l'IMC, les répartitions sont les suivantes :

	Âge	IMC
Moyenne (sd)	61.6 (8.9)	30.7 (4.8)
Médiane (Q_1, Q_3)	61 (54, 69)	30.2 (27.5, 33.6)

TABLE 5.9 – Moyenne et médiane de l'âge et de l'IMC dans la base FNIH

Les critères issus d'IRM Des critères similaires au WORMS, comme détaillé dans le tableau 5.3 sont mesurés dans l'étude FNIH. Ils sont mesurés selon le MOAKS, ce qui modifie légèrement les règles de calculs et le codage des variables. Nous utiliserons les scores cumulés dans les mêmes zones que pour l'étude SEKOIA. On passe alors de 108 variables à 47 variables mesurées selon le MOAKS (contre 33 dans la base SEKOIA).

D'autres critères ont été mesurés par différents laboratoires :

- Chondometrics : 42 variables de volume, épaisseur du cartilage et pourcentage de surface osseuse dénudée
- Qmetrics : 12 variables de surfaces osseuses et leurs morphologies
- Imorphics : 12 variables de surface osseuse sous-chondrale
- Biomediq : 7 variables de volumes de cartilage et du ménisque

Les biomarqueurs 18 biomarqueurs sont mesurés parmi lesquels 7 sont urinaires et 11 sanguins. Nous avons utilisé la transformation Box-Cox la plus adaptée selon les don-

nées observées de la base FNIH sans prendre en compte les transformations utilisées dans SEKOIA. Le tableau 5.10 liste les transformations choisies.

Biomarqueurs	Transformations
MMP3	$\log(x)$
CSEPI846	$\log(x)$
HYALACID	$1/x$
SCTX	$\log(x)$
C12C	$\sqrt{x}$
C2C	$\log(x)$
HCOMP	$\log(x)$
COL21N2	$\log(x)$
P2CP	$\log(x)$
NTX	$\log(x)$
PIIANP	x
UCTXC	$\log(x)$
UC12C	$\log(x)$
UC2C	$\log(x)$
Uα	$1/x$
Uβ	$\log(x)$
UCOL21N2	$\log(x)$

TABLE 5.10 – Transformation choisie pour chaque biomarqueur de l'étude FNIH. Les marqueurs en caractères gras sont également dosés dans SEKOIA

5.3.2 Gestion des données manquantes

La base de données FNIH ne souffre pas d'attrition. Elle présente néanmoins des données manquantes ponctuelles et certains biomarqueurs sont soumis à un seuil de détection. Les données manquantes représentent 2.53% des données et sont réparties de la manière suivante :

- aucune données manquantes pour la réponse, le womac et les variables démographiques
- en moyenne 6.15% de données sous le seuil de détection et 0.28% de données manquantes pour les biomarqueurs
- 1% de données manquantes pour la base Imorphics
- 1.06% de données manquantes pour la base Biomediq
- 1.06% de données manquantes pour la base Chondometrics
- 2.46% de données manquantes pour la base Qmetrics

- 1.38% de données manquantes pour la base MOAKS

Parmi les biomarqueurs, les taux de données non observées car soumises à un seuil de détection sont variables passant de 0% à plus de 70% pour certaines. Le tableau 5.11 montre les différences d'un biomarqueur à l'autre.

Biomarqueurs	Taux de données sous le seuil de détection
MMP3	15%
CSEPI846	75.1%
HYALACID	64.7%
SCTX	06.8%
C12C	0.2%
C2C	0%
HCOMP	0.3%
COL21N2	1.1%
CPII	0%
NTX	0.5%
PIIANP	0%
UCTXC	15.1%
UC12C	44.7%
UC2C	5.8%
Uα	53.4%
Uβ	25.8%
UCOL21N2	36.6%

TABLE 5.11 – Taux de données manquantes car sous le seuil de détection pour les biomarqueurs de la base FNIH

Imputation Multiple Nous avons utilisé la même méthode que pour l'imputation multiple de la base SEKOIA. Nos analyses sont basées sur 10 imputations effectuées à l'aide du package `mice` de Van Buuren and Groothuis-Oudshoorn [2011]. Les fonctions d'imputations utilisées sont le *predictive mean matching* pour les variables continues, la régression logistique pour les variables binaires, et la régression Tobit pour les variables soumises à un seuil de détection. La réponse et la variable *visite* qui représente le numéro de la visite effectuée sont imposées comme prédicteurs pour chaque variable qui présente des données manquantes. Nous avons utilisé la fonction `quickpred` afin de n'utiliser que des prédicteurs ayant au moins 30% de cases utilisables et un coefficient de corrélation supérieur à 0.1.

5.3.3 Sélection de variables par MI-PGEE

Nous avons restreint la base de données aux mêmes sources d'information que la base SEKOIA : les facteurs de risque, les biomarqueurs et les critères IRM mesurés par MOAKS. Le MOAKS propose trois mesures pour les oedèmes de la moelle osseuse : le nombre d'oedèmes, le pourcentage d'os souffrant d'oedèmes et la taille des oedèmes. Ces variables sont très corrélées. Aussi, nous les utilisons toutes pour l'imputation multiple mais nous ne conservons que les variables de taille qui correspondent aux mêmes définitions que dans le WORMS utilisé par l'étude SEKOIA. En conséquence, les 43 scores cumulés sont réduits à 35.

La base de données analysée est donc composée de 61 variables dont la réponse, 4 facteurs de risque, 3 scores du womac, 18 biomarqueurs et 35 critères issus d'IRM. La figure 5.16 représente l'évolution des coefficients estimés et du BIC.

Comme pour l'étude de la base SEKOIA, nous avons tracé cette évolution en fonction du taux de réduction s. On obtient un BIC optimal pour $s = 0.167$ soit un taux de réduction moins fort que pour l'étude SEKOIA. Les coefficients estimés par MI-PGEE sont plus importants que ceux estimés dans la base SEKOIA. De plus, le MI-PGEE sélectionne plus de variables. Le modèle final est composé de 15 variables dont les coefficients associés sont estimés grâce aux règles de Rubin (table 5.12).

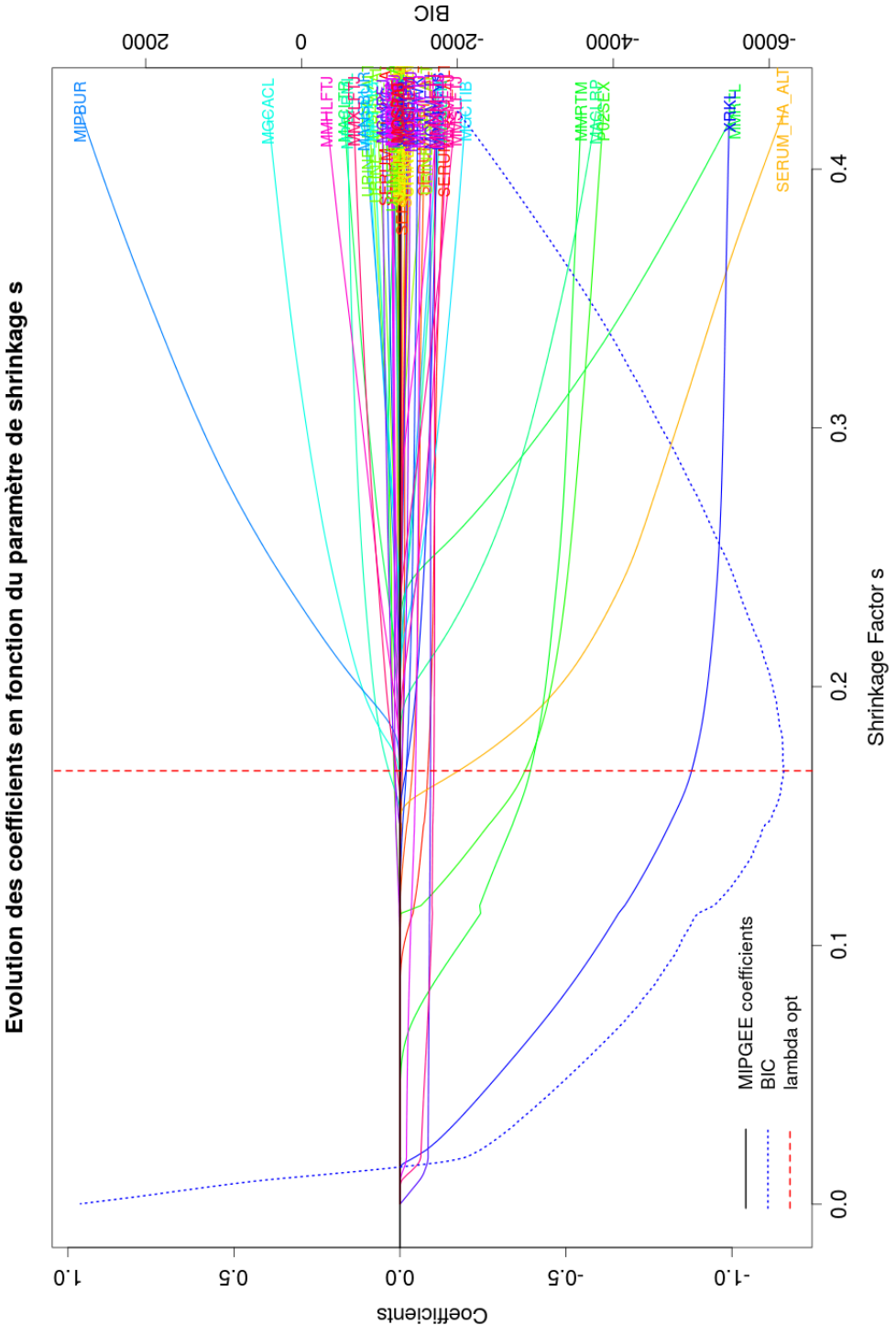


FIGURE 5.16 – Chemin des coefficients estimés par MI-PGEE sur la base FNIH

Variables	$\hat{\beta}_j$ (et)	$\lvert\hat{\beta}_j\rvert/et$
C2C Serum	−0.096 (0.179)	0.535
Coll2-1NO2 Serum	−0.023 (0.055)	0.426
P2CP Serum	−0.082 (0.132)	0.621
HYALACID	−0.957 (1.063)	0.900
Beta Urine	−0.048 (0.074)	0.652
Sexe*	0.559 (0.076)	7.345
Échelle KL*	−0.997 (0.105)	9.483
Épanchement	−0.038 (0.031)	1.254
Kyste poplité	−0.101 (0.092)	1.101
Déchirure de la racine méniscale médiale*	−0.605 (0.217)	2.791
Cartilage Morphologie (région MFT)*	−0.089 (0.018)	5.077
Oedème de la moelle osseuse (région LFT)	−0.035 (0.042)	0.816
Dommage méniscal (région MFT)*	−0.065 (0.016)	4.122
Dommage méniscal (région LFT)	−0.028 (0.032)	0.878
Extrusion du ménisque (région MFT)*	−0.099 (0.033)	2.970

TABLE 5.12 – Sélection obtenue par MI-PGEE sur la base FNIH, estimation (écart type) par MI-GEE. Le symbole * représente les variables pour lesquelles l'intervalle de confiance ne comprend pas zéro. Les régions MFT et LFT représentent les zones médiale et latérale fémoro-tibiale

Le modèle final est composé de 5 biomarqueurs sanguins et urinaires, deux facteurs de risque, la présence d'épanchement, de kystes, de déchirure du ménisque ainsi que 8 scores cumulés. Les patients n'ont pas été sélectionnés pour avoir une arthrose sur la zone médiale contrairement à l'étude SEKOIA. Cette absence de biais de sélection explique que les scores sélectionnés soient mesurés dans diverses régions (médiale et latérale). Parmi ces scores, 4 portent sur l'état du ménisque ce qui montre son importance dans la maladie de l'arthrose. Il a été démontré qu'une extrusion du ménisque est fortement liée à la maladie et plus particulièrement à une arthrose douloureuse (Gale et al. [1999]; Lerer et al. [2004]). De nombreux scores cumulés différents sont sélectionnés, et le modèle final est composé de nombreuses variables qui reflètent les différents aspects de la maladie.

Tous les coefficients estimés sont négatifs sauf pour la variable sexe comme dans l'analyse de la base SEKOIA. Avoir des scores cumulés importants, des kystes, un épanchement ou une déchirure du ménisque est en lien avec une largeur de l'espace articulaire plus faible. Parmi les variables sélectionnées, on retrouve l'acide hyaluronique, le C-Terminal P Propeptide of type II procollagen, les scores cumulés de cartilage, d'oedème et de dommage

méniscal sélectionnés sur SEKOIA ainsi que l'effet protecteur du sexe. Le score cumulé de cartilage morphologie dans la zone médiale, seule variable significative de la sélection sur SEKOIA, est sélectionné et significatif sur le projet FNIH.

Bien que les scores sélectionnés soient mesurés dans diverses régions, les seuls à ressortir significatifs avec des coefficients plus importants, sont les scores de la zone médiale fémoro-tibiale, comme pour la base SEKOIA. Il semble donc que la maladie ait plus d'impact dans cette zone. En effet, pendant la marche normale, les lieux d'adduction forcent principalement sur le compartiment médial, ce qui explique que l'articulation ait tendance à plus facilement se dégrader dans cette zone. La perte de largeur de l'espace articulaire dans cette région conduit souvent à une position de jambes arquées. C'est une zone d'intérêt qui peut être une zone précocement atteinte.

Trois autres biomarqueurs sanguins sont sélectionnés, le C2C et le Coll2-1NO2 qui jouent un rôle dans le catabolisme du collagène et le Beta Urine qui est lié à la résorption osseuse. Puisque leurs intervalles de confiance sont larges, aucune conclusion ne peut être établie mais le rôle de ces biomarqueurs dans l'arthrose est étudié par la communauté scientifique (Ameye et al. [2007]; Conrozier et al. [2008]). Le modèle final est composé de nombreuses variables dont 6 obtiennent un intervalle de confiance significatif. Certains coefficients sont associés à un écart type du même ordre que l'estimation. On retrouve les problématiques de l'étude SEKOIA. Cette analyse confirme une de nos hypothèses : une multitude de critères est en lien avec la largeur de l'espace articulaire et chacune est associée à un faible impact.

Parmi les variables mesurées pour le projet FNIH mais pas pour l'étude SEKOIA, on retrouve les biomarquers Coll21NO2 dans le sérum et le Beta dans les urines ainsi que la présence d'épanchement, de kystes poplité, de déchirure de la racine méniscale et d'extrusion du ménisque. Ces deux dernières variables admettent un intervalle de confiance significatif, elles apportent de l'information. Dans les futures études internes, ajouter la mesure de ces critères pourrait être d'intérêt pour les cliniciens.

Remarques Les populations observées dans les bases FNIH et SEKOIA ne sont pas les mêmes que ce soit sur le plan démographique (une étude Européenne et une Nord Améri-

caine) ou le plan médicale (répartition différente du sexe, mais aussi niveau d'avancement de la maladie) ce qui explique que les sélections soient différentes sur les deux bases.

5.3.4 Analyses de sensibilité

Nous avons réalisé quatre analyses de sensibilité :

1. en imputant les biomarqueurs soumis à un seuil de détection au seuil
2. en utilisant les biomarqueurs sanguins et urinaires sans transformation Box-Cox
3. en intégrant les biomarqueurs qui ont plus de 30% d'observations sous le seuil de détection en binaire (1 au dessus du seuil, 0 sinon)
4. en imposant les facteurs de risque que sont l'âge, le sexe, l'échelle KL et l'IMC

Les résultats sont résumés dans le tableau 5.13. La sélection est inchangée que l'on utilise l'imputation multiple, l'imputation simple au seuil, ou les indicatrices au dessus du seuil. Lorsque les biomarqueurs sans transformation sont utilisés, aucun n'est sélectionné. Ce résultat est semblable à celui obtenu dans les analyses de sensibilité menées sur la base SEKOIA. Les transformations permettent d'homogénéiser les écarts de concentrations entre patients ce qui permet de mettre plus facilement en évidence les relations entre biomarqueurs et largeur de l'espace articulaire. Lorsque l'on impose les facteurs de risque connus, l'âge et l'IMC sont inclus au modèle avec des coefficients très proches de zéro. La variable dommage du ménisque cumulée dans la région latérale n'est alors plus sélectionnée. Les estimations des paramètres d'une analyse de sensibilité à l'autre sont très similaires, la sélection est robuste.

	mice	seuil	Sans BC	Binaire	Clinique
C2C Serum	−0.096 (0.179)	−0.099 (0.173)	∅	−0.096 (0.178)	−0.113 (0.173)
Coll2-1NO2 Serum	−0.023 (0.055)	−0.023 (0.055)	∅	−0.021 (0.055)	−0.017 (0.056)
P2CP Serum	−0.082 (0.132)	−0.076 (0.130)	∅	−0.086 (0.132)	−0.066 (0.135)
HYALACID	−0.957 (1.063)	−3.798 (2.747)	∅	−0.066 (0.044)	−1.220 (1.058)
Beta Urine	−0.048 (0.074)	−0.061 (0.095)	∅	−0.034 (0.031)	−0.045 (0.076)
Sexe	0.559* (0.076)	0.563* (0.075)	0.556* (0.075)	0.561* (0.076)	0.617* (0.077)
Echelle KL	−0.997* (0.105)	−0.995* (0.106)	−0.989* (0.108)	−0.996* (0.106)	−1.023* (0.105)
Epanchement	−0.038 (0.031)	−0.036 (0.030)	−0.038 (0.030)	−0.037 (0.031)	−0.052 (0.029)
Kyste poplité	−0.101 (0.092)	−0.105 (0.092)	−0.107 (0.093)	−0.099 (0.093)	−0.105 (0.092)
Déchirure méniscale médiale	−0.605* (0.217)	−0.606* (0.214)	−0.606* (0.219)	−0.606* (0.220)	−0.591* (0.208)
Cartilage Morphologie MFT	−0.089* (0.018)	−0.089* (0.018)	−0.089* (0.018)	−0.089* (0.017)	−0.090* (0.017)
Oedème de la moelle osseuse LFT	−0.035 (0.042)	−0.035 (0.042)	−0.035 (0.042)	−0.036 (0.042)	−0.031 (0.040)
Dommage méniscal MFT	−0.065* (0.016)	−0.065* (0.016)	−0.065* (0.016)	−0.065* (0.016)	−0.058* (0.017)
Dommage méniscal LFT	−0.028 (0.032)	−0.027 (0.031)	−0.028 (0.032)	−0.028 (0.032)	∅
Extrusion du ménisque MFT	−0.099* (0.033)	−0.098* (0.034)	−0.100* (0.034)	−0.098* (0.033)	−0.108* (0.032)
IMC	∅	∅	∅	∅	0.006 (0.007)
Age	∅	∅	∅	∅	0.002 (0.004)

TABLE 5.13 – Sélection par MI-PGEE et estimations (écarts types) par MI-GEE pour les différentes analyses de sensibilité. Le symbol ∅ indique que la variable n'a pas été sélectionnée

Analyse supplémentaire Lorsque l'on inclut les autres bases de critères issus d'IRM (Chondometrics, Imorphics, Qmetrics et Biomediq), on obtient une base de 118 variables dont 92 sont issues d'IRM. Le MI-PGEE sélectionne les mêmes biomarqueurs sanguins et urinaires, les mêmes critères du MOAKS ainsi que deux variables de la base Qmetrics : le ratio de signal entre surface osseuse et cartilage sur le fémur latéral d'une part et sur le fémur médial d'autre part. Toutes les deux sont associées à des coefficients négatifs, ainsi que 4 variables de la base Chondometrics (surface de cartilage et surface d'os recouvert de cartilage dans les zones médiale centrale et médiale latérale du fémur) ce qui conduit à un modèle composé de 21 variables.

Conclusion Ces analyses montrent que l'arthrose est une maladie complexe qui impacte de nombreuses structures de l'articulation. Deux biomarqueurs sanguins sont sélectionnés dans la base SEKOIA et FNIH : l'acide hyaluronique et le C-Terminal P Propeptide of type II procollagen. Ces biomarqueurs sont en lien avec la production de collagène et sont produits en plus grande quantité lorsque l'articulation se dégrade. Leurs implications dans le mécanisme de la maladie doivent être étudiées plus précisément. De même, le score de morphologie du cartilage cumulé dans la zone médiale est sélectionné dans les deux études avec un intervalle de confiance significatif. De nombreux critères sont sélectionnés dans la zone médiale de l'articulation, son implication dans le processus de dégradation de l'articulation doit être explorée. Cette région de l'articulation est à cibler en priorité sur les examens d'imagerie.

Le critère principal de l'étude : la largeur de l'espace articulaire minimale est un critère controversé. L'arthrose est une maladie complexe dans laquelle la douleur et le ressenti du patient sont très importants. Certains patients ne souffrent pas, ne sont pas limités par la maladie alors que leur articulation est dégradée avec une largeur de l'espace articulaire très fine. D'autres souffrent beaucoup et sont handicapés au quotidien mais présentent des largeurs de l'espace articulaire dans la norme. Certaines études proposent alors de combiner scores de douleur et largeur de l'espace articulaire et de s'intéresser à la notion de progesseur douleur et progresseur radio.

Conclusion et perspectives

Les principales questions qui ont motivé cette thèse sont : Comment prendre en compte les sorties d'étude ? Comment prendre en compte les variables soumises à un seuil de détection ? Lorsque l'on choisit une méthode d'imputation multiple pour combler les données manquantes la problématique devient alors : Comment intégrer les jeux de données multi-imputées dans la sélection de variables pour GEE ?

Apports de la thèse

Pour répondre à ces questions, nous avons étudié l'impact de différents types de données manquantes sur les GEE et mis au point une fonction d'imputation pour les biomarqueurs soumis à un seuil de détection. Notre fonction est proposée pour l'imputation multiple par algorithme mice. Une fois les jeux de données imputées, nous proposons le MI-PGEE. Cet opérateur est capable de sélectionner le sous-groupe de covariables en lien avec la réponse tout en intègrant les jeux de données multi-imputées et les corrélations intra-patient.

Imputation des observations sous le seuil de détection Notre fonction d'imputation propose d'utiliser la régression Tobit afin de prendre en compte les observations sous le seuil dans l'estimation des paramètres. Cette méthode permet une modélisation complexe du biomarqueur et les estimateurs obtenus sont non biaisés. Combinée à l'algorithme mice, l'imputation proposée permet d'intégrer l'information d'autres covariables. On peut alors utiliser l'information portée par toute autre variable, qui présente ou non des données manquantes, qu'elle soit binaire, continue ou de comptage. Nous avons montré par simulations que la méthode pouvait être avantageuse par rapport aux méthodes de référence avec un

biais relatif absolu proche de celui obtenu sans données manquantes.

Sélection de variables par MI-PGEE La méthode de sélection de variables proposée permet d'intégrer les données manquantes grâce aux jeux de données multi-imputées. De plus, elle permet de prendre en compte les corrélations intra-patient via la matrice de corrélation de travail. Nous proposons un opérateur capable de réaliser une sélection consistante à travers les jeux de données muti-imputées, ce qui permet l'utilisation des règles de Rubin pour l'inférence. Comparé à son analogue sur imputation simple, le MI-PGEE admet de meilleurs résultats et montre sa robustesse aux données manquantes MAR.

Appliquées à l'étude de l'arthrose du genou, ces méthodologies ont permis de démontrer l'implication de la maladie dans une multitude de mécanismes de l'articulation. Nous avons mis en avant le lien entre certains biomarqueurs sangins et urinaires, des critères d'imagerie et la largeur de l'espace articulaire au cours du temps.

Limites rencontrées

Lorsque le taux de données manquantes est important, les estimateurs peuvent être biaisés. Dans les bases de données, on ne sait pas toujours quel est le meilleur modèle à utiliser pour l'imputation, le choix des prédicteurs est une question délicate (Van Buuren et al. [2006]). Comme pour toute méthode statistique, la méthode d'imputation proposée pour les variables soumises à un seuil de détection dépend du bon choix du prédicteurs. Nos études par simulations montrent qu'à partir d'un certain taux de données manquantes, notre méthode peut produire des estimateurs biaisés.

La méthode du MI-PGEE présente des limites quant à la sélection de coefficients *faibles*. Il est difficile de mettre en place une imputation multiple automatique avec une sélection précise des prédicteurs pour une étude par simulations. Or, un faible coefficient combiné à du bruit rend difficile la sélection de la variable associée. Lors de l'analyse de bases de

données réelles, lorsque le nombre de variables observées est raisonnable, les prédicteurs doivent être choisis en concertation avec les experts cliniques.

Les simulations du chapitre 4 sont conduites avec une réponse continue puisque c'est ce que nous observons sur nos bases de données. Toutefois, la méthode peut facilement être adaptée aux réponses discrètes. Nous utilisons une structure de corrélation auto-régressive puisque que nous sommes dans un contexte longitudinal, mais l'utilisateur est libre de choisir la structure de corrélation qui s'ajuste au mieux aux données qu'il traite. La section 1.2.5 liste quelques critères qui permettent de sélectionner la *meilleure* structure. Cependant, il est recommandé (Wang and Carey [2003]) de ne choisir la structure de corrélation qu'une fois la structure de la moyenne définie. La méthode du MI-PGEE peut être utilisée avec une structure de corrélation définie à l'avance à l'aide d'*a priori* de l'utilisateur et des spécialistes. Une fois le modèle de régression déterminé, choisir la matrice de corrélation peut se faire à l'aide d'un critère de qualité. Cette méthode peut également être appliquée sur des jeux de données corrélées qui ne sont pas longitudinaux.

Perspectives

Imputation par régression Tobit Nous avons proposé d'utiliser la méthode d'imputation par classe afin de ne pas alourdir cette dernière avec des effets aléatoires. Une extension possible est donc d'utiliser la forme longitudinale de la régression Tobit pour l'imputation des valeurs sous le seuil. Comparer les résultats issus de cette méthode à ceux obtenus par imputation par classe permettrait d'évaluer le gain de précision par rapport à l'augmentation du temps de calcul.

Hypothèses de simulation Nous avons formulé des hypothèses MCAR et MAR pour nos simulations ; hypothèses qui ne sont pas toujours réalistes. L'imputation multiple peut être utilisée pour des données MNAR (Molenberghs et al. [2014] chapitre 9), et de nombreuses études de sensibilité permettent de comprendre l'impact de telles hypothèses sur

les inférences. Une étude par simulations avec de telles données manquantes pourrait être utile. Nous pourrions imaginer une étude par simulations où les covariables sont soumises à un seuil de détection et évaluer les performances du MI-PGEE en fonction du taux de données manquantes, comparées à celles du PGEE avec une imputation au seuil.

Sélection de variables par MI-PGEE Face aux difficultés rencontrées en présence de faibles coefficients, une extension possible serait l'utilisation de pénalité de type SCAD (Fan and Li [2001]). Il faudrait pour cela réécrire la formulation de la pénalité pour l'utiliser de la même manière que nous avons utilisé la pénalité Group-LASSO. Ainsi, la réduction des coefficients serait proportionnelle à leur importance. Nous avons vu que le MI-PGEE est mis en difficulté lorsque les corrélations entre variables actives sont importantes. Cette limite est connue pour le LASSO et le Group-LASSO (Zou and Hastie [2005]) pour lesquels il est difficile de sélectionner des groupes sans information *a priori*. Une alternative capable de meilleures performances dans ce domaine est l'utilisation de pénalités combinées. Nous pourrions utiliser une pénalité SCAD par exemple, combinée à une pénalité Ridge comme proposé par Zeng and Xie [2014] et par Blommaert et al. [2014] pour les GEE. De cette façon, nous pourrions réduire les coefficients tout en prenant en compte le *grouping effect* dû aux corrélations. Cette extension combinerait une pénalité Group-SCAD et une pénalité Group-Ridge à définir.

Bibliographie

H. Akaike. Information theory and extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, pages 267–281, 1973. 44, 47

L.G. Ameye, M. Deberg, M. Oliveira, A. Labasse, J.M. Aeschlimann, and Y. Henrotin. The chemical biomarkers c2c, coll2-1, and coll2-1no2 provide complementary information on type ii collagen catabolism in healthy and osteoarthritic mice. *Arthritis & Rheumatism*, 56(10) :3336–3346, 2007. 142

B. C. Arnold and D. Strauss. Pseudolikelihood estimation : Some examples. *Sankhyā : The Indian Journal of Statistics, Series B*, pages 233–243, 1991. 37

V. Audigier, F. Husson, and J. Josse. Multiple imputation for continuous variables using a bayesian principal component analysis. *Journal of Statistical Computation and Simulation*, 86(11) :2140–2156, 2016. 81

P. W. Bernhardt. *Statistical Modeling with Covariates Subject to Detection Limits*. PhD thesis, North Carolina State University, 2013. 87

A. Blommaert, N. Hens, and P. Beutels. Data mining for longitudinal data under multicollinearity and time dependence using penalized generalized estimating equations. *Computational Statistics & Data Analysis*, 71 :667–680, 2014. 59, 150

A.-L. Boulesteix and W. Sauerbrei. Added predictive value of high-throughput molecular data to clinical data and its validation. *Briefings in Bioinformatics*, 12(3) :215–229, 2011. 132

M. J. Bradburn, J. J. Deeks, J. A. Berlin, and A. Russell L. Much ado about nothing : a

comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine*, 26(1) :53–77, 2007. 83

L. Breiman et al. Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24(6) :2350–2383, 1996. 54

R. W. Browne and B. W. Whitcomb. Procedures for determination of detection limits : Application to high-performance liquid chromatography analysis of fat-soluble vitamins in human serum. *Epidemiology*, 21(4) :S4–S9, 2010. 84

P. Bühlmann and S. Van De Geer. *Statistics for High-Dimensional Data : Methods, Theory and Applications*. Springer Science & Business Media, New York, 2011. 109

E. Cantoni, J. M. Flemming, and E. Ronchetti. Variable selection for marginal longitudinal generalized linear models. *Biometrics*, 61(2) :507–514, 2005. 46, 47, 65

E. Cantoni, C. Field, J. M. Flemming, and E. Ronchetti. Longitudinal variable selection by cross-validation in the case of many covariates. *Statistics in Medicine*, 26(4) :919–930, 2007. 46

J. R. Carpenter, M. G. Kenward, and S. Vansteelandt. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, 169(3) :571–584, 2006. 72

R.J. Carroll, S. Wang, D.G. Simpson, A.J. Stromberg, and D. Ruppert. The sandwich (robust covariance matrix) estimator. 1998. 44

Q. Chen and S. Wang. Variable selection for multiply-imputed data with application to dioxin exposure study. *Statistics in Medicine*, 32(21) :3646–3659, 2013. 100, 101, 103, 105

T. Conrozier, A.R. Poole, F. Ferrand, P. Mathieu, F. Vincent, M. Piperno, C. Verret, M. Ionescu, and E. Vignon. Serum concentrations of type ii collagen biomarkers (c2c, c1, 2c and cpii) suggest different pathophysiologies in patients with hip osteoarthritis. *Clinical and Experimental Rheumatology*, 26(3) :430, 2008. 142

C. Cooper, J.-Y. Reginster, R. Chapurlat, C. Christiansen, H. Genant, N. Bellamy, W. Bensen, F. Navarro, J. Badurski, E. Nasonov, et al. Efficacy and safety of oral strontium ranelate for the treatment of knee osteoarthritis : Rationale and design of randomised, double-blind, placebo-controlled trial. *Current Medical Research & Opinion*, 28(2) :231–239, 2012. 116

M. Davidian, A. A. Tsiatis, and S. Leon. Semiparametric estimation of treatment effect in a pretest–posttest study with missing data. *Statistical Science : a review journal of the Institute of Mathematical Statistics*, 20(3) :261, 2005. 73

C. S. Davis. *Statistical Methods for the Analysis of Repeated Measurements*. Springer Science & Business Media, New York, 2002. 38

P. Diggle, D. Farewell, and R. Henderson. Analysis of longitudinal data with drop-out : Objectives, assumptions and a proposal. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 56(5) :499–550, 2007. 72

J. J. Dziak and R. Li. Variable selection with penalized generalized estimating equations. Technical report, Pennsylvania State University, The Methodology Center, 2006. 49, 54, 65, 112

J. J. Dziak and R. Li. An overview on variable selection for longitudinal data. *Quantitative Medical Data Analysis*, pages 5–24, 2007. 45, 47

B. Efron. Bootstrap methods : Another look at the jackknife. *Annals of Statistics*, 7 :1–26, 1979. 44

B. Efron. How biased is the apparent error rate of a prediction rule ? *Journal of the American Statistical Association*, 81(394) :461–470, 1986. 46

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993. 87

B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al. Least angle regression. *The Annals of Statistics*, 32(2) :407–499, 2004. 57, 109

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456) :1348–1360, 2001. 57, 60, 62, 150

G. M. Fitzmaurice, N. M. Laird, and J. H. Ware. *Applied Longitudinal Analysis*, volume 998. John Wiley & Sons, 2012. 33

L. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2) :109–135, 1993. 54

J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer Series in Statistics Springer, 2001. 56

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1) :1, 2010. 58

W. J. Fu. Penalized regressions : the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3) :397–416, 1998. 54, 57, 63

W. J. Fu. Penalized estimating equations. *Biometrics*, 59(1) :126–132, 2003. 49, 59, 64, 106, 112

W. J. Fu. Nonlinear gcv and quasi-gcv for shrinkage models. *Journal of Statistical Planning and Inference*, 131(2) :333–347, 2005. 64, 105

D.R. Gale, C.E. Chaisson, S.M.S. Totterman, R.K. Schwartz, M.E. Gale, and D. Felson. Meniscal subluxation : Association with osteoarthritis and joint space narrowing. *Osteoarthritis and Cartilage*, 7(6) :526–532, 1999. 141

M. Garland, J. S. Morris, B. A. Rosner, M. J. Stampfer, V. L Spate, C. J. Baskett, W. C. Willett, and D. J. Hunter. Toenail trace element levels as biomarkers : Reproducibility over a 6-year period. *Cancer Epidemiology Biomarkers & Prevention*, 2(5) :493–497, 1993. 85

S. Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350) :320–328, 1975. 44

J. Geronimi and G. Saporta. The effect of missing visits on gee, a simulation study. In *ASMDA 2015*, pages 269–276, Piraeus, Greece, July 2015. The 16th Conference of the Applied Stochastic Models and Data Analysis International Society. 74

H. Geys, G. Molenberghs, and L. M. Ryan. Pseudolikelihood modeling of multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association*, 94(447) :734–745, 1999. 37

R. O. Gilbert. *Statistical Methods for Environmental Pollution Monitoring*. John Wiley & Sons, 1987. 86

A. Gleit. Estimation for small normal data sets with detection limits. *Environmental Science & Technology*, 19(12) :1201–1206, 1985. 85

M. Gosho. Model selection in the weighted generalized estimating equations for longitudinal data with drop-out. *Biometrical Journal*, 2015. 98

Y. Guo, O. Harel, and R. J. Little. How well quantified is the limit of quantification ? *Epidemiology*, 21(4) :S10–S16, 2010. 84

U. Halekoh, S. Hojsgaard, and J. Yan. The r package geepack for generalized estimating equations. *Journal of Statistical Software*, 15(2) :1–11, 2006. 73, 76, 90

J. W. Hardin and J. M. Hilbe. *Generalized Linear Models and Extensions*. Stata press, 2003. 43

M. Hebiri, S. Van De Geer, et al. The smooth-lasso and other l1+ l2 penalized methods. *Electronic Journal of Statistics*, 5 :1184–1226, 2011. 57

D. R. Helsel. Less than obvious-statistical treatment of data below the detection limit. *Environmental Science & Technology*, 24(12) :1766–1774, 1990. 85

D. R. Helsel and T. A. Cohn. Estimation of descriptive statistics for multiply censored water quality data. *Water Resources Research*, 24(12) :1997–2004, 1988. 85

A. Henningsen. Estimating censored regression models in r using the censreg package. *University of Copenhagen*, 2010. 86

L.-Y Hin and Y.-G. Wang. Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine*, 28(4) :642–658, 2009. 50

L.-Y. Hin, V. J. Carey, and Y.-G. Wang. Criteria for working correlation structure selection in gee. *The American Statistician*, 2012. 50, 51

A. E. Hoerl and R. W. Kennard. Ridge regression : Applications to nonorthogonal problems. *Technometrics*, 12(1) :69–82, 1970. 56

N. J. Horton and S. R. Lipsitz. Multiple imputation in practice : Comparison of software packages for regression models with missing variables. *The American Statistician*, 55 (3) :244–254, 2001. 83

J. P. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, number 1, pages 221–233. L.M. LeCam and J. Neyman, 1967. 44

D. J. Hunter, A. Guermazi, G. H. Lo, A. J. Grainger, P. G. Conaghan, R. M. Boudreau, and F. W. Roemer. Evolution of semi-quantitative whole joint assessment of knee oa : Moaks (mri osteoarthritis knee score). *Osteoarthritis and Cartilage*, 19(8) :990–1002, 2011. 135

J. Josse and F. Husson. Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique*, 153(2) :79–99, 2012. 81

M. G. Kenward and J. Carpenter. Multiple imputation : Current perspectives. *Statistical Methods in Medical Research*, 16(3) :199–218, 2007. 82

M. G. Kenward and G. Molenberghs. Likelihood based frequentist inference when data are missing at random. *Statistical Science*, pages 236–247, 1998. 71

J. P. Klein and M. L. Moeschberger. *Survival Analysis :Ttechniques for Censored and Truncated Data*. Springer Science & Business Media, 2005. 85

M. Lee, L. Kong, and L. Weissfeld. Multiple imputation for left-censored biomarker data based on gibbs sampling method. *Statistics in Medicine*, 31(17) :1838–1848, 2012. 87

D.B. Lerer, H.R. Umans, M.X. Hu, and M.H. Jones. The role of meniscal root pathology and radial meniscal tear in medial meniscal extrusion. *Skeletal Radiology*, 33(10) :569–574, 2004. 141

K.-Y Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1) :13–22, 1986. 38, 39, 43, 44, 70, 71

G. Lin and R. N. Rodriguez. Weighted methods for analyzing missing data with the gee and causaltrt procedures. In *Proceedings of the SAS Global Forum 2014 Conference. URL http ://support. sas. com/resources/papers/proceedings14/SAS166-2014. pdf*, 2014. 73

S. R. Lipsitz, G. Molenberghs, G. M. Fitzmaurice, and J. Ibrahim. Gee with gaussian estimation of the correlations when data are incomplete. *Biometrics*, 56(2) :528–536, 2000. 73

R. J. Little. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431) :1112–1121, 1995. 19, 69, 70, 71

R. J. Little and D. B. Rubin. *Statistical Analysis with Missing Data.* John Wiley & Sons, 1987. 15, 19, 68, 70, 71, 80

J. H. Lubin, J. S. Colt, D. Camann, S. Davis, J. R. Cerhan, R. K. Severson, L. Bernstein, and P. Hartge. Epidemiologic evaluation of measurement data in the presence of detection limits. *Environmental Health Perspectives*, pages 1691–1696, 2004. 85, 86, 87

C. L. Mallows. Some comments on cp. *Technometrics*, 15(4) :661–675, 1973. 44, 46

S. McConnell, P. Kolopack, and A. M. Davis. The western ontario and mcmaster universities osteoarthritis index (womac) : a review of its utility and measurement properties. *Arthritis Care & Research*, 45(5) :453–461, 2001. 117

P. McCullagh. Quasi-likelihood functions. *The Annals of Statistics*, pages 59–67, 1983. 37, 39

W. Q. Meeker and L. A. Escobar. *Statistical Methods for Reliability Data.* John Wiley & Sons, 2014. 85

A. Miller. *Subset Selection in Regression*. Chapman and Hall, 2002. 54

G. Molenberghs, G. Fitzmaurice, M. G. Kenward, A. Tsiatis, and G. Verbeke. *Handbook of Missing Data Methodology*. Chapman and Hall, 2014. 67, 71, 72, 82, 149

K. G.M. Moons, R. A.R.T. Donders, T. Stijnen, and F. E. Harrell. Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology*, 59(10) :1092–1101, 2006. 83

D. J. Moschandreas, S. Karuchit, Y. Kim, H. Ari, M. D. Lebowitz, M. K. O'Rourke, S. Gordon, and G. Robertson. On predicting multi-route and multimedia residential exposure to chlorpyrifos and diazinon. *Journal of Exposure Analysis & Environmental Epidemiology*, 11(1), 2001. 85

J. A. Nelder and R. J. Baker. Generalized linear models. *Encyclopedia of Statistical Sciences*, 1972. 34

L. Nie, H. Chu, C. Liu, S. R. Cole, A. Vexler, and E. F. Schisterman. Linear regression with an independent variable subject to a detection limit. *Epidemiology (Cambridge, Mass.)*, 21(Suppl 4) :S17, 2010. 85

W. Pan. Akaike's information criterion in generalized estimating equations. *Biometrics*, 57(1) :120–125, 2001a. 47, 48, 65, 98

W. Pan. Model selection in estimating equations. *Biometrics*, 57(2) :529–534, 2001b. 46

W. Pan. On the robust variance estimator in generalised estimating equations. *Biometrika*, 88(3) :901–906, 2001c. 44

C. G. Peterfy, A. Guermazi, S. Zaim, P. F. J. Tirman, Y. Miaux, D. White, M. Kothari, Y. Lu, K. Fye, S. Zhao, et al. Whole-organ magnetic resonance imaging score (worms) of the knee in osteoarthritis. *Osteoarthritis and Cartilage*, 12(3) :177–190, 2004. 119

J. S. Preisser, K. K. Lohman, and P. J. Rathouz. Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in Medicine*, 21(20) :3035–3054, 2002. 73, 98

B. F. Qaqish. A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika*, 90(2) :455–463, 2003. 76

C. R. Rao and H. Toutenburg. *Linear Models*. Springer, 1995. 32

J. N. K. Rao. On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91(434) :499–506, 1996. 80

J.-Y. Reginster, J. Badurski, N. Bellamy, W. Bensen, R. Chapurlat, X. Chevalier, C. Christiansen, H. Genant, F. Navarro, E. Nasonov, et al. Efficacy and safety of strontium ranelate in the treatment of knee osteoarthritis : Results of a double-blind, randomised placebo-controlled trial. *Annals of the Rheumatic Diseases*, 2012. 116

D. B. Richardson and A. Ciampi. Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *American Journal of Epidemiology*, 157(4) :355–363, 2003. 85

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427) :846–866, 1994. 72, 80

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429) :106–121, 1995. 72

A. Rotnitzky and N. P. Jewell. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, 77(3) :485–497, 1990. 50, 51

A. Rotnitzky, J. M. Robins, and D. O. Scharfstein. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93(444) :1321–1339, 1998. 72

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3) :581–592, 1976. 15, 19, 68, 70, 71

D. B. Rubin. Multiple imputation for nonresponse in surveys. *Wiley Series in Probability and Statistics*, 1987. 72, 81, 87, 97, 100

D. B. Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434) :473–489, 1996. 82

J. L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapam & Hall, 1997. 82

J. L. Schafer. Multiple imputation : a primer. *Statistical Methods in Medical Research*, 8 (1) :3–15, 1999. 82

J. L. Schafer and J. W. Graham. Missing data : Our view of the state of the art. *Psychological Methods*, 7(2) :147, 2002. 71

D. O. Scharfstein, A. Rotnitzky, and James M. R. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448) :1096–1120, 1999. 73

G. Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2) : 461–464, 1978. 44, 48

S. R. Seaman, I. R. White, A. J. Copas, and L. Li. Combining multiple imputation and inverse-probability weighting. *Biometrics*, 68(1) :129–137, 2012. 96

C.-W. Shen and Y.-H. Chen. Model selection for generalized estimating equations accommodating dropout missingness. *Biometrics*, 68(4) :1046–1054, 2012. 97

C.-W. Shen and Y.-H. Chen. Model selection of generalized estimating equations with multiply imputed longitudinal data. *Biometrical Journal*, 55(6) :899–911, 2013. 99

C.-W. Shen and Y.-H. Chen. Model selection for marginal regression analysis of longitudinal data with missing observations and covariate measurement error. *Biostatistics*, 16(4) : 740–753, 2015. 98

J. Shults, W. Sun, X. Tu, K. Kim, J. Amsterdam, J. M. Hilbe, and T. Ten-Have. A comparison of several approaches for choosing between working correlation structures

in generalized estimating equation analysis of longitudinal binary data. *Statistics in Medicine*, 28(18) :2338–2355, 2009. 51

R. H. Shumway, R. S. Azari, and M. Kayhanian. Statistical approaches to estimating mean water quality concentrations with detection limits. *Environmental Science & Technology*, 36(15) :3345–3353, 2002. 85

J. A.C. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter. Multiple imputation for missing data in epidemiological and clinical research : Potential and pitfalls. *British Medical Journal*, 338 :b2393, 2009. 82

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 111–147, 1974. 44

M. Stone. Asymptotics for and against cross-validation. *Biometrika*, pages 29–35, 1977. 46

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 56, 57, 58, 63, 64

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67(1) :91–108, 2005. 57

J. Tobin. Estimation of relationships for limited dependent variables. *Econometrica : Journal of the Econometric Society*, pages 24–36, 1958. 86

A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer Science & Business Media, 2007. 73

J-Y. Tzeng, Wenbin Lu, M. W. Farmen, Y. Liu, and P. F. Sullivan. Haplotype-based pharmacogenetic analysis for longitudinal quantitative traits in the presence of drop-out. *Journal of Biopharmaceutical Statistics*, 20(2) :334–350, 2010. 99, 130

S. Van Buuren. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3) :219–242, 2007. 82

S. Van Buuren. *Flexible Imputation of Missing Data*. Chapam & Hall, 2012. 82

S. Van Buuren and K. Groothuis-Oudshoorn. mice : Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3), 2011. 68, 83, 107, 123, 138

S. Van Buuren, J. P.L. Brand, C.G.M. Groothuis-Oudshoorn, and D. B. Rubin. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12) :1049–1064, 2006. 83, 148

S. Van Buuren et al. Multiple imputation of multilevel data. *Handbook of advanced multilevel analysis*, pages 173–196, 2011. 84, 90

W. M. van der Wal, R. B. Geskus, et al. ipw : An r package for inverse probability weighting. *Journal of Statistical Software*, 43(13) :1–23, 2011. 73

G. Verbeke and G. Molenberghs. *Models for Discrete Longitudinal Data*. Springer, 2005. 38

E. Vittinghoff, D. V. Glidden, S. C. Shiboski, and C. E. McCulloch. *Regression Methods in Biostatistics : Linear, Logistic, Survival, and Repeated Measures Models*. Springer Science & Business Media, 2011. 32

H. Wang, R. Li, and C. L. Tsai. A consistent tuning parameter selector for scad. *Biometrika*, 2006. 49

L. Wang, J. Zhou, and A. Qu. Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, 68(2) :353–360, 2012. 59

Y.-G. Wang and V. Carey. Working correlation structure misspecification, estimation and covariate design : Implications for generalised estimating equations performance. *Biometrika*, 90(1) :29–41, 2003. 49, 51, 149

R. W. M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, 61(3) :439–447, 1974. 37, 39

H. White. Maximum likelihood estimation of misspecified models. *Econometrica : Journal of the Econometric Society*, pages 1–25, 1982. 44

A. M. Wood, I. R. White, and P. Royston. How should variable selection be performed with multiply imputed data ? *Statistics in Medicine*, 27(17) :3227–3246, 2008. 99, 100

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 68(1) :49–67, 2006. 57, 58, 105

L. Zeng and J. Xie. Group variable selection via scad-l2. *Statistics*, 48(1) :49–66, 2014. 150

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476) :1418–1429, 2006. 57

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67(2) :301–320, 2005. 57, 59, 150

Liste des communications et publications

Congrès internationaux

J. Geronimi, G. Saporta. The Effect of Missing Visits on GEE, a Simulation Study. In *ASMDA 2015*, pages 257-264, Piraeus, Greece, July 2015. The 16th Conference of the Applied Stochastic Models and Data Analysis International Society.

J. Geronimi, G. Saporta. Variable selection for multiply-imputed data with penalized estimating equations. 28th International Biometric Conference (IBC), 10-5 juillet 2016, Victoria, Canada.

J. Geronimi, G. Saporta. Variable selection for longitudinal biomarkers constrained by a detection limit. 22th International Conference on Computational Statistics (COMPSTAT), 23-26 aout 2016, Oviedo, Espagne.

Congrès nationaux

J. Geronimi, G. Saporta. L'effet de visites manquantes sur l'estimateur des GEE, une étude par simulation. 47èmes Journées de Statistique, 1-5 juin 2015, Lille, France.

J. Geronimi, G. Saporta. L'effet de visites manquantes sur l'estimateur des GEE, une étude par simulation. 6èmes Rencontres Jeunes Statisticiens, 28 aout-2 septembre 2015, Le

Teich, France.

J. Geronimi, G. Saporta. Intégrer les données manquantes dans la sélection de variables pour données longitudinales. 48èmes Journées de Statistique, 30 mai-3 juin 2016, Montpellier, France.

Posters

J. Geronimi, G. Saporta. The Effect of Missing Visits on GEE, a Simulation Study. Miss Data, 18-19 juin 2015, Rennes, France.

Publication en révision

J. Geronimi, G. Saporta. Variable Selection for Multiply-Imputed data with Penalized Generalized Estimating Equations. *Computational Statistics & Data Analysis.*

Publication en cours de rédaction

J. Geronimi, G. Saporta. Covariates Constrained by a Detection Limit : Multiple Imputation by Chained Equations and Tobit Regression.

Annexes

Annexe A

Résultats des simulations de la section 3.3 pour covariables binaires

A.1 Simulations pour données manquantes MCAR

		5% MCAR		10% MCAR	
	PGEE & Comp	PGEE & IS	MIPGEE	PGEE & IS	MIPGEE
1st Pattern					
MSE	0.424	**0.501**	0.548	0.773	**0.677**
SEN	0.865	**0.812**	0.731	**0.760**	0.715
SPE	0.693	0.726	**0.863**	0.783	**0.836**
2nd Pattern					
MSE	0.222	0.325	**0.272**	0.453	**0.315**
SEN	0.798	0.781	**0.807**	0.685	**0.764**
SPE	0.763	0.750	**0.848**	0.738	**0.805**
3rd Pattern					
MSE	0.408	0.531	**0.491**	0.634	**0.551**
SEN	0.778	**0.689**	0.640	**0.666**	0.622
SPE	0.703	0.739	**0.859**	0.786	**0.834**

TABLE A.1 – Mean square error (MSE), sensibilité (SEN) et spécificité (SPE) de la sélection pour des données manquantes MCAR avec PGEE sur données complètes (PGEE & Comp), PGEE sur imputation par la moyenne (PGEE & IS) et MI-PGEE sur 10 imputations avec **covariables binaires**. Les caractères gras représentent les meilleurs résultats entre imputation simple et MI-PGEE

A.2 Simulations pour données manquantes MAR

		5% MAR		10% MAR	
	PGEE & Comp	PGEE & IS	MIPGEE	PGEE & IS	MIPGEE
1st Pattern					
MSE	0.424	0.566	**0.536**	0.780	**0.643**
SEN	0.865	**0.817**	0.756	**0.738**	0.714
SPE	0.693	0.762	**0.864**	0.791	**0.843**
2nd Pattern					
MSE	0.222	0.368	**0.290**	0.448	**0.347**
SEN	0.798	0.745	**0.867**	0.701	**0.802**
SPE	0.763	0.743	**0.851**	0.751	**0.836**
3rd Pattern					
MSE	0.408	0.575	**0.504**	0.654	**0.507**
SEN	0.778	**0.693**	0.634	**0.661**	0.615
SPE	0.703	0.664	**0.859**	0.789	**0.844**

TABLE A.2 – Mean square error (MSE), sensibilité (SEN) et spécificité (SPE) de la sélection pour des données manquantes MAR avec PGEE sur données complètes (PGEE & Comp), PGEE sur imputation par la moyenne (PGEE & IS) et MI-PGEE sur 10 imputations avec **covariables binaires**. Les caractères gras représentent les meilleurs résultats entre imputation simple et MI-PGEE

Annexe B

Résultats détaillés des simulations de la section 3.5.3

B.1 Comparaison des Biais Relatifs Absolus

B.1.1 Pour le 1er scénario de données manquantes

	LOD	LOD/2	LOD/$\sqrt{2}$	Tobit	Tobit-A	MI-Tobit	MI-Tobit-A	Mice-Tobit
0%					0.262			
2%	0.269	0.293	0.285	0.263	0.263	**0.259**	0.263	**0.259**
5%	0.287	0.311	0.304	0.275	0.265	0.269	0.268	**0.265**
10%	0.304	0.325	0.322	0.273	0.264	0.274	0.267	**0.263**
15%	0.331	0.376	0.362	0.269	**0.259**	0.267	0.263	0.262
25%	0.395	0.462	0.435	0.288	**0.256**	0.270	0.264	0.261
35%	0.482	0.556	0.526	0.280	**0.251**	0.289	0.261	0.268
50%	0.708	0.691	0.698	0.296	0.247	0.312	**0.256**	0.274
60%	0.905	0.780	0.833	0.362	**0.255**	0.362	0.261	0.280
75%	1.290	0.864	1.032	0.407	**0.277**	0.383	0.304	0.283
85%	1.826	0.900	1.231	0.493	**0.274**	0.494	0.351	0.298

TABLE B.1 – Biais Relatifs Absolus estimé sur 500 réplications pour les 8 méthodes comparées, pour le 1er scénario de données manquantes

B.1.2 Pour le 2nd scénario de données manquantes

	LOD	LOD/2	LOD/$\sqrt{2}$	Tobit	Tobit-A	MI-Tobit	MI-Tobit-A	Mice-Tobit
0%					0.235			
2%	0.239	0.245	0.243	0.235	0.235	**0.234**	0.235	0.235
5%	0.246	0.265	0.258	0.234	0.235	**0.233**	0.235	0.234
10%	0.262	0.298	0.285	0.234	0.234	**0.232**	0.234	0.234
15%	0.283	0.332	0.314	0.239	0.234	**0.232**	0.233	0.234
25%	0.340	0.404	0.380	0.249	**0.232**	0.238	**0.232**	0.236
35%	0.421	0.480	0.456	0.268	0.237	0.254	0.239	**0.241**
50%	0.606	0.595	0.599	0.308	0.271	0.289	0.280	**0.245**
60%	0.794	0.667	0.720	0.340	0.324	0.322	0.356	**0.269**
75%	1.252	0.783	0.966	0.407	0.541	0.388	0.691	**0.334**
85%	1.842	0.887	1.225	0.481	0.826	0.459	0.849	**0.446**

TABLE B.2 – Biais Relatifs Absolus estimé sur 500 réplications pour les 8 méthodes comparées, pour le 2nd scénario de données manquantes

B.2 Comparaison des estimateurs $\hat{\boldsymbol{\beta}}$

B.2.1 Pour le 1er scénario de données manquantes

	LOD	LOD/2	LOD/$\sqrt{2}$	Tobit	Tobit-A	MI-Tobit	MI-Tobit-A	Mice-Tobit
0%					0,846			
2%	0.842	0.825	0.832	0.847	0.846	0.847	0.846	0.844
5%	0.831	0.793	0.809	0.846	0.846	0.847	0.846	0.844
10%	0.804	0.733	0.762	0.850	0.843	0.848	0.845	0.843
15%	0.763	0.680	0.714	0.834	0.840	0.831	0.839	0.840
25%	0.686	0.575	0.621	0.830	0.850	0.841	0.848	0.831
35%	0.560	0.458	0.501	0.830	0.844	0.817	0.847	0.828
50%	0.294	0.301	0.298	0.809	0.859	0.791	0.860	0.821
60%	0.076	0.206	0.151	0.813	0.872	0.795	0.865	0.794
75%	-0.291	0.111	-0.067	0.813	0.904	0.805	0.860	0.757
85%	-0.613	0.139	-0.164	0.827	0.949	0.794	0.904	0.706

TABLE B.3 – $\hat{\beta}_1$ estimé sur 500 réplications pour les 8 méthodes comparées, pour le 1er scénario de données manquantes. $\beta_1 = 1$

B.2.2 Pour le 2nd scénario de données manquantes

	LOD	LOD/2	LOD/$\sqrt{2}$	Tobit	Tobit-A	MI-Tobit	MI-Tobit-A	Mice-Tobit
0%					1,051			
2%	1.067	1.089	1.088	1.044	1.051	1.042	1.052	1.045
5%	1.092	1.101	1.109	1.037	1.052	1.040	1.050	1.043
10%	1.134	1.179	1.175	1.012	1.048	1.013	1.044	1.043
15%	1.179	1.222	1.219	1.010	1.047	1.009	1.044	1.041
25%	1.265	1.324	1.310	0.966	1.037	0.943	1.025	1.054
35%	1.358	1.406	1.390	0.908	1.022	0.908	1.011	1.063
50%	1.539	1.529	1.534	0.816	0.987	0.802	0.977	1.061
60%	1.694	1.585	1.633	0.743	0.967	0.706	0.961	1.073
75%	1.987	1.626	1.784	0.567	0.906	0.594	0.942	1.083
85%	2.308	1.622	1.900	0.429	0.864	0.450	0.806	1.103

TABLE B.4 – $\hat{\beta}_2$ estimé sur 500 réplications pour les 8 méthodes comparées, pour le 1$^{\text{er}}$ scénario de données manquantes. $\beta_2 = 1$

	LOD	LOD/2	LOD/$\sqrt{2}$	Tobit	Tobit-A	MI-Tobit	MI-Tobit-A	Mice-Tobit
0%					0.194			
2%	0.199	0.215	0.207	0.199	0.196	0.197	0.196	0.192
5%	0.213	0.263	0.240	0.204	0.201	0.200	0.201	0.188
10%	0.230	0.264	0.248	0.218	0.209	0.225	0.212	0.182
15%	0.244	0.283	0.266	0.219	0.215	0.232	0.215	0.179
25%	0.280	0.328	0.307	0.251	0.217	0.253	0.220	0.184
35%	0.332	0.376	0.357	0.280	0.219	0.269	0.224	0.194
50%	0.418	0.410	0.413	0.311	0.240	0.311	0.243	0.206
60%	0.472	0.439	0.452	0.297	0.247	0.330	0.248	0.203
75%	0.592	0.482	0.524	0.281	0.270	0.275	0.248	0.209
85%	0.736	0.496	0.579	0.199	0.255	0.205	0.277	0.205

TABLE B.5 – $\hat{\beta}_3$ estimé sur 500 réplications pour les 8 méthodes comparées, pour le 1$^{\text{er}}$ scénario de données manquantes. $\beta_3 = 0.2$

	LOD	LOD/2	LOD/$\sqrt{2}$	Tobit	Tobit-A	MI-Tobit	MI-Tobit-A	Mice-Tobit
0%					-0.733			
2%	-0.741	-0.754	-0.753	-0.725	-0.732	-0.727	-0.734	-0.725
5%	-0.760	-0.792	-0.785	-0.711	-0.734	-0.722	-0.733	-0.721
10%	-0.792	-0.829	-0.822	-0.707	-0.733	-0.720	-0.733	-0.716
15%	-0.820	-0.865	-0.854	-0.688	-0.738	-0.707	-0.731	-0.716
25%	-0.880	-0.915	-0.907	-0.700	-0.739	-0.692	-0.727	-0.719
35%	-0.946	-0.982	-0.969	-0.675	-0.717	-0.648	-0.721	-0.733
50%	-1.077	-1.069	-1.073	-0.598	-0.704	-0.574	-0.706	-0.742
60%	-1.189	-1.113	-1.147	-0.624	-0.687	-0.522	-0.696	-0.742
75%	-1.401	-1.133	-1.246	-0.400	-0.661	-0.439	-0.713	-0.716
85%	-1.684	-1.157	-1.362	-0.331	-0.632	-0.331	-0.724	-0.733

TABLE B.6 – $\hat{\beta}_4$ estimé sur 500 réplications pour les 8 méthodes comparées, pour le 1$^{\text{er}}$ scénario de données manquantes. $\beta_4 = -0.8$

	LOD	LOD/2	LOD/$\sqrt{2}$	Tobit	Tobit-A	MI-Tobit	MI-Tobit-A	Mice-Tobit
0%					-0.511			
2%	-0.522	-0.533	-0.530	-0.518	-0.516	-0.515	-0.513	-0.506
5%	-0.533	-0.541	-0.541	-0.523	-0.518	-0.517	-0.518	-0.498
10%	-0.545	-0.564	-0.560	-0.527	-0.521	-0.511	-0.520	-0.491
15%	-0.562	-0.596	-0.586	-0.506	-0.511	-0.495	-0.515	-0.488
25%	-0.610	-0.660	-0.641	-0.486	-0.505	-0.491	-0.511	-0.485
35%	-0.667	-0.709	-0.692	-0.475	-0.508	-0.480	-0.503	-0.503
50%	-0.786	-0.779	-0.782	-0.476	-0.509	-0.473	-0.493	-0.502
60%	-0.894	-0.820	-0.850	-0.390	-0.503	-0.441	-0.486	-0.509
75%	-1.159	-0.894	-0.997	-0.429	-0.496	-0.393	-0.449	-0.527
85%	-1.470	-0.940	-1.125	-0.315	-0.484	-0.301	-0.432	-0.505

TABLE B.7 – $\hat{\beta}_5$ estimé sur 500 réplications pour les 8 méthodes comparées, pour le 1$^{\text{er}}$ scénario de données manquantes. $\beta_5 = -0.4$

	LOD	LOD/2	LOD/$\sqrt{2}$	Tobit	Tobit-A	MI-Tobit	MI-Tobit-A	Mice-Tobit
0%					0.592			
2%	0.599	0.599	0.603	0.593	0.593	0.590	0.593	0.590
5%	0.610	0.618	0.621	0.583	0.589	0.585	0.589	0.586
10%	0.632	0.661	0.657	0.570	0.585	0.569	0.584	0.580
15%	0.659	0.696	0.687	0.562	0.584	0.564	0.581	0.580
25%	0.708	0.751	0.738	0.541	0.573	0.543	0.574	0.581
35%	0.774	0.816	0.801	0.514	0.569	0.524	0.570	0.586
50%	0.871	0.869	0.870	0.501	0.560	0.508	0.552	0.589
60%	0.932	0.893	0.912	0.404	0.542	0.433	0.541	0.594
75%	1.007	0.861	0.931	0.322	0.507	0.320	0.517	0.583
85%	1.111	0.760	0.897	0.193	0.471	0.201	0.461	0.577

TABLE B.8 – $\hat{\beta}_6$ estimé sur 500 réplications pour les 8 méthodes comparées, pour le 1$^{\text{er}}$ scénario de données manquantes. $\beta_6 = 0.6$

	LOD	LOD/2	LOD/$\sqrt{2}$	Tobit	Tobit-A	MI-Tobit	MI-Tobit-A	Mice-Tobit
0%					1.020			
2%	1.013	0.995	1.002	1.020	1.019	1.020	1.019	1.020
5%	1.000	0.963	0.977	1.020	1.018	1.020	1.018	1.021
10%	0.974	0.914	0.938	1.020	1.018	1.020	1.018	1.022
15%	0.943	0.867	0.897	1.021	1.020	1.020	1.020	1.024
25%	0.864	0.775	0.811	1.022	1.032	1.022	1.033	1.026
35%	0.759	0.682	0.714	1.024	1.055	1.023	1.054	1.025
50%	0.531	0.545	0.539	1.026	1.103	1.025	1.108	1.017
60%	0.304	0.456	0.394	1.026	1.128	1.025	1.135	1.017
75%	- 0.256	0.320	0.097	1.034	1.158	1.031	1.176	1.031
85%	-1.016	0.194	-0.235	1.033	1.144	1.031	1.141	1.029

TABLE B.9 – $\hat{\beta}_1$ estimé sur 500 réplications pour les 8 méthodes comparées, pour le 2$^{\text{nd}}$ scénario de données manquantes. $\beta_1 = 1$

	LOD	LOD/2	LOD/$\sqrt{2}$	Tobit	Tobit-A	MI-Tobit	MI-Tobit-A	Mice-Tobit
0%					1.000			
2%	1.013	1.014	1.020	0.994	1.001	0.995	1.001	0.998
5%	1.031	1.037	1.046	0.985	1.001	0.985	1.001	0.993
10%	1.061	1.074	1.082	0.966	0.996	0.967	0.996	0.985
15%	1.092	1.113	1.118	0.943	0.985	0.944	0.984	0.975
25%	1.156	1.189	1.185	0.893	0.945	0.894	0.945	0.959
35%	1.227	1.263	1.252	0.833	0.887	0.838	0.885	0.947
50%	1.360	1.352	1.355	0.732	0.757	0.739	0.748	0.928
60%	1.477	1.390	1.429	0.655	0.641	0.661	0.614	0.866
75%	1.737	1.409	1.550	0.518	0.363	0.521	0.219	0.723
85%	2.053	1.406	1.659	0.397	0.059	0.400	0.040	0.544

TABLE B.10 – $\hat{\beta}_2$ estimé sur 500 réplications pour les 8 méthodes comparées, pour le 2$^{\text{nd}}$ scénario de données manquantes. $\beta_2 = 1$

	LOD	LOD/2	LOD/$\sqrt{2}$	Tobit	Tobit-A	MI-Tobit	MI-Tobit-A	Mice-Tobit
0%					0.200			
2%	0.206	0.229	0.218	0.202	0.201	0.202	0.201	0.198
5%	0.215	0.250	0.234	0.209	0.205	0.208	0.204	0.196
10%	0.231	0.278	0.256	0.218	0.210	0.218	0.211	0.191
15%	0.248	0.297	0.275	0.230	0.216	0.230	0.218	0.185
25%	0.284	0.330	0.310	0.251	0.230	0.251	0.231	0.177
35%	0.327	0.360	0.346	0.269	0.239	0.267	0.242	0.173
50%	0.410	0.405	0.407	0.289	0.239	0.286	0.240	0.181
60%	0.484	0.439	0.457	0.286	0.217	0.286	0.212	0.199
75%	0.657	0.506	0.562	0.257	0.125	0.264	0.081	0.239
85%	0.864	0.566	0.668	0.211	0.021	0.221	0.012	0.215

TABLE B.11 – $\hat{\beta}_3$ estimé sur 500 réplications pour les 8 méthodes comparées, pour le 2$^{\text{nd}}$ scénario de données manquantes. $\beta_3 = 0.2$

	LOD	LOD/2	LOD/$\sqrt{2}$	Tobit	Tobit-A	MI-Tobit	MI-Tobit-A	Mice-Tobit
0%					-0.801			
2%	-0.813	-0.818	-0.821	-0.799	-0.803	-0.799	-0.803	-0.800
5%	-0.828	-0.837	-0.841	-0.793	-0.803	-0.793	-0.804	-0.795
10%	-0.852	-0.870	-0.873	-0.780	-0.800	-0.781	-0.801	-0.788
15%	-0.878	-0.904	-0.903	-0.764	-0.794	-0.766	-0.794	-0.780
25%	-0.935	-0.972	-0.963	-0.729	-0.767	-0.733	-0.768	-0.763
35%	-1.000	-1.036	-1.023	-0.693	-0.726	-0.693	-0.725	-0.752
50%	-1.120	-1.113	-1.116	-0.621	-0.629	-0.625	-0.624	-0.739
60%	-1.225	-1.151	-1.183	-0.564	-0.535	-0.567	-0.508	-0.693
75%	-1.466	-1.182	-1.302	-0.452	-0.304	-0.458	-0.179	-0.586
85%	-1.756	-1.197	-1.412	-0.347	-0.056	-0.359	-0.033	-0.434

TABLE B.12 – $\hat{\beta}_4$ estimé sur 500 réplications pour les 8 méthodes comparées, pour le 2$^{\text{nd}}$ scénario de données manquantes. $\beta_4 = -0.8$

	LOD	LOD/2	LOD/$\sqrt{2}$	Tobit	Tobit-A	MI-Tobit	MI-Tobit-A	Mice-Tobit
0%					-0.405			
2%	-0.412	-0.428	-0.422	-0.406	-0.407	-0.405	-0.407	-0.404
5%	-0.424	-0.453	-0.442	-0.407	-0.409	-0.407	-0.409	-0.401
10%	-0.444	-0.482	-0.468	-0.410	-0.413	-0.409	-0.413	-0.395
15%	-0.464	-0.507	-0.491	-0.412	-0.413	-0.412	-0.415	-0.388
25%	-0.508	-0.551	-0.534	-0.416	-0.414	-0.415	-0.414	-0.377
35%	-0.557	-0.590	-0.577	-0.413	-0.407	-0.415	-0.405	-0.370
50%	-0.652	-0.647	-0.649	-0.396	-0.372	-0.401	-0.370	-0.365
60%	-0.739	-0.682	-0.706	-0.376	-0.325	-0.382	-0.314	-0.358
75%	-0.934	-0.741	-0.818	-0.326	-0.188	-0.332	-0.113	-0.348
85%	-1.164	-0.781	-0.920	-0.263	-0.029	-0.269	-0.017	-0.282

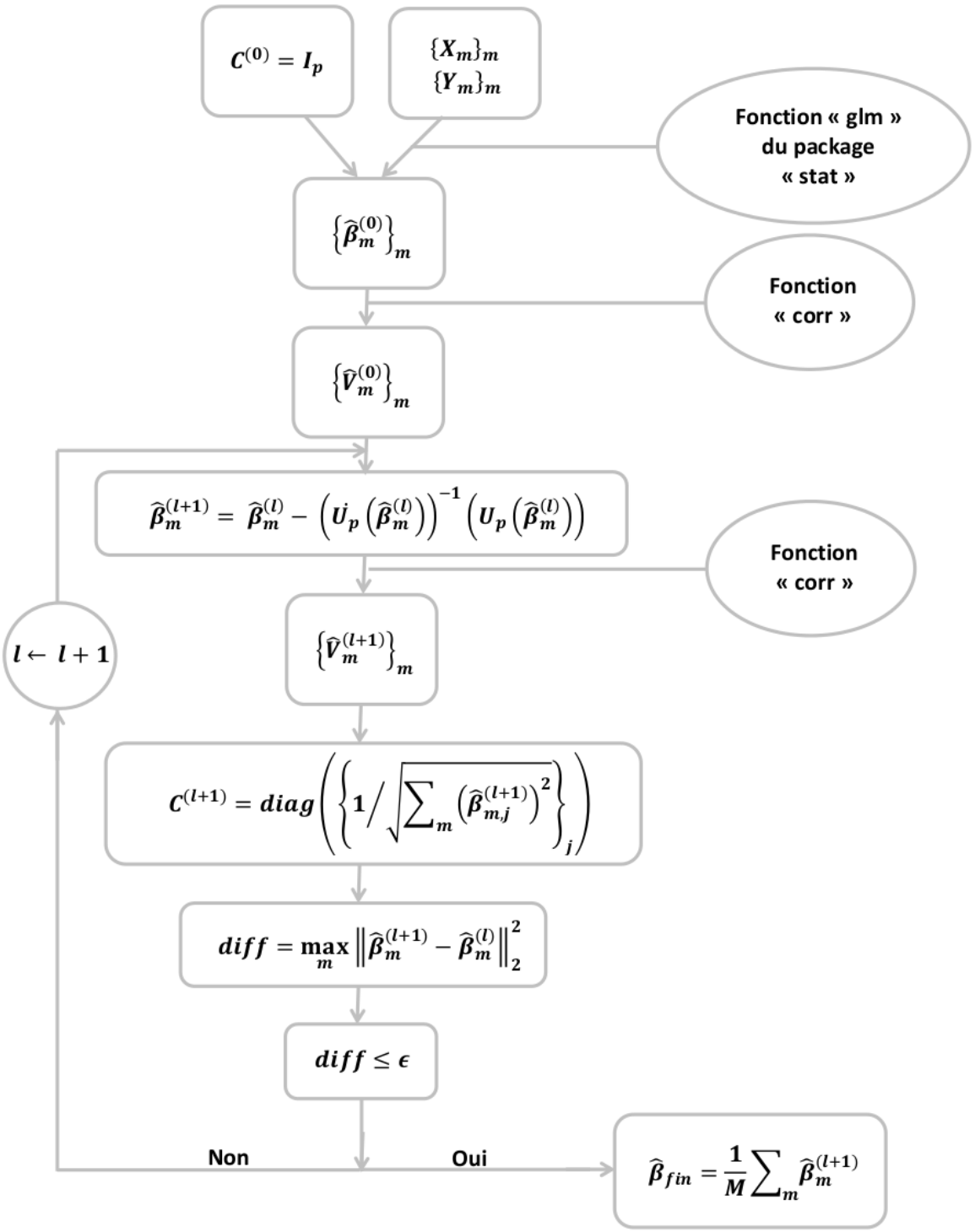
TABLE B.13 – $\hat{\beta}_5$ estimé sur 500 réplications pour les 8 méthodes comparées, pour le 2$^{\text{nd}}$ scénario de données manquantes. $\beta_5 = -0.4$

	LOD	LOD/2	LOD/$\sqrt{2}$	Tobit	Tobit-A	MI-Tobit	MI-Tobit-A	Mice-Tobit
0%					0.608			
2%	0.617	0.622	0.624	0.606	0.609	0.605	0.609	0.606
5%	0.629	0.640	0.642	0.601	0.609	0.602	0.609	0.603
10%	0.650	0.666	0.667	0.592	0.608	0.593	0.608	0.597
15%	0.670	0.689	0.689	0.583	0.603	0.583	0.604	0.591
25%	0.713	0.739	0.734	0.557	0.581	0.556	0.583	0.577
35%	0.760	0.785	0.777	0.520	0.546	0.524	0.546	0.561
50%	0.847	0.843	0.845	0.462	0.464	0.463	0.459	0.536
60%	0.922	0.867	0.892	0.408	0.387	0.412	0.367	0.472
75%	1.086	0.882	0.970	0.318	0.205	0.320	0.122	0.357
85%	1.273	0.880	1.036	0.237	0.031	0.241	0.022	0.245

TABLE B.14 – $\hat{\beta}_6$ estimé sur 500 réplications pour les 8 méthodes comparées, pour le 2$^{\text{nd}}$ scénario de données manquantes. $\beta_6 = 0.6$

Annexe C

Organigramme : calcul de l'estimateur par MI-PGEE


$C^{(0)} = I_p$
$\{X_m\}_m$
$\{Y_m\}_m$
Fonction « glm » du package « stat »
$\left\{\widehat{\beta}_m^{(0)}\right\}_m$
Fonction « corr »
$\left\{\widehat{V}_m^{(0)}\right\}_m$
$\widehat{\beta}_m^{(l+1)} = \widehat{\beta}_m^{(l)} - \left(\dot{U}_p\left(\widehat{\beta}_m^{(l)}\right)\right)^{-1}\left(U_p\left(\widehat{\beta}_m^{(l)}\right)\right)$
Fonction « corr »
$l \leftarrow l + 1$
$\left\{\widehat{V}_m^{(l+1)}\right\}_m$
$C^{(l+1)} = diag\left(\left\{1\Big/\sqrt{\sum_m \left(\widehat{\beta}_{m,j}^{(l+1)}\right)^2}\right\}_j\right)$
$diff = \max_m \left\|\widehat{\beta}_m^{(l+1)} - \widehat{\beta}_m^{(l)}\right\|_2^2$
$diff \leq \epsilon$
Non
Oui
$\widehat{\beta}_{fin} = \frac{1}{M}\sum_m \widehat{\beta}_m^{(l+1)}$

Annexe D

Publications

D.1 Publication parue

J. Geronimi, G. Saporta. The Effect of Missing Visits on GEE, a Simulation Study. In *ASMDA 2015*, pages 257-264, Piraeus, Greece, July 2015. The 16th Conference of the Applied Stochastic Models and Data Analysis International Society.

The Effect of Missing Visits on GEE, a Simulation Study

Julia Geronimi[1,2] and Gilbert Saporta[2]

[1] Institut de de Recherches Internationales SERVIER, 50 rue Carnot 92150 Suresnes (E-mail: geronimi.julia@gmail.com)
[2] Cedric-Cnam, 292 rue Saint Martin 75141 Paris Cedex 03 (E-mail: gilbert.saporta@cnam.fr)

Abstract. Clinical research is often interested in longitudinal follow-up over several visits. All scheduled visits are not carried out and it is not unusual to have a different number of visits by patient. The Generalized Estimating Equations can handle continuous or discrete autocorrelated response. The method allows a different number of visits by patients. The GEE are robust to missing completely at random data, but when the last visits are fewer, the estimator may be biased. We propose a simulation study to investigate the impact of missing visits on the estimators of the model parameters under different missing data patterns. Different types of responses are studied with an exchangeable or autoregressive of order one structure. The number of subjects affected by the missing data and the number of visits removed, vary in order to assess the impact of the missing data. Our simulations show that the estimators obtained by GEE are resistant to a certain rate of missing data. The results are homogeneous regardless to the imposed missing data structure.
Keywords: Longitudinal data, repeated correlated data, correlation, missing data, simulations, Generalized Estimating Equations.

1 Introduction

Clinical follow-up provides information on changing pattern of diseases. This allows for biological measurements and clinical criterion observation over several visits. Therefore, it is possible to study the link between several potential biological covariates and a clinical response on repeated measurements.

However, observations from the same patient cannot be handled as independent and the correlation among visits must be taken into account. Two of the most common methods which are able to deal with longitudinal data are the Generalized Linear Mixed Model, GLMM as describe by McCulloch [6] and the Generalized Estimating Equations, GEE from Liang and Zeger [5].

GLMM are a subject specific method which introduces a random effect per patient to take into account the longitudinal aspect of observations. Unfortunately, the integration over these random effects distribution may be numerically untractable. GEE are a population specific method which consider the

16^{th} *ASMDA Conference Proceedings, 30 June – 4 July 2015, Piraeus, Greece*

intra-subject correlations by imposing a correlation structure to the response. Advantage of the GEE method is that only correct specification of marginal means is needed for having a consistent and asymptotically normal parameter estimator. We will use this method in this paper. For a discussion on GEE, GLMM and relation between marginal and mixed effect models, reader can refer to the work of Park [9], Heagerty and Zeger[3] and Nelder and Lee[7].

Studies'design provides for a number of visits per patient which is regrettably not always complied. In the case of intermittent missing data this results in blank lines in observation matrix. No classical parametric imputation shall be performed since no information is collected at this date. Moreover the interpolation of these values is difficult because there are often few widely spaced visits which means the prediction is blurred.

Missing data, as defined by Rubin [15], are divided into three categories :

- Missing Completely at Random, like a visit randomly deleted by loss record
- Missing At Random, as a missed visit linked to the length of the study
- Missing Not At Random, such as non presence of a patient related to the latent seriousness of his condition

The GEE estimator is robust to the first case but biased in the other two as explained by Liang and Zeger[5] and Robins *et al.*[13]. In case of dropouts Robins *et al.*[13] introduced an inverse probability of censoring weighted GEE which have been studied by Preisser *et al.*[10]. They proposed a modified version of GEE in which observations or person-visits have weights inversely proportional to their probability of being observed, which is unfortunately not suitable here.

Within this context questions may arise :

- How much the GEE estimator is robust to missing visits?
- Which bias should we consider in case of MAR data?

We provide a simulation study to measure the impact of different missing data patterns on GEE estimators. Second part of this paper gives the GEE approach outline. Simulations plan and their results are shown in section 3 and 4. The paper ends by a conclusion in section 5.

2 Generalized Estimating Equation

When the population-average effect is of interest, the marginal model is commonly used to analyzing longitudinal data. Liang and Zeger[5] proposed the Generalized Estimating Equations to estimate the regression parameter, by only specifying the marginal distribution of the outcome variables in the marginal model. Both continuous and binary responses can be modeled.

Let y_{it}, of expectation μ_{it}, be the response of interest for the subject i at the visit t for $i \in \{1, .., K\}$ and $t \in \{1, .., n_i\}$. Each subject has a set of p measured covariates at each time t denoted x_{it}. For a known function $V(.)$ and a given mean-link function $g(.)$ we have :

$$\mathbb{V}ar(y_{it}) = \phi V(\mu_{it}) \tag{1}$$

$$g(\mu_{it}) = x_{it}^t \beta \tag{2}$$

β is the regression parameter to be estimated, ϕ is the dispersion parameter. We will note Y_i, the $n_i \times 1$ independent response vector and X_i, the $n_i \times p$ measured covariates matrix for subject i. Generalized Estimating Equations are defined by :

$$U(\beta) = \sum_{i=1}^{K} D_i^t V_i^{-1} (Y_i - \mu_i) = 0 \tag{3}$$

D_i is the matrix of partial derivatives with $\partial \mu_{it} / \partial \beta_k$ as its (t, k)-th element. V_i is the working covariance matrix defined by :

$$V_i = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2} \tag{4}$$

where $R_i(\alpha)$ is a working correlation matrix completely described by the parameter vector α of size $s \times 1$. A_i is the diagonal matrix with elements equal to the variance terms $V(\mu_{it})$. If $R_i(\alpha)$ is the true correlation matrix of Y_i then V_i is the true covariance matrix.

Liand and Zeger[5] propose an iterative estimation method. A consistent method (as the moments method) is used to estimate the couple (α, ϕ) for fixed values of $\hat{\beta}$. Then equation (3) is used to estimate $\hat{\beta}$ for fixed values of $(\hat{\alpha}, \hat{\phi})$. This leads to a consistent estimate of β.

The choice of $R_i(\alpha)$ is important. Classic structures are independent, exchangeable or auto-regressive of order 1. Selection criterion for the choice of the working correlation matrix are useful. We quote here just a few : the Quasi-log-likelihood under the independence model Information Criteria from Pan [8], the Correlation Information Criteria from Hin and Wang[4] and Rotnitzky-Jewell's criterion[14]. In order to simplify, we will suppose the working correlation known and of exchangeable or auto-regressive of order one structure.

3 Simulations plan/structure

Two types of responses are studied, a continuous and a binary outcome. Both cases introduce 4 covariates which have been simulated by a Gaussian distribution with an auto-regressive of order one with parameter $\rho = 0.3$. We denote Σ this correlation structure.

3.1 Gaussian response

The response Y_i is a multivariate normal vector with intra-subject correlation structure $R_i(\alpha)$ following the model :

$$Y_i = X_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \tag{5}$$

where $x^l \sim \mathcal{N}(0, \Sigma)$ for $l \in \{2, ..., 5\}$. The error vector $\boldsymbol{\epsilon}_i$ is a multivariate normal vector with mean zero and variance matrix $\sigma^2 R_i(\alpha)$. The mean parameter vector is imposed equal to $\beta = (1, 0.5, -0.2, 1, -1)$, where the first component is the intercept. The variance parameter σ^2 is chosen for having a signal/noise ratio of 0.5 as described by Fu[1].

$$\frac{V(x_{it}^t\beta)}{\sigma^2} = \frac{1}{2} \Leftrightarrow \sigma^2 = 2\sum_{l=2}^{5}\beta_l^2 = 4.58 \tag{6}$$

3.2 Binary response

To simulate a binary response, the *logit* link is used and an intra-subject correlation structure equal to $R_i(\alpha)$ is imposed thanks to Qaquish[11].

$$logit(\mathbb{E}(y_{it})) = x_{it}^t\beta \tag{7}$$

where $x^l \sim \mathcal{N}(0, \Sigma)$ for $l \in \{2, ..., 5\}$. The mean parameter vector is imposed equal to $\beta = (1, 0.5, -0.2, 0.3, -0.4)$. The first component is the intercept.

For both kinds of data, the parameters vary as follows according to a full factorial design.

- K, the number of subjects on $\mathcal{K} = \{50, 100, 200, 300\}$
- n, the number of scheduled visits on $\mathcal{N} = \{4, 6, 9\}$
- $R_i(\alpha)$, the correlation structure is either exchangeable or auto-regressive of order one (both admit a scalar $\alpha \to s = 1$)
- α, the unique parameter of correlation on $\mathcal{A} = \{0.1, 0.3, 0.5, 0.6\}$

We simulated 1000 samples that we will called *completed* for each of these 96 scenarios. All of the subjects in these samples get the same number of visits. In order to evaluate the effect of missing visits on the GEE estimators we simulated 1000 other samples that we will called *uncompleted* ou *unbalanced* where we deleted some of the visits on some subjects. The percentage of concerned subjects varies according to $\mathcal{P} = \{10\%, 20\%, 30\%, 50\%\}$ and the number of deleted visits varies according to $\mathcal{V} = \{1, 2, 3\}$.

With the aim of evaluating how robust the GEE estimator is in MCAR and MAR situations, we imposed two different schemes of visits removal. First, we consider a scheme where visits follow a uniform distribution. In that case we can speak of MCAR data. In a second time we consider a probability of

deletion that will increase with the follow-up (i.e. with the number of visits). Last case imposed MAR data. We will talk about uniform unbalanced and increasing unbalanced respectively. All computations are performed using R [12] and GEE fitting performed by the package `geepack` of Halekoh *et al.*[2].

4 Results

A useful criterion for assessing the goodness of an estimator $\hat{\boldsymbol{\theta}}$ is the Absolute Relative Bias defined by $ARB(\hat{\boldsymbol{\theta}}) = \frac{||\mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}||}{||\boldsymbol{\theta}||}$. We estimate this criterion by :

$$\widehat{ARB}(\hat{\boldsymbol{\theta}}) = \frac{1}{1000} \sum_{b=1}^{1000} \frac{||\hat{\boldsymbol{\theta}}_b - \boldsymbol{\theta}||}{||\boldsymbol{\theta}||} \tag{8}$$

where $||.||$ is the euclidean norm which boils down to the absolute value when the parameter is a scalar. $\hat{\boldsymbol{\theta}}_b$ is the estimate of $\boldsymbol{\theta}$ on the b-th sample. The mean of the absolute relative gap between the estimator and its target is thus estimated on 1000 samples.

4.1 Continuous response results

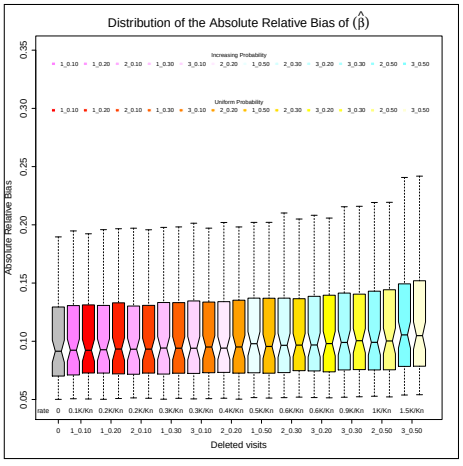


Fig. 1. $\hat{\beta}$ ARB evolution by missing rate for 96 models with a continuous response

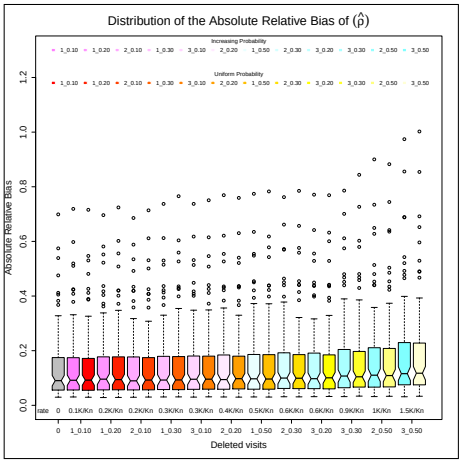


Fig. 2. $\hat{\rho}$ ARB evolution by missing rate for 96 models with a continuous response

Figures 1 and 2 show the distribution of the ARB for the GEE estimator of the parameter β and ρ on the 96 tested models for a continuous response. These graphs compare the two deletion schemes : uniform and increasing.
The boxplots show no differences between the two deletion schemes. Precisely, the difference is between $[-0.005, 0.005]$ for the Absolute Relative Bias of $\hat{\beta}$ and between $[-0.06, 0.06]$ for the ABR of $\hat{\rho}$.

The ARB slightly increases with the missing rate. The median ARB switches from 0.091 to 0.101 for $\hat{\beta}$ and from 0.09 to 0.117 for $\hat{\rho}$.
More precisely, graphics 3, 4 and 5 present the evolution of the Absolute Relative Bias for $\hat{\beta}$ in the case $K = 100$ and $n \in \{4, 6, 9\}$ with increasing unbalanced scheme.

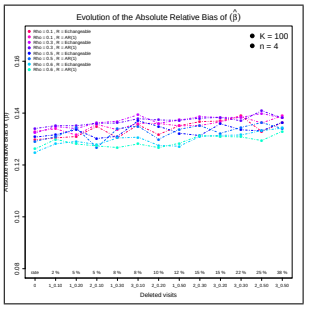


Fig. 3. $\hat{\beta}$ ARB evolution by missing rate for $K = 100$ and $n = 4$ for a continuous response

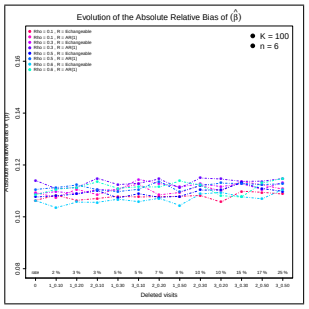


Fig. 4. $\hat{\beta}$ ARB evolution by missing rate for $K = 100$ and $n = 6$ for a continuous response

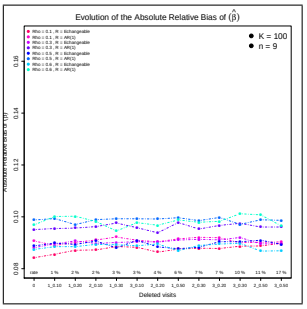


Fig. 5. $\hat{\beta}$ ARB evolution by missing rate for $K = 100$ and $n = 9$ for a continuous response

4.2 Binary response results

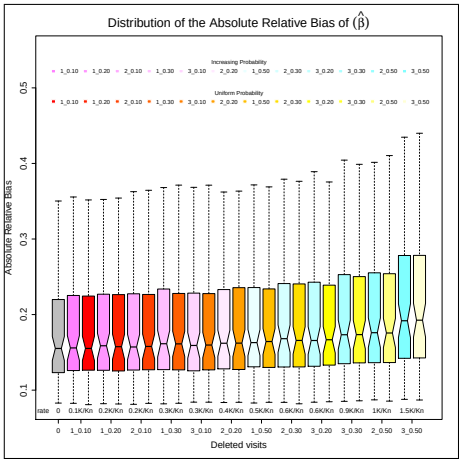


Fig. 6. $\hat{\beta}$ ARB evolution by missing rate for 96 models with a binary response

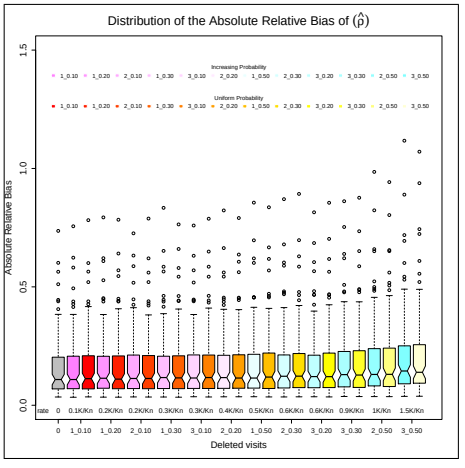


Fig. 7. $\hat{\rho}$ ARB evolution by missing rate for 96 models with a binary response

Graphs 6 and 7 show the distribution of the ARB for the GEE estimator of the parameter β and ρ on the 96 tested models for a binary response. These

graphs compare the two deletion schemes : uniform and increasing.
There are no differences between the two deletion schemes. Some differences in the range of $[-0.005, 0.005]$ and $[-0.015, 0.015]$ have been noted for the Absolute Relative Bias of $\hat{\beta}$ and $\hat{\rho}$ respectively.

The small increase of the ARB is more important for a binary response whith a median ARB switching from 0.155 to 0.193 for $\hat{\beta}$ and from 0.101 to 0.131 for $\hat{\rho}$. Graphs 8, 9 and 10 give more details about the evolution of the Absolute Relative Bias.

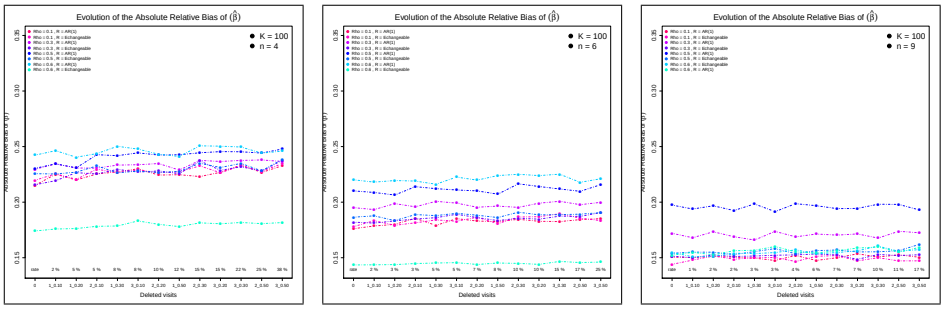


Fig. 8. $\hat{\beta}$ ARB evolution by missing rate for $K = 100$ and $n = 4$ for a binary response

Fig. 9. $\hat{\beta}$ ARB evolution by missing rate for $K = 100$ and $n = 6$ for a binary response

Fig. 10. $\hat{\beta}$ ARB evolution by missing rate for $K = 100$ and $n = 9$ for a binary response

Results on binary response show higher Absolute Relative Bias meaning worst results. Such results were expected since it is more complicated to have an accurate estimator with a binary outcome. Nevertheless both responses, binary and continuous, show the same evolution according to the rate of missing visits. Moreover, both responses point the same lack of differences between uniform unbalanced and increasing unbalanced structure.
Figures 3, 4, 5, 8, 9 and 10 demonstrate how small the increase is with the rate of missing data. The decrease with the number of scheduled visits was expected since it means a lower rate and better estimations.

5 Conclusion

Our simulations show two important issues. First of all, the evolution of the absolute relative bias is similar regardless of the imposed missing data structure. This means that no differences have been highlighted between both schemes. Secondly, the absolute relative bias increases slowly with the missing rate, which means that our imposed missing rate does not disrupt the efficacy of GEE estimator.
We may infer that GEE estimators can be used in studies where MCAR and MAR data are present. Bias induced by MAR is negligible. However, users should pay attention to the missing data scheme and rates used here.

Since it is very complicated to prove the presence of MNAR data, this missing structure has not been studied here. Nevertheless, a complementary study with this type of missing data could bring some more information about expected bias.

References

1. W. Fu. Penalized estimating equations. *Biometrics*, 59:126–132, 2003.
2. U. Halekoh, S. Hojsgaard, and J. Yan. The R package geepack for generalized estimating equations. *Journal of Statistical Software*, 15(2):1–11, 2006.
3. P. Heagerty and S. Zeger. Marginalized multilevel models and likelihood inference (with comments and a rejoinder by the authors). *Statistical Science*, 15(1):1–26, 2000.
4. L.-Y. Hin and Y.-G. Wang. Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine*, 28(4):642–658, 2009.
5. K.-Y. Liang and S. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 38:13–22, 1986.
6. C. McCulloch and J. Neuhaus. *Generalized linear mixed models*. Wiley Online Library, 2001.
7. J. Nelder and Y. Lee. Conditional and marginal models: another view. *Statistical Science*, 19(2):219–238, 2004.
8. W. Pan. Akaike's information criterion in generalized estimating equations. *Biometrics*, 57:120–125, 2001.
9. T. Park. A comparison of the generalized estimating equation approach with the maximum likelihood approach for repeated measurements. *Statistics in Medicine*, 12(18):1723–1732, 1993.
10. J. Preisser, K. Lohman, and P. Rathouz. Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in Medicine*, 21(20):3035–3054, 2002.
11. F. Qaqish. A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika*, 90 (2):455–463, 2003.
12. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`. ISBN 3-900051-07-0.
13. J. Robins, A. Rotnitzky, and L. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.
14. A. Rotnitzky and N. Jewell. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, 77 (3):485–497, 1990.
15. D. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

D.2 Publication en révision

J. Geronimi, G. Saporta. Variable Selection for Multiply-Imputed data with Penalized Generalized Estimating Equations. *Computational Statistics & Data Analysis.*

Variable Selection for Multiply-Imputed Data with Penalized Generalized Estimating Equations

J. Geronimi[a,b,*], G. Saporta[b]

[a] *IRIS, 50 rue Carnot, 92284 Suresnes, France*
[b] *Cedric, CNAM, 292 rue Saint-Martin, 75141 Paris, France*

Abstract

Generalized estimating equations (GEE) are useful tools for marginal regression analysis for longitudinal data. Having a high number of variables along with the presence of missing data presents complex issues when working in a longitudinal context. In variable selection for instance, penalized generalized estimating equations have not been systematically developed to integrate missing data. The MI-PGEE: multiple imputation-penalized generalized estimating equations, an extension of the multiple imputation-least absolute shrinkage and selection operator (MI-LASSO) is presented. MI-PGEE allows integration of missing data and within-subject correlation in variable selection procedures. Missing data are dealt with using multiple imputation, and variable selection is performed using a group LASSO penalty. Estimated coefficients for the same variable across multiply-imputed datasets are considered as a group while applying penalized generalized estimating equations, leading to a unique model across multiply-imputed datasets. In order to select the tuning parameter, a new BIC-like criterion is proposed. In a simulation study, the advantage of using MI-PGEE compared to simple imputation PGEE are shown. The usefulness of the new method is illustrated by an application to a subgroup of the placebo arm of the strontium ranelate efficacy in knee osteoarthritis trial study.

Keywords: Generalized Estimating Equations; LASSO; Longitudinal data;

*Corresponding author
Email addresses: geronimi.julia@gmail.com (J. Geronimi), gilbert.saporta@cnam.fr (G. Saporta)

Missing data; Multiple imputation; Variable selection.

1. Introduction

Longitudinal data from clinical studies gives us the opportunity to study the link between a clinical criterion of interest and covariates, through data collected repeatedly from each study subject over time. The Generalized Estimating Equations (GEE), introduced by Liang and Zeger (1986) in a seminal paper, are a very popular marginal method for regression. Avoiding explicit specification of a likelihood, the GEE provide an extension of general linear models which are robust for inference. The method is easy to implement and only requires specification of the marginal mean function and a working correlation structure for the within-subject correlation.

However, there is difficulty when it comes to model selection; most traditional model selection criteria, such as AIC, BIC and Mallow's C_p, need to be newly defined to take into account the clustered structure of observations and lack of explicit likelihood. Pan (2001) has proposed the quasi-likelihood under independence model Criterion (QIC) as an extension of AIC to GEE model selection. Cantoni et al. (2005) have proposed the Generalized C_p (GC_p) for robust GEE model selection. A more detailed list, with comparisons by simulation, can be found in Dziak and Li (2006).

When the number of variables is large, penalized regression has been extended to GEE through the penalized generalized estimating equations (PGEE). Fu (2003) has proposed a generalization of the bridge penalty to GEE, Dziak and Li (2006) generalized LASSO and smoothly clipped absolute deviation (SCAD) penalties, Blommaert et al. (2014) have suggested a combination of penalties such as in the elastic net, composed of a LASSO and ridge penalty for longitudinal Gaussian data, and Wang et al. (2012) have proposed the SCAD-penalized GEE in which the number of parameters is allowed to diverge. Li et al. (2013) have proposed the smooth-threshold GEE (SGEE) as an extension of the work of Ueki (2009).

A major limitation for existing methods is that they cannot accommodate missing data. The complete case method, which deletes every row with missing values before analysis, is wasteful and may induce bias (Little, 1992). Moreover, the data that motivates our work suffers from missing values scattered haphazardly throughout the data.

Multiple Imputation (MI) is a popular and useful tool to address the problem of missing data (Rubin, 1996; Schafer, 1999; Schafer and Graham, 2002). The key idea of MI is to first replace each missing observation by a set of plausible values, as predicted by a chosen imputation method's model. Each slightly different imputed dataset is then analyzed by the method of our choice, and then results from all of them are combined to produce final parameter estimates, using Rubin's rules (Rubin, 1987). Consequently, inference reflects the uncertainty in the missing-data imputation. Good statistical inference can be obtained with only three to ten imputations in many cases, which makes the method fast and efficient (Rubin, 2004). Joint modeling and sequential regression imputation are two well-known methods that can handle missing data with a general missing pattern (Schafer, 1997; Van Buuren et al., 2006). Unfortunately, jointly modeling multivariate data can be challenging with large datasets and several types of variables (e.g., continuous, categorical, etc.) (Van Buuren, 2007; Molenberghs et al., 2014). Sequential regression imputation uses separate conditional models for each variable, given the others. These models, also known as multivariate imputation by chained equation (mice), are flexible, and implemented in the `R` package `mice` from Van Buuren and Groothuis-Oudshoorn (2011).

Little is known about how to perform effective and reliable model selection with multiply-imputed datasets. Shen and Chen (2013) suggest MI quasi-likelihood under the independence model information criterion (MI-QIC) and MI missing longitudinal information criterion(MI-MLIC) for model selection in MI-GEE analysis, which may become computationally intensive when the number of covariates is large. For univariate outcome regression, Rubin's rules stepwise procedure was proposed by Wood et al. (2008). Multiple imputation LASSO (MI-LASSO), introduced by Chen and Wang (2013) for independent

data, proposes a shared selection across the multiply-imputed datasets which has been demonstrated to be better, in simulations, than Rubin's rules stepwise. However, none of these methods are able to take into account within-subject correlation. The aim of this paper is therefore to extend MI-LASSO to accommodate longitudinal data.

Given that missing data arises often in longitudinal studies, and that variable selection is important in many applications, we aim to propose an effective procedure for MI-GEE variable selection. For this purpose, we propose MI-PGEE: multiple imputation-penalized generalized estimating equations, an extension of MI-LASSO to correlated data. The paper is organized as follows. Section 2 briefly reviews the basics of GEE, MI-GEE and mice algorithm, while MI-PGEE and algorithmic issues are detailed in Section 3. Simulations and results are presented in Section 4, and an application to the SEKOIA study is given in Section 5. Concluding remarks are provided in Section 6.

2. Multiple Imputation Generalized Estimating Equations

2.1. Generalized estimating equations

Let $Y_i = (Y_{i1}, \ldots, Y_{iT_i})$ and $X_i = (X_{i1}, \ldots, X_{iT_i})^T$ be the data collected for individual $i \in \{1, \ldots, K\}$. Y_{it} and X_{it} are the observed response and vector of p covariates at time $t \in \{1, \ldots, T_i\}$. In GEE, the marginal mean $\mu_{it} = \mathbf{E}(Y_{it}|X_{it})$ is usually related to the covariates through the link function g, such that $g(\mu_{it}) = X_{it}^T\beta$, where β is a $p \times 1$ vector of regression parameters. The marginal variance depends on the marginal mean according to $Var(Y_{it}) = \phi\nu(\mu_{it})$, where ϕ is the dispersion parameter and ν a variance function that defines the mean-variance relationship. Estimates given by the GEE of Liang and Zeger (1986) are solutions of

$$S(\beta) = \sum_{i=1}^{K} S_i(\beta) = \sum_{i=1}^{K} D_i^T V_i^{-1}(Y_i - \mu_i) = 0, \tag{1}$$

where $D_i = \partial\mu_i/\partial\beta^T$, and $V_i = A_i^{1/2} R_i(\alpha) A^{1/2}$ with A_i a $T_i \times T_i$ diagonal matrix composed of marginal variances $Var(Y_{it})$. $R_i(\alpha)$ is the working correlation

matrix chosen by the user which depends on a set of parameters α. Liang and Zeger (1986) suggest computing the solution by iterating between a modified Newton algorithm for β and a moment estimation for the nuisance parameters α and ϕ, until convergence.

The covariance matrix of the estimate $\hat{\beta}$, solution of (1), can be approximated by the so-called sandwich estimator, robust to misspecification of the variance-covariance model:

$$W = I^{-1} \left\{ \sum_{i=1}^{K} D_i^T V_i^{-1} Cov(Y_i) V_i^{-1} D_i \right\} I^{-1}. \tag{2}$$

Here, $I = \sum_{i=1}^{K} D_i^T V_i^{-1} D_i$, and β, α and ϕ can be replaced by their estimates. The covariance matrix $Cov(Y_i)$ can be estimated by $(Y_i - \mu_i)(Y_i - \mu_i)^T$.

2.2. Multiple imputation generalized estimating equations

The key idea of multiple imputation is to use a chosen random distribution to replace each missing observation by a set of M plausible values, thus obtaining M datasets. The combined estimate $\overline{\beta}$ is then given by the average of the M individual estimates from the imputed datasets, thanks to the well-known Rubin's rules (Rubin, 1987):

$$\overline{\beta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\beta}_m. \tag{3}$$

The estimator $\hat{\beta}_m$ is the estimate of β in the mth imputed dataset, $m \in \{1, \ldots, M\}$. The variance of $\overline{\beta}$ is composed of a within-imputation variance $\overline{W} = \frac{1}{M} \sum_{m=1}^{M} \hat{W}_m$, where $\hat{W}_m$ is the covariance matrix of $\hat{\beta}_m$ estimated as in (2), and a between-imputation variance $B = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\beta}_m - \overline{\beta})(\hat{\beta}_m - \overline{\beta})^T$. The combined variance is then :

$$\Sigma = \overline{W} + \frac{M+1}{M} B. \tag{4}$$

As remarked in Chen and Wang (2013), Rubin's rules can only be applied if the same covariates are retained in each imputed dataset. This idea motivates the need to have a selection procedure that provides a joint model across all

imputed datasets, to be able to apply Rubin's rules after an optimal model has been selected.

2.3. Multivariate imputation by chained equations

When confronted with a large number of covariates – which may be of different types – joint modeling can be challenging. Multivariate imputation by chained equations (MICE), with the associated `R` package `mice`, has been used to respect the heterogeneity of databases (Van Buuren and Groothuis-Oudshoorn, 2011). Here, each variable X_j, composed of the observed X_j^o and missing part X_j^m, is first imputed thanks to its marginal distribution, then at each iteration l of the algorithm, missing values of the variable are imputed with a univariate conditional model given the others, leading to the following:

$$\begin{aligned}
\theta_1^{*(l)} &\sim \mathbb{P}(\theta_1 | X_1^o, X_2^{(l-1)}, \ldots, X_p^{(l-1)}),\\
X_1^{*(l)} &\sim \mathbb{P}(X_1 | X_1^o, X_2^{(l-1)}, \ldots, X_p^{(l-1)}, \theta_1^{*(l)}),\\
&\vdots\\
\theta_p^{*(l)} &\sim \mathbb{P}(\theta_p | X_p^o, X_1^{(l)}, \ldots, X_{p-1}^{(l)}),\\
X_p^{*(l)} &\sim \mathbb{P}(X_p | X_p^o, X_1^{(l)}, \ldots, X_{p-1}^{(l)}, \theta_p^{*(l)}).
\end{aligned}$$

Here, ${X_j}^{(l)} = (X_j^o,\ X_j^{*(l)})$ is the jth imputed variable at iteration l, and the parameters $\theta_1, \ldots, \theta_p$ are specific to the respective conditional densities, using notation from Van Buuren and Groothuis-Oudshoorn (2011). This algorithm use a concatenation of univariate procedures, making it easy to implement. In this way, we can use different imputation procedure for different variable types. This method is well-known and widely used (Horton and Lipsitz, 2001; Van Buuren et al., 2006; White et al., 2011).

Many situations only require 3 to 10 imputations to yield good results (Rubin, 1987), but in order to achieve better estimates of standard errors and better confidence intervals, more imputations are usually performed. The rule of thumb suggested by Bodner (2008) is to use the average percentage of missingness as the number of imputations. If there is 10% of missing data, on average, in a

dataset, 10 imputations are performed. This idea is discussed further by White et al. (2011).

3. Multiple imputation penalized generalized estimating equations

The new method we present (MI-PGEE) integrates intra-subject correlation and missing data into the variable selection framework. This requires defining a new operator to help extend MI-LASSO to longitudinal data.

3.1. Multiple imputation LASSO

For cross-sectional studies, Chen and Wang (2013) have proposed the multiple imputation LASSO (MI-LASSO) to handle missing data when performing variable selection. Consider a cross-sectional study with K subjects, where Y and X are respectively the response and matrix of p covariates. In the presence of missing data, filled-in by multiple imputation, we will denote Y_m and $X_m = (X_{m,1}, \ldots, X_{m,p})$ the response and covariates imputed on the mth sample. Chen and Wang (2013) use a group LASSO penalty on the imputed data sets, where the estimated regression coefficients of the same variable across all the imputed datasets are treated as a group. They consider the following optimization approach:

$$\min_{\beta_{m,j}} \left\{ \sum_{m=1}^{M} ||Y_m - X_m^T \beta_m||_2^2 + \lambda \sum_{j=1}^{p} \sqrt{\sum_{m=1}^{M} \beta_{m,j}^2} \right\}. \tag{5}$$

The quantity defined by $\sum_{j=1}^{p} \sqrt{\sum_{m=1}^{M} \beta_{m,j}^2}$ is the group LASSO penalty, introduced by Yuan and Lin (2006). Because of a singularity at the origin, a local quadratic approximation (LQA) as in Fan and Li (2001) is used. At the lth iteration of the algorithm, we already have estimates $\hat{\beta}_{m,j}^{(l)}$, $m \in \{1, \ldots, M\}$. When $\sqrt{\sum_{m=1}^{M} \beta_{m,j}^2} > 0$, the LQA can use the following approximation:

$$\sqrt{\sum_{m=1}^{M} \beta_{m,j}^2} \approx \frac{\sum_{m=1}^{M} \beta_{m,j}^2}{\sqrt{\sum_{m=1}^{M} (\beta_{m,j}^{(l)})^2}}.$$

Therefore, (5) can be approximated by a sum of M ridge regressions:

$$\min_{\beta_{m,j}} \sum_{m=1}^{M} \left\{ ||Y_m - X_m^T \beta_m||_2^2 + \lambda \sum_{j=1}^{p} c_j \beta_{m,j}^2 \right\}. \tag{6}$$

Weights are given by $c_j = 1/\sqrt{\sum_{m=1}^{M} (\beta_{m,j}^{(l)})^2}$, and λ is chosen by minimizing a BIC-like criterion.

3.2. Multiple imputation penalized generalized estimating equations

The penalized GEE (PGEE) method, developed by Fu (2003), involves solutions to

$$S^p(\beta) = S(\beta) - \dot{P}(\beta) = 0.$$

The vector $\dot{P}(\beta) = \partial P(\beta)/\partial\beta$ is the derivative of the penalty function, which could be the bridge penalty as in Fu (2003), or a combination of known penalties, as in Blommaert et al. (2014). Explicit specification of the joint likelihood is avoiding by the use of GEE, which turns the optimisation problem of (5) into a system of equations. Therefore, multiple imputation PGEE (or MI-PGEE) as a modification of (6) to PGEE, may be approximated by M penalized GEE with ridge penalties and adaptive weights:

$$S^p(\beta_m) = \sum_{i=1}^{K} D_{m,i}^T V_{m,i}^{-1} (Y_{m,i} - \mu_{m,i}) - \lambda \mathbf{C} \beta_m = 0, \tag{7}$$

for $m \in \{1, ..., M\}$. The matrix $\mathbf{C} = diag(c_j)$ is the only shared quantity across the imputed datasets, as in (6). Solutions of the PGEE can be approximated by the Newton-Raphson algorithm, as suggested in Dziak and Li (2006). Our method allows the user to fit M models on all imputed datasets jointly, which means we obtain one overall selection across the imputed datasets that integrates within-subject correlation.

3.3. Selection of tuning parameters

Penalized regression traditionally uses cross validation for tuning parameter selection (Tibshirani, 1996). A more general form is given by Fu (2005), in which

generalized cross validation is extended to GEE. This method is computational intensive in the multiple imputation case. The major problems are: how to define the degree of freedom, how many subjects to use, and how to integrate the working correlation matrix. The BIC-like criterion BIC_G we propose is given by:

$$BIC_G = MN \log(\sum_{m=1}^{M} WRSS_m/MN) + df \log(\sum_{m=1}^{M} \tilde{N}_m). \quad (8)$$

The weighted residual sum of squares (WRSS) estimated on the mth imputed dataset, $WRSS_m$, is given by

$$WRSS_m = \sum_{i=1}^{K} (Y_{m,i} - \hat{\mu}_{m,i})^T \hat{R}_{m,i}^{-1}(\alpha)(Y_{m,i} - \hat{\mu}_{m,i}).$$

If the response is not continuous and the link function not the identity, the first part of the criterion could be replaced by the weighted deviance, as defined by Fu (2005), and equation (8) adjusted accordingly .

$\hat{R}_{m,i}$ is the estimated working matrix of correlations for the mth imputed dataset. As explained in Dziak and Li (2006), the fitting criterion is now affected by the determinant term, as it will give a different correlation term for each subset considered. One way for dealing with this is to estimate $R_i(\alpha)$ once, from the largest model (i.e., when $\lambda = 0$).

The degree of freedom, noted *df*, is estimated as for the group LASSO in Yuan and Lin (2006) and Chen and Wang (2013):

$$df = \sum_{j=1}^{p} I\left(\sqrt{\sum_{m=1}^{M} \hat{\beta}_{m,j}^2} > 0\right) + \sum_{j=1}^{p} \frac{\sqrt{\sum_{m=1}^{M} \hat{\beta}_{m,j}^2}}{\sqrt{\sum_{m=1}^{M} \tilde{\beta}_{m,j}^2}}(M-1).$$

Here, $\tilde{\beta}_{m,j}$ is the jth estimated coefficient on the mth imputed dataset for the complete model (i.e., GEE with no penalization).

A *light* BIC or a *heavy* BIC, using respectively the number of subjects K or the number of observations $N = \sum_{i=1}^{K} T_i$, could be used, as described in Dziak and Li (2006). A compromise, suggested by Fu (2005), defines an *effective*

sample size which accounts for the amount of within-subject correlation:

$$\tilde{N} = \sum_{i=1}^{K} \frac{T_i^2}{\sum_{i=1}^{K}\sum_{k=1}^{K} \hat{R}_{ik}}.$$

Thus $K \leq \tilde{N} \leq N$, and $\tilde{N}$ decreases as observations become more correlated. We estimate this quantity on each imputed dataset, leading to M effective sample sizes $\tilde{N}_m$, for $m \in \{1, \ldots, M\}$.

The tuning parameter λ is now chosen by minimizing our BIC-like criterion over a grid of possible values. Each value of the grid gives an estimate of β that we can use to trace profiles of estimates, depending on the tuning parameter.

3.4. Computational issues

The algorithm starts with initial values $\beta_m^0 = (\beta_{m,1}^0, \ldots, \beta_{m,p}^0)$ close to the solution estimated by GEE without penalization. Then at each iteration l, equation (7) can be estimated by a Taylor series expansion, leading to the following iterative algorithm :

$$\beta_m^{(l+1)} = \beta_m^{(l)} - \left(\frac{\partial S^p(\beta_m^{(l)})}{\partial \beta} \right)^{-1} S^p(\beta_m^{(l)}),$$

$$\beta_m^{(l+1)} = \beta_m^{(l)} - \left[\frac{\partial S(\beta_m^{(l)})}{\partial \beta} - \lambda \mathbf{C}^{(l)} \right]^{-1} \left[S(\beta_m^{(l)}) - \lambda \mathbf{C}^{(l)} \beta_m^{(l)} \right].$$

Once the group of coefficient associated with the jth covariate is shrunken to zero, the process can not reverse itself, as c_j will be large. To overcome this limit, we fix $\hat{\beta}_{m,j}^{(l)} = \delta$ for $m \in \{1, \ldots, M\}$ when $\sum_{m=1}^{M} \left(\hat{\beta}_{m,j}^{(l)} \right)^2 \leq M\delta^2$, and choose $\delta = 10^{-10}$ following the suggestion of Chen and Wang (2013). Because $S(\beta_m^{(l)})$, as defined in Equation (1), and its derivative depend on the working covariance matrix, we iterate between this modified Newton algorithm, to estimate $\hat{\beta}_m^{(l)}$, and the method of moments to estimate the nuisance parameters, until convergence.

At the end of iterations, small coefficients are set to zero. We fix $\hat{\beta}_{m,j} = 0$ for $m \in \{1, \ldots, M\}$ when $\sum_{m=1}^{M} \hat{\beta}_{m,j}^2 \leq 5^{-10}$. This allows the ridge penalty to shrink and select coefficients as λ grows. For a given tuning parameter λ, this process is repeated until convergence.

4. Simulations

Datasets were simulated with $K = 100$ individuals and 4 time points. Responses Y_{it} were generated as $Y_{it} = 1 + X_{it}^T\beta + \epsilon_{it}$, with ϵ_{it} a multivariate normal with an AR(1) correlation structure given by parameter $\rho_y = 0.7$ for within-subject correlation, and variance σ_y^2. X is the matrix of $p = 30$ covariates, of which 5 are active. The true regression coefficient is $\beta = (1, 0.5, -0.2, 1, -1, 0, \ldots, 0)$, and σ_y^2 was chosen to get a signal-to-noise ratio of one. Three patterns of correlation were used for covariates. The first uses independent covariates, which means $cor(X_l, X_k) = 0, \forall\ l \neq k$, and the second used a positive uniform correlation, leading to $cor(X_l, X_k) = 0.5, \forall\ l \neq k$. The third pattern involves strong correlation among actives variables, with $cor(X_1, X_3) = 0.9$, $cor(X_2, X_5) = -0.8$, $X_4 = X_1 - X_2 + e$, where $e \sim U(0, 0.005)$, and $cor(X_l, X_k) = 0.1, \forall\ l \neq k \in \{6, \ldots, 30\}$. This is similar to the simulations performed by Fu (2003). Continuous covariates are simulated with a multivariate Gaussian with zero mean and unit variance. Covariates that are repeated across visits usually exhibit a non-zero correlation across time points; to ensure plausible simulations, we imposed an exchangeable structure with $\rho_x = 0.3$.

Missing data was generated in the covariates and in the response, with missing completely at random (MCAR) and missing at random (MAR) mechanisms. Ten imputed datasets were generated using predictive mean matching for continuous covariates and logistic regression for binary ones, using the `R` package `mice` (Van Buuren and Groothuis-Oudshoorn, 2011). For each incomplete covariate, the imputation model is composed of all the other covariates, the response and the time variable. This is called multiple imputation separate classes, or per-group imputation.

We compared: the performance of PGEE with complete data before generating missingness, PGEE with simple imputation by the mean, and our method MI-PGEE, with an autoregressive working correlation matrix, the true structure of the response. The regularization parameter λ was chosen by minimizing a BIC-type criterion for the three methods. We conducted an initial step with

fifteen replications to determine the small interval, by increments, of 0.05 for the grid of values for λ.

Four criteria are used. The first two are the mean squared error $MSE(\hat{\beta}) = (\hat{\beta}-\beta)^T X^T X(\hat{\beta}-\beta)$ and the relative MSE, defined by $RMSE = MSE(\hat{\beta})/MSE(\tilde{\beta})$, which allows us to interpret the result as the percentage of the error that would have been made with unpenalized GEE. The other two criteria are the sensitivity and specificity of selection, which represent the rate of selected active covariates, and the rate of not-selected unimportant covariates. Lower MSE and RMSE as well as higher sensitivity and specificity reflect a better final selection.

Simulations were performed using `R` R Development Core Team (2008) with the help of packages `geepack` from Halekoh et al. (2006) and `mice` from Van Buuren and Groothuis-Oudshoorn (2011).

4.1. Simulation 1. Missing Completely at Random

MCAR was imposed by independently dropping 5% or 10% of the data, which can affect the response and covariates, leading to about only 60% complete cases. Table 1 presents results for the complete data base with no missing data, simple imputation by the mean, and MI-PGEE, averaged over 200 simulated datasets. It shows that MI-PGEE has a higher specificity than PGEE for all scenarios, and smaller errors as soon as covariates are correlated. The sensitivity is similar for the first correlation pattern, and higher for the second.

4.2. Simulation 2. Missing at Random

Missing data indicators R_{ijt} for subject i, covariate j at time t were generated using the following logistic regression model:

$$logit\left[Pr(R_{ijt}=0|X_{ij(t-1)},Y_{i(t-1)})\right] = \alpha_0 + \frac{1}{2}X_{ij(t-1)} + \frac{1}{2}Y_{i(t-1)}.$$

The intercept α_0 was chosen to generate around 60% complete cases. Missing data in responses were generated using $logit\left[Pr(R_{it}=0|Y_{i(t-1)})\right] = \alpha_0 + \frac{1}{2}Y_{i(t-1)}$. We obtained a missing data scheme depending on previous observations and responses for each variable. Results are shown in Table 2. These are

Table 1: Mean squared error (MSE), relative MSE (RME), sensitivity (SEN) and specificity (SPE) with continuous covariates and Missing Completely at Random data. Comp is for PGEE with complete data before MCAR is imposed and SI is for PGEE with simple imputation. Best results between SI-PGEE and MI-PGEE are in bold.

		5% MCAR		10% MCAR	
	Comp	SI	MI-PGEE	SI	MI-PGEE
1st Pattern					
MSE	0.440	**0.610**	0.633	0.774	**0.756**
RME	0.325	**0.427**	0.569	0.529	**0.484**
SEN	0.922	**0.895**	0.852	**0.874**	0.851
SPE	0.632	0.686	**0.834**	0.704	**0.784**
2nd Pattern					
MSE	0.587	0.612	**0.252**	0.718	**0.339**
RME	0.893	0.930	**0.486**	0.868	**0.447**
SEN	0.790	0.720	**0.839**	0.723	**0.827**
SPE	0.686	0.719	**0.795**	0.757	**0.766**
3rd Pattern					
MSE	0.890	0.987	**0.495**	1.097	**0.590**
RME	0.340	0.409	**0.343**	0.413	**0.304**
SEN	0.726	**0.677**	0.609	**0.670**	0.610
SPE	0.643	0.697	**0.824**	0.754	**0.798**

Table 2: Mean squared error (MSE), relative MSE (RME), sensitivity (SEN) and specificity (SPE) with continuous covariates and Missing at Random data. Comp is for PGEE with complete data before MAR is imposed and IS is for PGEE with simple imputation. Best results between SI-PGEE and MI-PGEE are in bold.

		5% MAR		10 % MAR	
	Comp	SI	MI-PGEE	SI	MI-PGEE
1st Pattern					
MSE	0.412	0.592	**0.580**	**0.687**	0.700
RME	0.331	**0.402**	0.480	0.456	**0.445**
SEN	0.919	**0.882**	0.852	**0.854**	0.840
SPE	0.635	0.735	**0.825**	0.756	0.795
2nd Pattern					
MSE	0.569	0.572	**0.235**	0.608	**0.289**
RME	0.921	0.858	**0.447**	0.798	**0.408**
SEN	0.801	0.705	**0.842**	0.709	**0.842**
SPE	0.686	0.689	**0.809**	0.701	**0.786**
3rd Pattern					
MSE	0.890	1.180	**0.480**	1.062	**0.599**
RME	0.340	0.378	**0.301**	0.376	**0.297**
SEN	0.726	**0.678**	0.622	**0.645**	0.621
SPE	0.643	0.641	**0.816**	0.626	**0.783**

similar to the MCAR case, with higher specificity for MI-PGEE across correlation patterns, and smaller errors with correlated covariates.

4.3. Simulation 3. Binary covariates

Here, the same structures as in Section 4.1 and 4.2 have been used for binary covariates using multivariate Gaussians and zero as a cut-off. We obtained similar results as for continuous candidate covariates, so only the MAR results are shown in Table 3. Results for MCAR structure can be found in Appendix B.

Table 3: Mean squared error (MSE), relative MSE (RME), sensitivity (SEN) and specificity (SPE) with binary covariates and Missing at Random data. Comp is for PGEE with complete data before MAR is imposed and IS is for PGEE with simple imputation. Bold characters are the best results between SI-PGEE and MI-PGEE.

		5%MAR		10%MAR	
	Comp	SI	MI-PGEE	SI	MI-PGEE
1st Pattern					
MSE	0.424	0.566	**0.536**	0.780	**0.643**
RME	0.305	**0.380**	0.489	0.441	**0.420**
SEN	0.865	**0.817**	0.756	**0.738**	0.714
SPE	0.693	0.762	**0.864**	0.791	**0.843**
2nd Pattern					
MSE	0.222	0.368	**0.290**	0.448	**0.347**
RME	0.232	**0.302**	0.460	**0.372**	0.413
SEN	0.798	0.745	**0.867**	0.701	**0.802**
SPE	0.763	0.743	**0.851**	0.751	**0.836**
3rd Pattern					
MSE	0.408	0.575	**0.504**	0.654	**0.507**
RME	0.299	**0.355**	0.451	**0.350**	0.358
SEN	0.778	**0.693**	0.634	**0.661**	0.615
SPE	0.703	0.664	**0.859**	0.789	**0.844**

4.4. Overall analysis of simulations

Across all our simulations, MI-PGEE had the highest specificity, which means that the method is better able to delete truly unimportant covariates. For the third pattern, the sensitivity is small which is mostly due to the non-selection of the covariate X_3, which has the smallest valued coefficient. We remark that this third pattern affects all methods; high correlation among active covariates leads to non-selection of important covariates. For the three patterns, an increase in missing data had little impact on sensitivity, but increased the error and impacted the specificity. When comparing simulations 1 and 2, it appears that MI-PGEE is stable while PGEE with simple imputation has poorer results. Similar conclusions can be made for both continuous and binary covariates, with higher specificity in the latter case.

Better results are observed with SI for the first pattern in terms of some of the criteria considered. This pattern is the only design where covariates are independent. In this context, multiple imputation uses uninformative predictors for the imputation model, providing worse selection in terms of certain criteria. Conversely, when covariates are correlated, the multiple imputation procedure uses predictors which carry information. MI-PGEE then provides better variable selection than PGEE with single imputation.

5. Application to SEKOIA data

This data comes from the strontium ranelate efficacy in knee osteoarthritis trial (SEKOIA), whose design and major results can be found in Reginster et al. (2012) and Cooper et al. (2012). In this application, we used the data of 166 patients, with 4 scheduled visits, from the placebo arm and the biomarkers study. Our goal was to identify markers associated with joint space width (JSW), which indicates osteoarthritis severity. Candidate covariates we considered included health and demographic ones measured at the first visit (age, sex, body mass index (BMI) and Kellgren-Lawrence stage), which are subject-level covariates that do not depend on time, the Western Ontario and McMaster

Universities Osteoarthritis Index (WOMAC) divided in three categories (pain, stiffness and physical function), as explained in McConnell et al. (2001), as well as eleven blood biomarkers. Magnetic resonance imaging (MRI) results were also included: the cumulative score of marginal osteophytes, cartilage morphology, cartilage signals, bone marrow edema, patella, subarticular cysts and bone attrition in several sub-regions of the knee joint, as detailed in Appendix Appendix A. Other MRI criteria include an index for specific cysts, and intra-articular loose bodies. More information about how these criteria were measured can be found in Peterfy et al. (2004). The final data for analysis contained four binary subject levels, four binary time-dependent covariates, and forty continuous time-dependent candidate covariates. A detailed list of candidate covariates can be found in Appendix A. Some correlations are high in the dataset with 2.6% above 0.4, but most of the correlations (89.7%) are between -0.2 and 0.2.

The study has missing and censored data. Section 6 suggests ways to integrate these censored cases, but we do not study them here since they can be put aside (our investigations show that the missingness is covariate-dependent but not related to outcomes at any time instant). Censored data are therefore not imputed nor compensated for, but represent a loss of 29.52%, 12.65% and 7.23% of patients for the second, third and fourth visits respectively. There is 2.95% of data missing, with 80.56% of rows having no missing data, and 60% of patients with no missing data. Subject-dependent covariates are not affected by missing data; the most affected variable had 8.64% of data missing.

We based our analyses on five imputations performed with the `R` package `mice`. Predictive mean matching and logistic regression were used for continuous and binary variables respectively. The imputation model for each covariate included all other candidate covariates, the response, and the time variable, as stated in our simulation protocol. After specifying a series of tuning parameters λ, the evolution of MI-PGEE coefficients and BIC values as λ changes are shown in Figure 1.

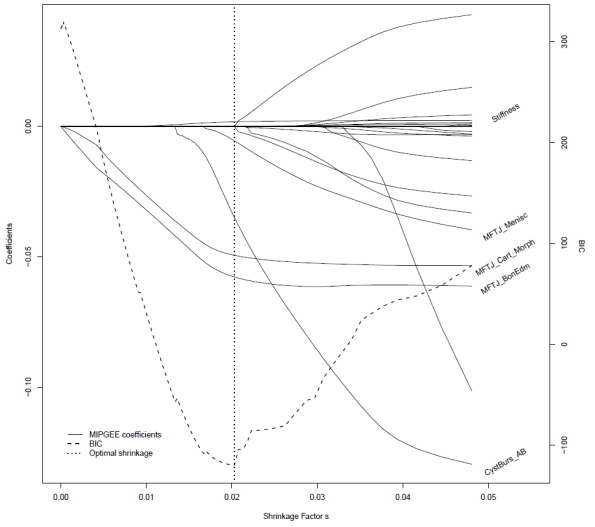


Figure 1: Evolution of MI-PGEE coefficients. The vertical dotted line represents the standard shrinkage rate $s = 0.02$ for the model selected by the smallest BIC. Only the names of selected covariates are displayed.

The y-axis (left) shows the average regression coefficient estimate:

$$\frac{1}{5}\sum_{m=1}^{5}\hat{\beta}_{m,j},$$

for $j = 1, \ldots, 48$, calculated from the MI-PGEE given in (7) with an autoregressive structure. The y-axis (right) shows the BIC values, defined in (8). The x-axis is the standardized shrinkage factor s, defined as

$$s = \frac{\sum_{j=1}^{48}\sqrt{\sum_{m=1}^{5}\hat{\beta}_{m,j}^2}}{\sum_{j=1}^{48}\sqrt{\sum_{m=1}^{5}\tilde{\beta}_{m,j}^2}}.$$

Here, $\tilde{\beta}_{m,j}$ is the GEE estimate for the jth covariate for the mth dataset. This is similar to the one used in Tibshirani (1996) and Yuan and Lin (2006). Drawing estimates over s allows for a better interpretation of results. When the shrinkage factor is equal to zero, no covariates are included in the final model. Conversely, when $s = 1$, no shrinkage is applied, leading to $\hat{\beta}_{m,j} = \tilde{\beta}_{m,j}$ (i.e., all covariates are included in the final model).

Table 4: Estimates (standard errors) and relative standard error (standard error over absolute value of estimate) with MI-GEE for the SEKOIA study.

	estimate (s.e.)	relative s.e.
intercept	3.886 (0.123)*	0.032
MFTJ Bone Edema	-0.094 (0.033)*	0.352
MFTJ Menisc	-0.102 (0.043)*	0.424
MFTJ Cart Morph	-0.070 (0.021)*	0.299
Stiffness	0.004 (0.001)*	0.260
Cysts Bursae AB	-0.119 (0.133)	1.121

*Statistically significant effect is observed based on 95% confidence intervals (estimate $\pm 1.96 \times$standard error) that do not contain 0.

Profiles are shown for s between 0 and 0.05 for visibility purposes. BIC achieves its optimal value when $s = 0.02$, which is represented by a vertical dashed line in Figure 1. Covariates with non-zero coefficients were bone marrow edema, meniscal index and cartilage morphology measured in the medial tibiofemoral joint, stiffness of the Womac index, and AB-type cyst bursae.

We used MI-PGEE for variable selection and then MI-GEE and Rubin's rules, detailed in equations (3) and (4), to obtain combined regression coefficient estimates and 95% CIs; results are shown in Table 4. Using a LASSO-type method for variable selection and an estimation method for inference is similar to the work done by Efron et al. (2004) and discussed in Bühlmann and Van De Geer (2011). Our method chooses a model with only five covariates from the forty-eight present. The three cumulative scores on the medial tibiofemoral joint (MFTJ) variables have been selected, with statistically significant effects. As expected, these criteria are negatively associated with JSW, meaning that the more one has abnormalities in that region, the lower ones JSW will be, which indicates a more damaged knee joint. The three cumulative scores are bone marrow edema, menisc, and cartilage morphology. For clinicians, it could be of interest to observe that three important and different components of the

joint have a noticeable impact. This is in accordance with osteoarthritis being considered not only a disease of the cartilage, but of the whole knee joint. Specifically, the JSW is lower among people with higher bone oedema, cartilage morphology and meniscal scores in the medial tibiofemoral joint. Other things being equal, stiffness of the Womac index happens to be positively associated with JSW, which was not expected. A higher stiffness score is unlikely to be associated with a higher JSW and thus a healthier knee. AB-type cysts bursae is selected and negatively associated with JSW, with a non-significant effect. Interpretation of these two variables is challenging, and requires further investigation. The final model only includes medial tibiofemoral indices. Note that due to the inclusion criterion of the SEKOIA study, only patients with primary knee osteoarthritis of the medial tibiofemoral compartment are included.

6. Concluding Remarks

In this paper, we address the problem of variable selection in a longitudinal context with missing data. Missing data arises often in large complex studies and can be challenging for data analysis. The widely-used multivariate imputation by chained equation method is flexible, easy to implement and available in many software packages. We have proposed the MI-PGEE method, as an extension of MI-LASSO, for variable selection on longitudinal data with multiply-imputed datasets. MI-PGEE applies PGEE with ridge penalties and adaptive weights that are common to the group of estimated regression coefficients of the same variable across multiply-imputed datasets. In this manner, selection is unique across multiply-imputed datasets, which means that coefficients are either all zero or all non-zero for all imputations, and intra-subject correlations are integrated by the working correlation matrix defined for the GEE. Our simulations show that MI-PGEE can select important covariates while integrating missing data and within subject correlation. The use of PGEE on simple imputation data leads to smaller specificity and higher error with correlated covariates than MI-PGEE. Furthermore, MAR impacts to a higher degree simple imputation,

which results in worse specificity and sensibility than MI-PGEE. We therefore recommend the use of MI-PGEE, since it seems more robust to MAR data.

Our simulations were conducted with a continuous response because of the nature of the criterion of interest in our application. However, the method is applicable to discrete outcome cases since the BIC-like criterion can be modified thanks to the weighted deviance. GEE for longitudinal data often use an autoregressive structure for the working correlation matrix. Nevertheless, the user may choose the structure that best fits their data, using criteria that have been developed for that purpose (Pan, 2001; Hin and Wang, 2009). Also, the method is not applicable to longitudinal data only. In the same way GEE can be applied to any type of correlated or clustered data, MI-PGEE can be used on other kinds of correlated data.

We included autoregressive structure in our simulations and application because of the longitudinal aspect of databases. However, there could be a loss of efficiency when time-dependent covariates do not respect certain conditions (Fitzmaurice, 1995; Diggle, 2002; Lai and Small, 2007). We used exogenous covariates in our simulations, and supposed that covariates were linked to the clinical criterion at time t in our application, meaning that any correlation structure could be used. However, when the assumed model includes lagged covariates, as in Diggle (2002) or Lai and Small (2007), only an independent working structure will provide unbiased estimates. To deal with such issues, we suggest using an independent working matrix when the full covariate conditional mean assumption of Lai and Small (2007) is not satisfied.

We conducted our simulations with 10 imputations in order to combine efficiency and sufficiently good estimates of standard errors. However, changes in imputation may provide different model selection. We thus recommend using a number of imputation which provides a stable selection. Following Bodner (2008), if there is 10% missing data on average in the dataset, we suggest running MI-PGEE with 8 to 12 imputations several times, then fix the number of imputations based on this. If results tend to vary across different runs while having a reasonable number of imputations, variables that are not selected across

runs should be examined further, and their model for imputation should be examined closely.

When dealing with pointwise missing data, the MCAR or MAR assumption is often made. Unfortunately, this may not be valid, and further investigations with sensitivity analysis may be needed. Rubin (2004) described how to use multiple imputation with MNAR data and Molenberghs et al. (2014, chap. 19) detailed sensitivity analysis which could be of use to practicioners. More specifically, Van Buuren and Groothuis-Oudshoorn (2011) described some of the procedures that can be performed with `mice`. One possible limitation of the method is the rate of missing data. We studied the effect of having a low rate, but can easily imagine a database with 60% missing data or more (combined data with few common covariates). In this case, even if MI methods lead to acceptable imputation, the effect of certain covariates could be diminished. Further investigations on to what extent the method is robust as more and more missing data exists, would be required.

Censored data in the SEKOIA study was not the focus here. However, when a study suffers from dropout, the distribution of missing data should be examined. MI-PGEE relies on GEE, which usually assume non-informative dropout (MCAR). Nevertheless, informative dropout (MAR) could also interfere in variable selection. A possible improvement of the method is to incorporate an inverse probability weight matrix into our algorithm, to take into account such cases. This would allow us to manage instances where missingness depended on observed responses. Unfortunately, covariates linked to missingness patterns could themselves suffer from missing data. Therefore, weights could be estimated after the multiple imputation procedure on each imputed dataset, then the modified MI-PGEE could be applied. In the case of MNAR dropout, more hypotheses have to be made and specific sensitivity analyses are required.

We compared MI-PGEE with SI-PGEE, the method of reference when dealing with a large number of covariates. In a setting with a smaller number of covariates, one could investigate the quality of selection of MI-PGEE compare to model selection criterion such as MI-QIC or MILIC (Shen and Chen, 2013).

This will be the subject of future research.

Acknowledgements

We gratefully acknowledge C. Gabarroca for her support and C. Guinot and Professor A. Latouche for useful comments. This work was partially funded by grants from the ANRT (Association Nationale de la Recherche et de la Technologie), IRIS (Institut de Recherches Internationales Servier), CIFRE number 2013/0831. We also would like to thank reviewers and the associate editor for helpful comments and suggestions that substantially improved the manuscript.

References

Blommaert A, Hens N, Beutels P. Data mining for longitudinal data under multicollinearity and time dependence using penalized generalized estimating equations. Computational Statistics & Data Analysis 2014;71:667–80.

Bodner TE. What improves with increased missing data imputations? Structural Equation Modeling 2008;15(4):651–75.

Bühlmann P, Van De Geer S. Statistics for high-dimensional data: methods, theory and applications. Springer Science & Business Media, 2011.

Cantoni E, Flemming J, Ronchetti E. Variable selection for marginal longitudinal generalized linear models. Biometrics 2005;61(2):507–14.

Chen Q, Wang S. Variable selection for multiply-imputed data with application to dioxin exposure study. Statistics in medicine 2013;32(21):3646–59.

Cooper C, Reginster JY, Chapurlat R, Christiansen C, Genant H, Bellamy N, Bensen W, Navarro F, Badurski J, Nasonov E, et al. Efficacy and safety of oral strontium ranelate for the treatment of knee osteoarthritis: rationale and design of randomised, double-blind, placebo-controlled trial. Current Medical Research & Opinion 2012;28(2):231–9.

Diggle P. Analysis of longitudinal data. Oxford University Press, 2002.

Dziak J, Li R. Variable selection with penalized generalized estimating equations. Technical Report; The Methodology Center, Pennsylvania State University; 2006.

Efron B, Hastie T, Johnstone I, Tibshirani R, et al. Least angle regression. The Annals of statistics 2004;32(2):407–99.

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association 2001;96(456):1348–60.

Fitzmaurice GM. A caveat concerning independence estimating equations with multivariate binary data. Biometrics 1995;:309–17.

Fu W. Penalized estimating equations. Biometrics 2003;59(1):126–32.

Fu WJ. Nonlinear gcv and quasi-gcv for shrinkage models. Journal of statistical planning and inference 2005;131(2):333–47.

Halekoh U, Højsgaard S, Yan J. The r package geepack for generalized estimating equations. Journal of Statistical Software 2006;15(2):1–11.

Hin LY, Wang YG. Working-correlation-structure identification in generalized estimating equations. Statistics in medicine 2009;28(4):642.

Horton NJ, Lipsitz SR. Multiple imputation in practice: comparison of software packages for regression models with missing variables. The American Statistician 2001;55(3):244–54.

Lai TL, Small D. Marginal regression analysis of longitudinal data with time-dependent covariates: A generalized method-of-moments approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2007;69(1):79–99.

Li G, Lian H, Feng S, Zhu L. Automatic variable selection for longitudinal generalized linear models. Computational Statistics & Data Analysis 2013;61:174–86.

Liang K, Zeger S. Longitudinal data analysis using generalized linear models. Biometrika 1986;:13–22.

Little RJ. Regression with missing x's: a review. Journal of the American Statistical Association 1992;87(420):1227–37.

McConnell S, Kolopack P, Davis AM. The western ontario and mcmaster universities osteoarthritis index (womac): a review of its utility and measurement properties. Arthritis care & research 2001;45(5):453–61.

Molenberghs G, Fitzmaurice G, Kenward MG, Tsiatis A, Verbeke G. Handbook of Missing Data Methodology. CRC Press, 2014.

Pan W. Akaike's information criterion in generalized estimating equations. Biometrics 2001;57(1):120–5.

Peterfy C, Guermazi A, Zaim S, Tirman P, Miaux Y, White D, Kothari M, Lu Y, Fye K, Zhao S, et al. Whole-organ magnetic resonance imaging score (worms) of the knee in osteoarthritis. Osteoarthritis and Cartilage 2004;12(3):177–90.

R Development Core Team . R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria; 2008. URL: `http://www.R-project.org`; ISBN 3-900051-07-0.

Reginster JY, Badurski J, Bellamy N, Bensen W, Chapurlat R, Chevalier X, Christiansen C, Genant H, Navarro F, Nasonov E, et al. Efficacy and safety of strontium ranelate in the treatment of knee osteoarthritis: results of a double-blind, randomised placebo-controlled trial. Annals of the rheumatic diseases 2012;:annrheumdis–2012.

Rubin D. Multiple imputation for nonresponse in surveys. volume 81. John Wiley & Sons, 2004.

Rubin DB. Multiple imputation for nonresponse in surveys (wiley series in probability and statistics) 1987;.

Rubin DB. Multiple imputation after 18+ years. Journal of the American statistical Association 1996;91(434):473–89.

Schafer JL. Analysis of incomplete multivariate data. CRC press, 1997.

Schafer JL. Multiple imputation: a primer. Statistical methods in medical research 1999;8(1):3–15.

Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychological methods 2002;7(2):147.

Shen C, Chen Y. Model selection of generalized estimating equations with multiply imputed longitudinal data. Biometrical Journal 2013;55(6):899–911.

Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological) 1996;:267–88.

Ueki M. A note on automatic variable selection using smooth-threshold estimating equations. Biometrika 2009;96(4):1005–11.

Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical methods in medical research 2007;16(3):219–42.

Van Buuren S, Brand JP, Groothuis-Oudshoorn C, Rubin DB. Fully conditional specification in multivariate imputation. Journal of statistical computation and simulation 2006;76(12):1049–64.

Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in r. Journal of statistical software 2011;45(3).

Wang L, Zhou J, Qu A. Penalized generalized estimating equations for high-dimensional longitudinal data analysis. Biometrics 2012;68(2):353–60.

White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Statistics in medicine 2011;30(4):377–99.

Table A.5: Table of the MRI cumulative score for the medial tibiofemoral joint (MFTJ), lateral tibiofemoral joint (LFTJ), patellofemoral joint (PFJ), and the portion of the tibia beneath the tibial spines (Sub) measured in the SEKOIA study.

	MFTJ	LFTJ	PFJ	Sub
Bone marrow edema	X	X	X	X
Cartilage morphology	X	X	X	
Cartilage signal	X	X	X	
Marginal osteophytes	X	X	X	
Patella	X	X		
Bone attrition	X	X	X	
Sub-articular cysts	X	X	X	X

Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? Statistics in medicine 2008;27(17):3227–46.

Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2006;68(1):49–67.

Appendix A. Details about the SEKOIA study

Thirty MRI criteria were measured. Twenty-two were cumulative scores from various joint regions as detailed in Table A.5, seven are cysts, and the last is an index of intra-articular loose bodies. The medial tibiofemoral joint is composed of medial femoral central, medial femoral posterior, medial tibial anterior, medial tibial central, and medial tibial posterior subdivisions. The lateral tibiofemoral joint is composed of lateral femoral central, lateral femoral posterior, lateral tibial anterior, lateral tibial central, and lateral tibial posterior subdivisions. The patellofemoral joint comprised the medial patella and lateral patella regions. Details of subdivisions can be found in Figure A.2.

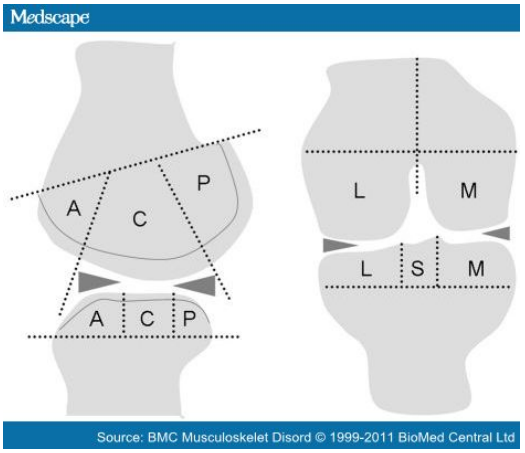


Figure A.2: Regional subdivision of the articular surfaces. The joint is composed of the femur (upper bone) and the tibia which are divided into the medial (M) and the lateral (L) part and the subspinous (S) for the tibia. Each of these regions are divided in sub regions : anterior (A), central (C) and posterior (P).

Appendix B. Simulation results for binary covariates and missing completely at random data

Table B.6: Mean squared error (MSE), relative MSE (RME), sensitivity (SEN) and specificity (SPE) with binary covariates and missing completely at random data. Comp is for PGEE with complete data before MCAR is imposed, and SI is for PGEE with simple imputation. In bold are the best results between SI-PGEE and MI-PGEE.

		5% MCAR		10% MCAR	
	Comp	SI	MI-PGEE	SI	MI-PGEE
1st Pattern					
MSE	0.424	**0.501**	0.548	0.773	**0.677**
RME	0.305	**0.365**	0.528	0.533	**0.468**
SEN	0.865	**0.812**	0.731	**0.760**	0.715
SPE	0.693	0.726	**0.863**	0.783	**0.836**
2nd Pattern					
MSE	0.222	0.325	**0.272**	0.453	**0.315**
RME	0.232	0.403	**0.379**	0.452	**0.410**
SEN	0.798	0.781	**0.807**	0.685	**0.764**
SPE	0.763	0.750	**0.848**	0.738	**0.805**
3rd Pattern					
MSE	0.408	0.531	**0.491**	0.634	**0.551**
RME	0.299	0.445	**0.308**	0.427	**0.421**
SEN	0.778	**0.689**	0.640	**0.666**	0.622
SPE	0.703	0.739	**0.859**	0.786	**0.834**

le cnam

Julia GERONIMI

Contribution à la sélection de variables en présence de données longitudinales

le cnam

Résumé :
Dans le cadre des études cliniques, de nombreuses variables peuvent être mesurées de façon répétée dans le temps. Lorsque l'objectif de l'analyse est de les relier à un critère clinique d'intérêt, les méthodes de régularisation de type LASSO, généralisées aux Generalized Estimating Equations (GEE) permettent de sélectionner un sous-groupe de variables en tenant compte des corrélations intra-patients. Cependant, les bases de données présentent souvent des données non renseignées et des problèmes de mesures ce qui entraîne des données manquantes inévitables. L'objectif de ce travail de thèse est d'intégrer ces données manquantes pour la sélection de variables en présence de données longitudinales. Nous utilisons la méthode d'imputation multiple et proposons une fonction d'imputation pour le cas spécifique des variables soumises à un seuil de détection. Nous proposons une nouvelle méthode de sélection de variables pour données corrélées qui intègre les données manquantes : le Multiple Imputation Penalized Generalized Estimating Equations (MI-PGEE). Notre opérateur utilise la pénalité group-LASSO en considérant l'ensemble des coefficients de régression estimés d'une même variable sur les échantillons imputés comme un groupe. Notre méthode permet une sélection consistante sur l'ensemble des imputations, et minimise un critère de type BIC pour le choix du paramètre de régularisation. Nous présentons une application sur l'arthrose du genoux où notre objectif est de sélectionner le sous-groupe de biomarqueurs qui expliquent le mieux les différences de largeur de l'espace articulaire au cours du temps.
Mots clés :
sélection de variables, données longitudinales, équations d'estimation généralisées, données corrélées, données manquantes, imputation multiple.

Abstract :
Clinical studies enable us to measure many longitudinales variables. When our goal is to find a link between a response and some covariates, one can use regularisation methods, such as LASSO which have been extended to Generalized Estimating Equations (GEE). They allow us to select a subgroup of variables of interest taking into account intra-patient correlations. Unfortunately, databases often have unfilled data and measurement problems resulting in inevitable missing data. The objective of this thesis is to integrate missing data for variable selection in the presence of longitudinal data. We use mutiple imputation and introduce a new imputation function for the specific case of variables under detection limit. We provide a new variable selection method for correlated data that integrate missing data : the Multiple Imputation Penalized Generalized Estimating Equations (MI-PGEE). Our operator applies the group-LASSO penalty on the group of estimated regression coefficients of the same variable across multiply-imputed datasets. Our method provides a consistent selection across multiply-imputed datasets, where the optimal shrinkage parameter is chosen by minimizing a BIC-like criteria. We then present an application on knee osteoarthritis aiming to select the subset of biomarkers that best explain the differences in joint space width over time.
Keywords :
variable selection, longitudinal data, generalized estimating equations, correlated data, missing data, multiple imputation.