Estimation dans des modèles de fragilité avec des structures de corrélation complexes via des algorithmes d'approximation stochastique

Ajmal Oodally

université
PARIS-SACLAY

Estimation in frailty models with complex correlation structures through stochastic approximation algorithms

Thèse de doctorat de l'université Paris-Saclay

Ecole Doctorale n° 574,
Ecole Doctorale de Mathématique Hadamard (EDMH)
Spécialité de doctorat: Mathématiques aux interfaces
Unité de recherche: Université Paris-Saclay, INRAE, Mathématiques et Informatique Appliquées du Génome à l'Environnement (MaIAGE), 78350, Jouy-en-Josas, France
Référent: Faculté des sciences d'Orsay

Thèse présentée et soutenue à Orsay, le 28 Septembre 2020, par

Ajmal OODALLY

Composition du jury:

Agathe GUILLOUX Professeure des Universités, Université d'Evry Val d'Essonne	Présidente
Adeline LECLERCQ SAMSON Professeure, Université Grenoble Alpes	Rapporteure et Examinatrice
Ingrid VAN KEILEGOM Professeure, Katholieke Universiteit Leuven	Rapporteure et Examinatrice
Aurélien LATOUCHE Professeur des Universités, Conservatoire national des arts et métiers	Examinateur
Andreas WIENKE Professeur, University Halle-Wittenberg	Examinateur
Estelle KUHN Directrice de recherche, INRAE	Directrice de thèse
Luc DUCHATEAU Professeur, Ghent University	Codirecteur de thèse

Thèse de doctorat

NNT: 2020UPASM003

“If I have seen further it is by standing on the shoulders of Giants.” - Isaac Newton

Ignis Vibrante Lumine

Remerciements

Je tiens tout d'abord à remercier mes directeurs de thèse qui m'ont accompagné tout au long de cette aventure. Ils ont fait preuve d'une grande patience et pédagogie et ont ainsi largement contribué au succès de cette thèse. Estelle, tu m'as non seulement guidé pendant ces trois ans et demi mais tu m'as aussi été d'une grande aide pour l'après thèse. J'ai beaucoup apprécié ta gentillesse et ta bienveillance. Merci pour tout.

Luc, ton accueil chaleureux lors de toutes mes visites à Ghent, ton humour et tes nombreuses autres qualités humaines vont beaucoup me manquer. C'était un vrai plaisir de travailler avec toi et Estelle. Je garde un bon souvenir de notre dîner au Chateaubriand quelques jours avant le confinement.

Je remercie Adeline Leclercq Samson et Ingrid Van Keilegom d'avoir accepté de rapporter ma thèse et pour leurs regards très appréciés sur mes travaux. Un grand merci à Aurélien Latouche, Andreas Wienke et Agathe Guilloux pour leur participation à ma soutenance de thèse dans ces conditions particulières et pour toutes leurs remarques très pertinentes. Je tiens aussi à remercier Christine Kéribin qui m'a suivi depuis mes débuts en Master et pour sa participation à mes comités de thèse.

J'ai passé trois très belles années dans un lieu atypique qui m'a tout de suite tapé dans l'oeuil. Et oui, ce n'est pas donné à tous les doctorants de préparer une thèse entouré de chevaux, de vaches, d'un potager et dans un bureau plein de vie; que ce soit les bactéries suite à des restes de desserts laissés à l'abandon, les graines et plantes disséminées un peu partout. Ce bureau aura aussi servi d'atelier à Maxime qui nous a régulièrement fait profiter de ses talents de pâtissier. Merci Romain pour ta bonne humeur, joie de vivre et ta sagesse (voix sarcastique). Merci à tous mes autres collègues de l'INRAE qui ont tous contribué à rendre mon environnement de travail fort agréable. Merci à Sandra, Gildas, François, Maxime, Patrick, Simon, Olivier, Béatrice, Elisabetha, Maud, Catherine, Laurent, Pierre, Ousmane, Henri, Lina et tant d'autres.

Comment ne pas mentionner mon collègue de bureau Léo, ce beau spécimen landais, un rêveur comme dirait son grand-père. Tu auras rendu cette thèse fort agréable et unique. Cette fameuse sortie à Trouville dans une combinaison trop grande aura été le début d'une nouvelle passion. Nos innombrables conversations sans filtres et totalement décomplexées à la cantine et au coin café avec un public (Romain, Lina, Henri) toujours à l'écoute.

Une pensée spéciale à Emile qui a été à mes côtés depuis mes premiers pas en France. On en aura vécu des choses, la super semaine de Noël passée chez toi à se goinfrer suivi d'une longue période de collocation pleine de

rebondissements. Trugarez ! Breizh atav !

Mille mercis à Valérie, Elisabeth, Sylvie et toute l'équipe de la MISS. Je pense notamment à tous les doctorants avec qui j'ai animé tant d'ateliers et les autres que j'ai souvent croisé pendant les pauses déjeuner et apéro de fin d'année.

J'adresse toute ma reconnaissance à ma femme pour son soutien sans faille que ce soit dans les moments difficiles ou les moments de joie. Tu as toujours fait preuve d'une patience hors norme à mon égard et tu as été à mes côtés depuis tellement longtemps que je ne peux plus imaginer ma vie sans toi.

Un grand merci à ma tante qui a relu toutes les lettres de motivations que j'ai écrites. Je serai toujours reconnaissant pour toute l'aide qu'elle m'a apportée

Enfin, je remercie ma mère, mon père, mon petit frère et ma petite soeur pour leur soutient indéfectible. Sans eux, la réussite de cette thèse ne serait pas possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Analyse de survie

L'analyse de survie est une branche des statistiques visant à analyser la durée attendue jusqu'à ce qu'un ou plusieurs événements se produisent. La première analyse de survie est apparue au début du vingtième siècle. Le premier domaine d'application concerné est celui de l'actuariat. Elle est utilisée dans le domaine médical pour la première fois en 1950. Par contre, la notion de table de survie (aussi appelé table de mortalité) est antérieure à ces domaines et a été introduite pour la première fois par John Graunt au XVII^{e} siècle, considéré par beaucoup comme l'un des premiers démographes (cf. Greenwood (1938)). Ayant pour objectif de détecter l'apparition de la peste bubonique à Londres, il avait analysé les bulletins de décès publiés hebdomadairement. Il est notamment reconnu pour avoir produit et largement diffusé la première table de mortalité, donnant des probabilités de survie en fonction des tranches d'âge. Depuis, la survenue d'un événement est souvent qualifiée d'échec, généralement attribuée au fait que l'événement soit un décès ou une maladie. Cependant, au cours des dernières décennies, les méthodes statistiques pour l'analyse des données de survie ont été étendues au-delà de la recherche biomédicale ou actuarielle à d'autres domaines tels que la criminologie, la sociologie et l'informatique. Les travaux de Canfora et al. (2011) concerne l'application de l'analyse de survie visant à étudier le risque de ne pas corriger un bug informatique dans un laps de temps donné. Dans le domaine de la criminologie, des détenus adultes libérés du Département correctionnel de l'Oklahoma ont été suivis et la récidive, mesurée en temps de retour à l'incarcération, a été étudiée à l'aide de méthodes d'analyse de survie par Spivak and Damphousse (2006).

Depuis, plusieurs modèles et des travaux s'orientant dans différentes directions ont permis d'enrichir ce domaine.

1.1.1 Fonctions de survie et de risque

Fonction de survie :

La quantité centrale en analyse de survie est la durée de survie. Le terme de durée de survie désigne le temps écoulé jusqu'à la survenue d'un événement d'intérêt. On note T ce temps écoulé. On suppose que T est une variable aléatoire de fonction de répartition F. On définit la fonction de survie S au temps t par la probabilité que l'événement d'intérêt survienne après un instant t fixé :

$$\forall t \geq 0,\, S(t) = P(T > t) = 1 - F(t) = \int_t^{\infty} f(x)dx$$

Par analogie, la fonction de répartition F représente, pour t fixé, la probabilité que l'événement d'intérêt survienne avant l'instant t.

Fonction de risque instantané :

La fonction de risque instantané h caractérise la probabilité que l'événement d'intérêt survienne au cours d'une courte durée dt après l'instant t donné sachant que l'événement ne s'est pas produit avant cet instant t.

$$\forall t \geq 0,\, h(t) = \lim_{dt \to 0^+} \frac{P(t \leq T < t + dt | T \geq t)}{dt}$$

Le numérateur représente la probabilité conditionnelle que l'événement survienne dans l'intervalle $[t, t+dt)$ étant donné qu'il n'est pas survenu avant l'instant t, et le dénominateur est égale à la largeur de l'intervalle. Le quotient permet donc d'obtenir un taux d'occurrence d'événements par unité de temps. La valeur limite quand la largeur de l'intervalle tend vers zéro donne le risque instantané au temps t. C'est une fonction positive ou nulle et son intégrale sur $[0, \infty]$ est infinie. Hormis ces deux contraintes, elle peut croître, décroître, être non-monotone, non continue.

Fonction de risque cumulé :

La fonction de risque cumulé, aussi appelé taux de hasard cumulé, est l'intégrale du risque instantané et s'écrit comme suit :

$$\forall t \geq 0,\, H(t) = \int_0^t h(u)du \tag{1.1}$$

Ces quantités sont reliées et peuvent donc s'exprimer les unes en fonction des autres.

$$\forall t \geq 0,\, S(t) = 1 - F(t) \tag{1.2}$$

$$\forall t \geq 0,\, f(t) = \frac{\partial}{\partial t} F(t) \tag{1.3}$$

$$\forall t \geq 0,\, h(t) = \frac{f(t)}{S(t)} \tag{1.4}$$

$$\forall t \geq 0,\, S(t) = \exp\,(-H(t)) \tag{1.5}$$

1.1.2 Observation censurée des durées de survie

Une caractéristique qui distingue l'analyse de survie des autres domaines des statistiques est la censure. La censure se produit lorsque des informations incomplètes sont disponibles sur la durée de survie. C'est un phénomène courant en analyse de survie et doit donc être pris en compte. Il existe plusieurs mécanismes qui peuvent conduire à des données censurées. On considère un échantillon de taille n composée d'observations i, $i = 1, \ldots, n$.

Censure de type I

Sous censure de type I, l'échantillon est étudié pendant un temps fixe τ. Le nombre d'observations pour lesquelles l'événement survient est aléatoire mais la durée totale de l'étude étant fixée, le temps maximal considérée est égal à τ.

Censure de type II

Sous censure de type II, l'échantillon de taille n est suivi jusqu'à ce que l'événement survienne pour m observations. Ce nombre m est fixé à l'avance. La durée totale de l'étude est alors aléatoire et inconnue.

Censure aléatoire

De façon plus générale, on considère la censure comme un phénomène aléatoire. Sous censure aléatoire, on associe à chaque observation un temps de censure C_i et une durée de survenue de l'événement T_i. Ces deux variables aléatoires sont usuellement supposées indépendantes. On observe alors $X_i = \min(T_i, C_i)$, et un indicateur de censure noté $\Delta_i = \mathbb{1}_{T_i \leq C_i}$ qui nous indique si l'observation i est censurée ou pas.

Une donnée peut-être censurée d'un côté comme de l'autre et aussi des deux côtés. Ces trois catégories de censure sont la censure à doite, la censure à gauche et la censure par intervalle. Nous illustrons ces différents types de censure dans le cas d'une étude clinique.

Censure à droite

La censure à droite se produit lorsqu'un patient quitte l'étude avant qu'un événement ne se produise ou l'étude se termine avant que l'événement ne se soit produit.

Censure à gauche

La censure à gauche se produit lorsque l'événement a lieu avant le début de l'étude pour un patient et le moment exact de survenue n'est pas connu.

Censure par intervalle

La censure par intervalle se produit lorsque l'on sait que l'événement a lieu dans un certain intervalle de temps mais que l'instant exact de survenue de l'événement n'est pas connu.

On fait l'hypothèse que le mécanisme de censure est non informatif, c'est-à-dire que la censure d'une observation ne doit fournir aucune information concernant la survenue de l'événement pour cette observation particulière au-delà de la période de censure. Cette hypothèse est indispensable pour l'analyse des modèles classiques d'analyse de survie. Dans la suite, on considère uniquement le cas de censure aléatoire à droite et non informative. Nous nous référons à Klein and Moeschberger (2006) pour une étude plus approfondie de la censure.

1.2 Estimateurs non paramétriques des fonctions de survie et de risque

1.2.1 Estimateur non paramétrique de la fonction de survie

L'estimation non paramétrique de la fonction de survie se fait assez facilement en s'inspirant de la définition de la fonction. Si les données ne sont pas censurées, l'estimateur empirique de cette fonction s'écrit :

$$\hat{S}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{T_i > t}$$

L'estimateur est tout simplement la proportion d'observations pour lesquelles l'événement n'est pas encore survenue au temps t. Cet estimateur a été adapté par Kaplan and Meier (1958) pour prendre en compte des données censurées. Nous définissons la statistique d'ordre des temps d'événements par $T_{(1)} < T_{(2)} < \cdots < T_{(n)}$, le nombre d'observations qui subissent l'événement au temps $T_{(i)}$ par d_i et le nombre d'observations à risque au temps $T_{(i)}$ par r_i. L'estimateur de Kaplan-Meier s'écrit alors :

$$\hat{S}(t) = \prod_{i:T_{(i)} \leq t} \left(1 - \frac{d_i}{r_i}\right)$$

Cet estimateur, aussi appelé estimateur produit-limite, est l'estimation du maximum de vraisemblance non-paramétrique de la fonction de survie $S(t)$ (cf. Kaplan and Meier (1958) pour plus de détails). Nous proposons une explication heuristique de $\hat{S}(t)$. Nous nous plaçons dans le contexte d'une étude de la durée de vie de patients malades qui luttent contre une maladie. En considérant l'objectif qui est de survivre jusqu'au temps t, il faut d'abord

être toujours vivant au temps $T_{(1)}$. La prochaine étape consiste à survivre de $T_{(1)}$ à $T_{(2)}$ sachant que le patient a survécu jusqu'au temps $T_{(1)}$ et ainsi de suite. On estime la probabilité conditionnelle de mourir au temps $T_{(i)}$ étant donné que le patient était vivant juste avant par $\frac{d_i}{r_i}$. La probabilité conditionnelle de survivre au temps $T_{(i)}$ est le complément de la quantité précédente: $1 - \frac{d_i}{r_i}$. La probabilité de survivre jusqu'au temps t est obtenue en multipliant les probabilités conditionnelles pour tous les temps pertinents jusqu'au temps t. La consistance de l'estimateur $\hat{S}(t)$ a été démontrée par Kaplan and Meier (1958) et la normalité asymptotique par Breslow and Crowley (1974).

1.2.2 Estimateur non paramétrique de la fonction de risque cumulé

Une première approche consiste à estimer $\hat{S}(t)$ et à utiliser l'équation (1.5) qui lie $S(t)$ à $H(t)$. Un estimateur possible du risque cumulé s'écrit tout simplement: $-\log(\hat{S}(t))$. Il existe aussi un estimateur qui permet d'estimer directement $H(t)$ sans passer par la fonction de survie $S(t)$. Cet estimateur est appelé estimateur de Nelson-Aalen et s'écrit comme suit :

$$\hat{H}(t) = \sum_{i:T_i \leq t} \frac{d_i}{r_i}$$

Ainsi, $\hat{H}(t)$ est une fonction en escalier croissante continue à droite avec des incréments de $\frac{d_i}{r_i}$ aux instants de survenue d'événement. C'est un estimateur du maximum de vraisemblance non paramétrique de $H(t)$; la consistance et la normalité asymptotique de l'estimateur ont été démontrées par Greenwood and Wefelmeyer (1990) dans le cadre du modèle à risque proportionnel de Cox.

Ces deux estimateurs peuvent être utilisés pour approcher la même fonction. Les deux sont asymptotiquement équivalents et le choix d'une approche au détriment de l'autre dépend du contexte comme indiqué dans une étude comparative menée par Colosimo et al. (2002).

1.3 Les modèles de survie paramétriques

Dans la section précédente, nous avons présenté les estimations non paramétriques des fonctions de survie et de risque. Dans certains cas, des informations préalables peuvent être disponibles sur les temps d'événement étudiés. Une deuxième approche consiste à supposer que les temps de survie suivent une certaine distribution. Les modèles de survie paramétriques sont souvent utilisés pour extrapoler des temps de survie au-delà des données de suivi disponibles. Cette particularité fait la popularité des modèles paramétriques dans le domaine de la santé où il est nécessaire de prendre en compte les effets et les coûts sur la survie suite à des interventions médicales (cf. Ishak et al. (2013) pour plus de détails).

Toute distribution de variables aléatoires définie pour $t \in [0, \infty)$ peut être utilisé pour décrire le temps de survenue de l'événement d'intérêt. On se place dans le cadre simple de temps d'événement aléatoire et i.i.d. Les temps

d'événement sont représentés par la variable aléatoire T. Nous donnons quelques exemples de distributions souvent utilisées dans la littérature. Les calculs des quantités associées (densité de probabilité de la variable T, risque instantané, risque cumulé, fonction de survie) ne seront pas explicités et nous ne considérons aucune covariable afin d'alléger les notations. Nous détaillons trois distributions qui seront utilisées dans les chapitres du manuscrit et renvoyons vers Duchateau and Janssen (2008) pour une liste plus exhaustive de distributions possibles.

1.3.1 Le modèle exponentiel

Le modèle exponentiel est le modèle paramétrique le plus simple et suppose un risque constant dans le temps, qui reflète une propriété implicite de la distribution. C'est la propriété d'absence de mémoire. La probabilité que l'événement survienne dans un intervalle de temps particulier dépend uniquement de la longueur de l'intervalle mais pas des valeurs des bornes de cet intervalle. Supposons $T \sim \text{Exp}(\lambda)$. Alors pour $t \geq 0$, $\lambda > 0$,

$$f(t) = \lambda \exp(-\lambda t)$$

$$h(t) = \lambda$$

$$H(t) = \lambda t$$

$$S(t) = \exp(-\lambda t)$$

Il n'y a qu'un paramètre qui caractérise la distribution, ici noté par λ et l'inverse de ce paramètre est égale à la fois à la moyenne et l'écart type de T. Ces caractéristiques en font un modèle très simple. Un exemple d'utilisation concerne la modélisation de la durée de vie d'un système où les pièces sont remplacées en cas de défaillance (cf. Mendenhall and Sincich (2016)). En revanche, en raison de la non flexibilité de la distribution, peu de travaux en font usage et optent plutôt pour des distributions plus flexibles.

1.3.2 Le modèle de Weibull

Le modèle de Weibull est une généralisation du modèle exponentiel avec deux paramètres positifs qu'on note λ et ρ. Le paramètre λ est le paramètre d'échelle et ρ le paramètre de forme. Le paramètre d'échelle caractérise la façon dont la densité est étirée alors que le paramètre de forme comme son nom l'indique est un paramètre qui permet à la densité (de façon équivalente le risque instantané) de prendre une variété de formes en fonction de la valeur du paramètre. Quand la valeur de ρ est inférieure à 1, h diminue de façon monotone en fonction du temps et inversement quand ρ est plus grand que 1, h augmente de façon monotone avec le temps. Dans le cas où $\rho = 1$, on retrouve le modèle exponentiel. Non seulement la distribution est plus flexible, mais les expressions des autres quantités restent simples comme on peut le voir ci-dessous. Supposons $T \sim \text{Weibull}(\lambda, \rho)$. Alors pour $t \geq 0$, $\lambda > 0$,

$\rho > 0$,

$$f(t) = \lambda\rho t^{\rho-1} \exp(-\lambda t^{\rho})$$

$$h(t) = \lambda\rho t^{\rho-1}$$

$$H(t) = \lambda t^{\rho}$$

$$S(t) = \exp(-\lambda t^{\rho})$$

Nous présentons graphiquement l'effet des paramètres ρ et λ sur le risque instantané en fonction du temps.

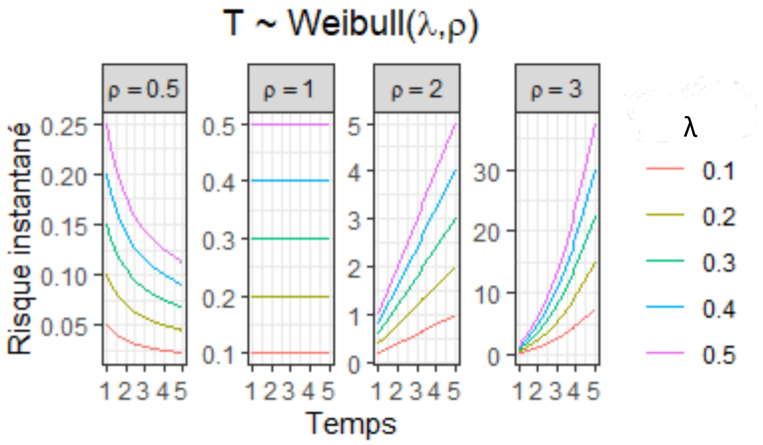

Figure 1.1: Risque instantané en fonction de ρ et λ

La distribution de Weibull peut être utilisée pour modéliser la distribution de survie d'une population à risque croissant, décroissant ou constant, et s'applique donc à de nombreux types de données. Le risque diminue pour $\rho < 1$, est constant pour $\rho = 1$ et augmente pour $\rho > 1$. On remarque que pour $\rho = 1$, la distribution de Weibull est équivalente à une distribution exponentielle paramétrée par λ. Ce modèle étant plus flexible, est plus souvent utilisé dans la littérature comme dans les travaux de Zhu et al. (2011) pour analyser les facteurs pronostiques chez les patients atteints de cancer gastrique. Une généralisation de la loi de Weibull et les nombreuses avantages de cette loi sont étudiées dans les travaux de Mudholkar et al. (1996).

1.3.3 Le modèle de Gompertz

La distribution de Gompertz trouve ses origines en 1825 et a été proposé par l'actuaire britannique Benjamin Gompertz. Il a remarqué une augmentation exponentielle progressive dans le taux de mortalité entre l'âge de maturation sexuelle et la vieillesse. Ces travaux sont encore d'actualités dans les études démographiques (cf. Wilson (1994)) et l'utilisation de cette distribution dans l'analyse de survie est courante. Supposons $T \sim$ Gompertz(α, λ). Alors

pour $t \geq 0$, $\alpha > 0$, $\lambda > 0$,

$$f(t) = \lambda \exp(\alpha t) \exp(-\frac{\lambda}{\alpha}(\exp(\alpha t) - 1))$$

$$h(t) = \lambda \exp(\alpha t)$$

$$H(t) = \frac{\lambda}{\alpha}(\exp(\alpha t) - 1)$$

$$S(t) = \exp(-\frac{\lambda}{\alpha}(\exp(\alpha t) - 1))$$

Nous présentons graphiquement l'effet des paramètres α et ρ sur le risque instantané en fonction du temps.

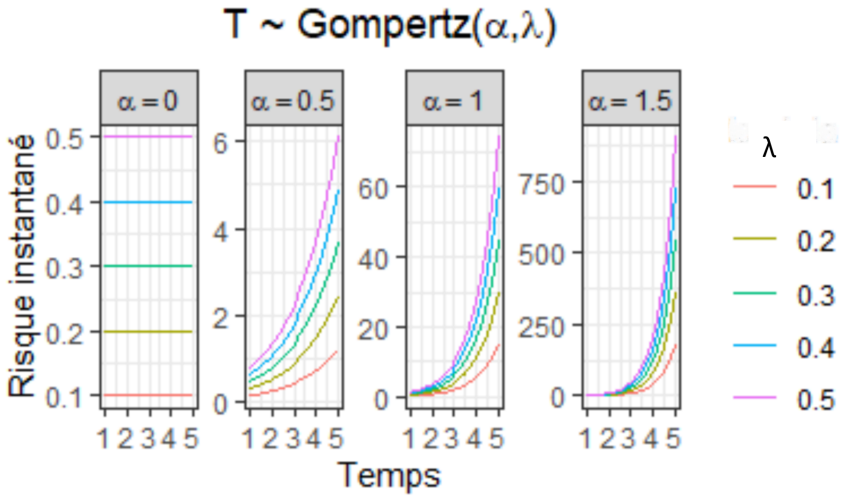

Figure 1.2: Risque instantané en fonction de ρ et λ

Le risque augmente pour $\alpha > 0$ et est constant pour $\alpha = 0$. Lorsque $\alpha = 0$, la loi de Gompertz est équivalente à une loi exponentielle paramétrée par λ. La distribution de Gompertz est caractérisée par le fait que le logarithme du risque instantané est linéaire en t et est donc étroitement liée à la distribution de Weibull où le logarithme du risque instantané est linéaire en logarithme de t.

1.4 Le modèle de Cox

Nous introduisons maintenant le modèle de Cox (1972) qui permet de modéliser le risque instantané et de quantifier l'effet des covariables sur les temps de survie.

1.4.1 Description du modèle

Le premier modèle proposé par Cox (1972), appelé modèle à risque proportionnel exprime le risque instantané comme un produit de deux quantités. Pour $i = 1, \ldots, n$,

$$\forall t \geq 0,\, h(t|Z_i) = h_0(t)\text{exp}\,(Z_i^t\beta) \tag{1.6}$$

où h_0 est la fonction de risque de base qui correspond au risque instantané lorsque toutes les covariables sont nulles, $\beta \in \mathbb{R}^p$ les paramètres d'effet, aussi appelés paramètres de régression, et $Z_i \in \mathbb{R}^p$ les covariables associées. On peut distinguer deux parties, la première ($h_0(t)$) est dépendante du temps contrairement à la seconde (exp $(Z_i^t\beta)$) qui elle ne dépend que des covariables. La forme de $h_0(t)$ n'étant pas précisée, on s'intéresse plutôt à l'association entre les covariables Z_i et la survenue de l'événement d'intérêt. Les hypothèses inhérentes à ce modèle sont les suivantes :

(C1) le rapport des risques instantanés de survenue de l'événement de deux observations doit être indépendant du temps.

(C2) le logarithme du risque est une fonction linéaire des covariables

L'hypothèse **(C1)**, aussi appelée hypothèses de proportionalité des risques, est la plus contraignante. Les données réelles ne permettent généralement pas de faire cette hypothèse. L'hypothèse **(C2)** est l'hypothèse de log-linéarité, i.e, $\log(h(t|Z_i)) = \log(h_0(t)) + Z_i^t\beta$ ce qui implique une relation linéaire entre le logarithme du risque et les covariables.

1.4.2 Estimation des paramètres par maximum de la vraisemblance partielle

La vraisemblance partielle a été définie permettant d'estimer le paramètre d'effet β en s'affranchissant de la fonction de risque de base h_0 (cf. Cox (1975)). Cette vraisemblance partielle ne fait plus intervenir la fonction de risque de base h_0. Le paramètre d'intérêt principal étant β puisqu'il permet de quantifier l'effet des covariables alors que h_0 n'est généralement pas étudié en survie.

On considère un n-échantillon $(X_i, \Delta_i)_{1\leq i\leq n}$ de variables aléatoires distribuées comme (X, Δ). Suivant Cox (1975), l'expression de la vraisemblance partielle s'écrit alors :

$$L^p(\beta; X, \Delta) = \prod_{i=1}^{n} \left(\frac{\text{exp}\,(Z_i^t\beta)}{\sum_{j\in R(X_i)} \text{exp}\,(Z_j^t\beta)} \right)^{\Delta_i} \tag{1.7}$$

où $R(X_{(i)}) = \{1 \leq j \leq n, \quad X_j \geq X_i\}$ est l'ensemble des observations à risque au temps X_i. Les observations à risque au temps X_i sont les observations pour lesquelles l'événement n'est pas encore survenue au temps X_i. L'estimateur de β qu'on note $\hat{\beta}$ est défini comme le maximum de cette vraisemblance partielle.

$$\hat{\beta} = \underset{\beta}{\text{argmax}}\; L^p(\beta; X, \Delta)$$

Sous des hypothèses de régularité, les propriétés asymptotiques telles que la consistance et la normalité asymptotique de l'estimateur $\hat{\beta}$ ont été démontrées par Tsiatis et al. (1981) et Andersen and Gill (1982). Ils ont montré avec élégance ces bonnes propriétés de l'estimateur à l'aide de processus de comptage et de martingales.

1.4.3 Propriétés asymptotiques de l'estimateur

Les travaux de Andersen and Gill (1982) nécessitant une reformulation du modèle avec des processus de comptage, nous présentons les résultats obtenus par Tsiatis et al. (1981). Nous commencons par écrire la log-vraisemblance partielle dans le modèle de Cox :

$$\begin{aligned} l^p(\beta; X, \Delta) &= \log(L^p(\beta; X, \Delta)) \\ &= \sum_{i=1}^{n} \Delta_i (Z_i^t \beta) - \log\Big(\sum_{j \in R(X_{(i)})} \exp\,(Z_j^t \beta)\Big) \end{aligned} \tag{1.8}$$

L'estimateur $\hat{\beta}$ est donc la solution de l'équation annulant la dérivée de la log-vraisemblance partielle définie dans l'équation (1.8) par rapport à β :

$$\frac{\partial l^p(\beta; X, \Delta)}{\partial \beta} = \sum_{i=1}^{n} \Delta_i Z_i^t - \frac{\sum_{j \in R(X_{(i)})} Z_j \exp\,(Z_j^t \beta)}{\sum_{j \in R(X_{(i)})} \exp\,(Z_j^t \beta)} \tag{1.9}$$

Nous explicitons maintenant les hypothèses nécessaires afin de prouver la consistance et la normalité asymptotique de $\hat{\beta}$.

(H1) $\mathbb{P}(X \geq \tau) > 0$

(H2) $\mathbb{E}[Z \exp(Z^t \beta)]^2$ est bornée uniformément dans un voisinage de β

Les temps d'événements sont supposés bornés et la quantité τ représente ici la fin de la période d'observation. L'hypothèse **(H1)** implique qu'à la fin de la période d'observation, il y a une probabilité non nulle pour qu'une observation qui n'a toujours pas subi l'événement ne soit pas censurée. C'est une condition qui est validée dans la plupart des études sur des données réelles. Sous les hypothèses **(H1)** et **(H2)**, le théorème 3.1 de Tsiatis et al. (1981) garanti l'existence d'une suite de solutions $\hat{\beta}_n$ de l'équation (1.9) tel que $\hat{\beta}_n$ converge p.s vers β_0. Ils prouvent également la normalité asymptotique de l'estimateur.

1.4.4 Relation entre les estimateurs de maximum de vraisemblance partielle et de maximum de vraisemblance non paramétrique dans le modèle de Cox

L'estimateur par maximum de vraisemblance partielle défini dans le modèle de Cox peut-être considéré comme un maximum de vraisemblance non paramétrique (NPMLE) (cf. Zeng and Lin (2007)). Nous écrivons la vraisemblance

jointe pour les paramètres β et H_0 dans le modèle de Cox :

$$L_j(\beta, H_0; X, \Delta) = \prod_{i=1}^{n} \Big(h_0(X_i)\ \exp(Z_i^t\beta)\Big)^{\Delta_i}\ \exp(-\ \exp(Z_i^t\beta)H_0(X_i)) \tag{1.10}$$

où $H_0(X_i) = \int_0^{X_i} h_0(t)dt$.

En considérant h_0 comme une fonction constante par morceaux entre les temps de survenue d'événement non-censurés, $L_j(\beta, H_0; X, \Delta)$ est maximisé simultanément par $\hat{\beta}$ défini comme le maximum de la vraisemblance partielle (cf. équation (1.7)) et l'estimateur de Breslow (cf. Zeng and Lin (2007)) :

$$\hat{H}_0(t) = \sum_{i=1}^{n} \frac{\mathbb{I}_{X_i \leq t}\Delta_i}{\sum_{j \in R(X_i)}\ \exp(Z_j^t\beta)}$$

Ainsi, l'estimateur NPMLE de β et de H_0 sont égaux à l'estimateur par maximum de vraisemblance partielle de β et l'estimateur de Breslow de H_0 respectivement.

1.5 Les modèles de fragilité

Le modèle de fragilité introduit par Vaupel et al. (1979) permet de s'affranchir de l'hypothèse de proportionnalité des risques du modèle de Cox. Ce modèle peut-être considéré comme une extension du modèle de Cox permettant de prendre en compte l'hétérogénéité qu'il peut y avoir dans les données. La notion de fragilité est un moyen pratique d'introduire des effets aléatoires, une hétérogénéité non observée ou des associations possibles dans les modèles d'analyse de survie. Dans sa forme la plus simple, une fragilité peut être considérée comme un effet aléatoire non observé qui modifie la fonction de risque instantané d'une observation ou de plusieurs observations liées les unes aux autres. Cet effet est modélisé par une variable aléatoire suivant une distribution de probabilité. Le ou les paramètres qui caractérisent cette distribution de probabilité sont également estimés avec les autres paramètres du modèle.

De nombreux modèles ayant chacun une structure de fragilité propre ont été proposés depuis Vaupel et al. (1979) et ces modèles permettent différentes modélisations. Dans cette section, nous décrivons quelques modèles fréquemment utilisés dans la littérature.

1.5.1 Modèles à fragilités univariées

Nous commençons par le modèle proposé par Vaupel et al. (1979) qui propose de gérer l'hétérogénéité présente dans les données par un effet aléatoire multiplicatif au modèle. Nous pouvons modéliser de manière équivalente l'effet aléatoire comme un effet additif dans la fonction de lien exponentiel. Ils ont introduit la notion de fragilité et l'ont appliquée à des données démographiques. Le modèle de fragilité classique qui est principalement utilisé

suppose un modèle à risques proportionnels qui est conditionnel à l'effet aléatoire (fragilité). Dans l'étude de Vaupel et al. (1979), le risque instantané d'un individu (observation) dépend en outre d'une variable aléatoire non observée, qui agit de manière multiplicative sur la fonction de risque de base. Les auteurs considèrent un modèle sans covariables et étudient les rapports de risque entre les observations. C'est un exemple de modèle univarié, du fait qu'il existe un effet aléatoire associé à chaque observation. Les covariables peuvent être naturellement incorporées au modèle pour obtenir une modélisation plus générale du risque instantané dans le modèle à fragilités univariées. On considère une population composée de n observations. Pour $1 \leq i \leq n$, le temps de survenue de l'événement et le temps de censure pour l'observation i sont modélisés par des variables aléatoires notées T_i et C_i respectivement. On observe alors pour $1 \leq i \leq n$ le temps censuré à droite et l'indicateur de censure notés respectivement X_i et Δ_i et définis par :

$$X_i = \min(T_i, C_i) \text{ et } \Delta_i = \mathbb{1}_{T_i \leq C_i}$$

Pour $1 \leq i \leq n$, le modèle s'écrit :

$$\forall t \geq 0 \;\; h_i(t|u_i) = h_0(t) u_i \exp(Z_i^t \beta) \tag{1.11}$$

où $h_i(t|u_i)$ est le risque instantané de survenue de l'événement pour l'observation i au temps X_i, $h_0(t)$ le risque de base au temps t, $\mathbf{u} = (u_i)_{1 \leq i \leq N}$ est le vecteur de fragilité, β le vecteur des paramètres de régression inconnu et Z_i les covariables associées à l'observation i.

On fait les hypothèses classiques suivantes :

(F1) Les temps de censure $(C_i)_{1 \leq i \leq n}$ sont indépendants des temps de survenue de l'événement $(X_i)_{1 \leq i \leq n}$ et des variables de fragilité $(u_i)_{1 \leq i \leq n}$.

(F2) Les temps de survenue de l'événement $(X_i)_{1 \leq i \leq n}$ sont indépendants et identiquement distribués.

(F3) Les fragilités $(u_i)_{1 \leq i \leq n}$ sont indépendantes et identiquement distribuées selon une loi de densité g paramétrée par γ.

Ce modèle est identifiable si $\mathbb{E}(u) < \infty$ et en présence de covariables (cf. Elbers and Ridder (1982)). Dans ce cas, aucune hypothèse sur la fonction de risque de base h_0 ou sur la classe de distribution de **u** est nécessaire. Nous pouvons aussi faire intervenir la fragilité de façon additive dans la fonction de lien exponentielle comme suit :

$$\forall t \geq 0 \;\; h_i(t|b_i) = h_0(t) \exp(Z_i^t \beta + b_i) \tag{1.12}$$

Les limites du modèle à fragilités univariées

Dans le modèle à fragilités univariées, nous ne supposons aucune corrélation entre les temps de survie ce qui

implique donc une population homogène. C'est la conséquence directe de l'hypothèse **(F2)**. Cela ne reflète pas toujours la réalité. Par exemple, dans un essai clinique mené sur plusieurs centres, les données collectées dans un même centre sont sûrement plus corrélées entres elles par rapport à des données collectées dans un autre centre. Cet "effet centre" doit être pris en compte lors de la modélisation de ce type de données.

1.5.2 Modèles à fragilités multivariées

En analyse de survie, la structure des données conduit souvent à des effets de groupe ou/et des corrélations fortes. Ce type de données se produit par exemple si l'on considère les durées de vie (ou les périodes d'apparition d'une maladie) de personnes d'une même famille (jumeaux, parents-enfants) ou des événements récurrents tels que des infections chez la même personne. Une première approche qui répond à la problématique d'un "effet groupe" tel que dans une étude clinique multi-centre consiste à associer à chaque groupe un effet aléatoire. Le modèle à fragilités partagées permet de prendre en compte la structure en groupe et permet d'aborder la nature multivariée des données. Ce type de modèle a cependant des limites qui seront détaillés. Des modèles plus flexibles permettent de contourner ces limites et offrent une alternative plus adaptées à certains types de données. Nous détaillons d'abord un exemple de modèle à fragilités partagées, puis un modèle à fragilités multivariées corrélées.

Modèles à fragilités partagées

Le modèle à fragilités partagées est pertinent quand les temps d'événements des observations étudiées sont liés. C'est un cas spécifique des modèles à fragilités multivariées. On suppose que les observations d'un groupe partagent la même fragilité, ce qui explique pourquoi ce modèle est appelé modèle à fragilités partagées. Il a été introduit par Clayton (1978) et plus largement étudié par Hougaard (2000).

On considère une population composée de N groupes. Pour $1 \leq i \leq N$, on note par n_i la taille du $i$ème groupe. Pour $1 \leq i \leq N$ et $1 \leq j \leq n_i$, le temps de survenue de l'événement et le temps de censure pour l'individu j du groupe i sont modélisés par des variables aléatoires notées T_{ij} et C_{ij} respectivement. On observe alors pour $1 \leq i \leq N$ et $1 \leq j \leq n_i$ le temps censuré à droite et l'indicateur de censure notés respectivement X_{ij} et Δ_{ij} et définis par :

$$X_{ij} = \min(T_{ij}, C_{ij}) \text{ et } \Delta_{ij} = \mathbb{1}_{T_{ij} \leq C_{ij}}$$

Le modèle de fragilité est défini pour $1 \leq i \leq N, 1 \leq j \leq n_i$ par :

$$\forall t \geq 0 \quad h_{ij}(t|b_i) = h_0(t) \exp(Z_{ij}^t \beta + b_i) \tag{1.13}$$

où $h_{ij}(t|b_i)$ est le risque instantané de survenue de l'événement pour l'individu j du groupe i au temps t, $h_0(t)$ le

risque de base au temps t, b_i le vecteur de fragilité du groupe i, β le vecteur des paramètres de régression inconnu et Z_{ij} les covariables associées à l'observation j du groupe i.

Nous reformulons les hypothèses faites dans le modèle à fragilités univariées pour prendre en compte la structure en groupes :

(F1) Les temps de censure $(C_{ij})_{1\leq i\leq N, 1\leq j\leq n_i}$ sont indépendants des temps de survenue de l'événement $(T_{ij})_{1\leq i\leq N, 1\leq j\leq n_i}$

(F2) Conditionnellement aux fragilités $(b_i)_{1\leq i\leq N}$, les temps de survenue de l'événement $(T_{ij})_{1\leq i\leq N, 1\leq j\leq n_i}$ sont indépendants.

(F3). Les fragilités $(b_i)_{1\leq i\leq N}$ sont indépendantes et identiquement distribuées selon une loi de densité g paramétrée par γ.

Ce modèle peut aussi être formulé avec une fragilité agissant de façon multiplicative sur la fonction de risque instantané. Dans ce cas, le modèle s'écrit pour $1 \leq i \leq N, 1 \leq j \leq n_i$:

$$\forall t \geq 0 \;\; h_{ij}(t|u_i) = h_0(t)u_i \exp(Z_{ij}^t\beta) \tag{1.14}$$

où u_i est le vecteur de fragilité du groupe i. Cette paramétrisation est courante dans la littérature, notamment dans les modèles de fragilité gamma. Cependant, modéliser la fragilité ainsi pose un problème si nous voulons considérer une covariable agissant sur le terme de fragilité par exemple.

Différents modèles à fragilités partagées peuvent être définis en utilisant différentes distributions de fragilité pour les effets aléatoires. Les paramètres de la fonction de risque de base h_0 sont estimés dans certains cas sous l'hypothèse d'une forme paramétrique de la fonction ou la fonction peut aussi être estimée de façon semi-paramétrique sur une base de splines (cf. Rondeau et al. (2012)). Nous verrons plus en détail les procédures d'estimation dans la section suivante.

Les limites du modèle à fragilités partagées

- Dans le modèle à fragilités partagées, l'effet aléatoire partagé par chaque observation d'un groupe explique la corrélation entre les membres du groupe. Cette modélisation impose les mêmes effets liés aux covariables non observées sur les observations du groupe ce qui peut être difficile à justifier pour des données réelles.
- En présence de covariables, les estimations du paramètre de la loi de fragilité et des coefficients de régression sont confondues (cf. Clayton and Cuzick (1985)).
- Dans la plupart des cas, comme avec une fragilité gamma par exemple, une fragilité unidimensionnelle ne peut qu'induire une association positive au sein du groupe.

Modèles à fragilités multivariées corrélées

La modélisation des temps d'événements de telle sorte que chaque membre d'un groupe partage la même fragilité comme dans le modèle à fragilités partagées n'est pas la meilleure solution lorsqu'il peut exister une hétérogénéité au sein du groupe. La difficulté que pose ce type de données est due à la dépendance des observations au sein des groupes, ou à des mesures répétées au sein des observations. La dépendance survient généralement lorsque les observations d'un même groupe sont liés les unes aux autres ou en raison de la récurrence de l'événement d'intérêt pour la même observation. Des modèles à fragilités multivariées corrélées ont été fréquemment utilisés pour modéliser cette dépendance dans les données (cf. Hougaard (2012)). Un premier modèle dans l'étude de données de survie de jumeaux danois a été proposé par Yashin and Iachine (1995). Les auteurs étendent le modèle à fragilités univariées et modélisent les temps de survie par un modèle bivarié à fragilités individuelles corrélées. Suivant les travaux de Yashin and Iachine (1995), pour $1 \leq i \leq N, 1 \leq j \leq 2$:

$$\forall t \geq 0 \quad h_{ij}(t|u_{ij}) = h_0(t)u_{ij} \exp(Z_{ij}^t \beta) \tag{1.15}$$

où $u_i = (u_{i1}, u_{i2})_{1 \leq i \leq N}$ sont i.i.d. au niveau des groupes mais les termes (u_{i1}, u_{i2}) sont corrélés et on note $\rho_u = \text{Corr}(u_{i1}, u_{i2})$. Ce modèle peut être considéré comme une version plus flexible du modèle à fragilités partagées. Dans le modèle à fragilités partagées, $u_{i1} = u_{i2}$ pour $i = 1, \ldots, N$ ce qui implique $\rho_u = 1$. Le modèle de Yashin and Iachine (1995) permet d'estimer le paramètre ρ_u dans l'intervalle $[0, 1]$ et d'être moins contraignant dans la modélisation des données. Des techniques similaires peuvent être appliquées afin de construire différents modèles de fragilité corrélés en fonction de la structure des données.

Les modèles de fragilité à corrélations spatiales

L'utilisation des statistiques spatiales en analyse de survie est relativement récente et s'avère cruciale dans l'analyse et la modélisation de certains types de données. Le travail de Snow sur la carte de Broad Street est considéré comme le premier travail spatial sur des données d'épidémiologie (cf. Shiode et al. (2015)). À ce jour, il existe peu de travaux sur les modèles de fragilités spatiales. Banerjee et al. (2003) ont proposé un modèle de fragilité paramétrique pour estimer les paramètres en utilisant une approche bayésienne sur des données de mortalité infantile au Minnesota. Les données considérées sont structurées en groupe et la dépendance spatiale entre les groupes est modélisée. Nous définissons un modèle de fragilité spatiale général qui peut être vu comme une extension du modèle à fragilités partagées qui prend en compte une dépendance spatiale entre les groupes. Ce modèle est dans la même veine que celui considéré dans Li and Ryan (2002). Pour $1 \leq i \leq N$, $1 \leq j \leq n_i$, on considère :

$$h_{ij}(t|b_i) = h_0(t)\ \text{exp}(Z_{ij}^t\beta + b_i) \tag{1.16}$$

où $(b_i)_{1\leq i\leq N} \sim \mathcal{N}(0_N, \Sigma(\rho))$. Le coefficient ρ est ici un coefficient de corrélation qui doit être estimé. Les modèles spatiales comme celui-ci considèrent souvent des régions comme groupe et les distances entre les groupes sont prises en compte dans le modèle. Les observations d'un même groupe partagent le même terme de fragilité et les fragilités des différents groupes sont corrélées. Par exemple, dans Li and Ryan (2002), une position (coordonnées géographiques) est mesurée par région. Dans le Chapitre 3, nous proposons un modèle spatiale qui permet de considérer les distances entre toutes les observations et nous discutons les avantages de cette modélisation par rapport à la structure en groupe en l'appliquant pour analyser des données de malaria.

1.5.3 Lois de fragilités

Nous décrivons dans cette section une liste non exhaustive de lois de fragilités couramment utilisées dans la littérature. Il existe bien sûr une plus grande palette de distributions (voir cf. Duchateau and Janssen (2008) pour plus de détails).

La fragilité gamma

La distribution gamma est l'un des choix de distribution les plus populaires et est donc très présent dans la littérature. La forme simple de la densité permet un calcul facile dans différentes approches d'estimation; que ce soit les approches de vraisemblance classiques ou quand il s'agit de trouver des expressions analytiques de la fonction de survie, du risque cumulé et de la fonction de risque instantané. Par exemple, nous verrons plus tard que l'intégrale de la vraisemblance complète par rapport à une fragilité gamma possède une forme analytique contrairement à d'autres distributions de fragilité. D'autre part, la simplicité de la transformée de Laplace s'avère également très utile dans de nombreuses applications. Il est très fréquent de considérer une distribution gamma avec une moyenne égale à 1 et d'estimer le seul paramètre qui caractérise la distribution qu'on note η. Supposons le vecteur de fragilité u qui suit une distribution gamma de moyenne égale à 1, la densité de la variable u s'écrit :

$$g(u) = \frac{u^{\eta-1}\eta^{\eta}\text{exp}(-\eta u)}{\Gamma(\eta)}$$

Comme on peut le voir dans la Figure 1.3, c'est une distribution flexible qui prend diverses formes selon la valeur du paramètre η. Le cas $\eta = 1$ correspond à la distribution exponentielle et lorsque η est grand, elle prend une forme en cloche rappelant la distribution normale. Cette loi de fragilité est favorisée dans la plupart des cas pour les facilités mathématiques et informatiques qu'elle offre. Il existe des tests d'adéquation à une loi de fragilité gamma (cf. Geerdens et al. (2012)) pour les modèles à fragilités partagées.

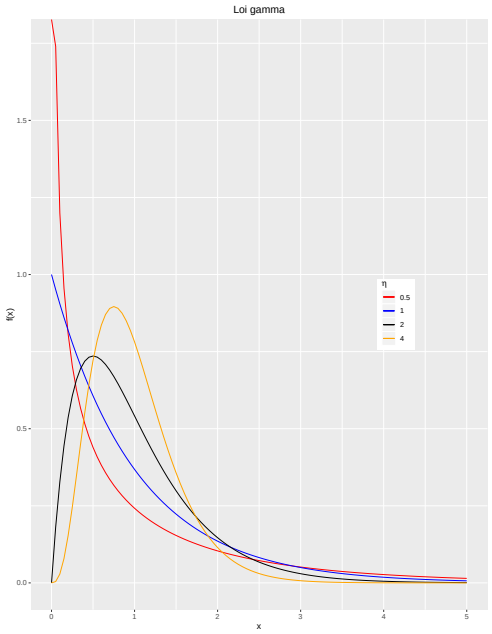

Figure 1.3: Densité de la distribution gamma pour différentes valeurs de η

La fragilité log-normale (multiplicative) ou normale (additive)

Le modèle de fragilité log-normale est défini suivant l'équation du modèle (1.11) et le modèle de fragilité normale selon le modèle (1.12). Dans les deux modèles, on impose souvent les contraintes suivantes : $\mathbb{E}(\mathbf{b}) = 0$ pour le modèle de fragilité normale et $\mathbb{E}(\mathbf{u}) = 1$ pour le modèle log-normale. Le modèle (1.12) est le plus souvent utilisé dans la littérature et c'est celui qu'on implémente dans tous les chapitres du manuscrit. Elle permet l'inclusion de covariables au niveau de la fragilité et offre donc plus de possibilités de modélisation. La fragilité log-normale est particulièrement utile pour modéliser les structures de dépendance dans les modèles de fragilité multivariés.

Cependant, il n'existe pas de forme explicite de la vraisemblance marginale pour ces deux modèles. Par conséquent, des stratégies d'estimation basées sur des approximations, intégrations numériques ou algorithme de type Expectation Maximization (EM) sont nécessaires dans une approche de maximisation de la vraisemblance marginale.

1.6 Méthodes d'estimation existantes pour les modèles de fragilité

Il est important de distinguer les différents objectifs de l'analyse de survie suivant les quantités qui nous intéressent. Nous considérons comme paramètres d'intérêts les paramètres de régression β, le paramètre de la loi de fragilité et la fonction de risque de base h_0 selon le contexte. Les deux principales méthodes comprennent l'approche paramétrique et l'approche semi-paramétrique. L'approche semi-paramétrique est plus riche car elle offre de nombreuses façons de gérer la fonction de risque de base h_0. Nous commençons par décrire l'approche paramétrique

qui consiste à faire une hypothèse paramétrique sur la fonction h_0. Nous encha$\hat{\imath}$nons ensuite avec les approches semi-paramétriques.

1.6.1 Estimation paramétrique

Dans les approches d'estimation paramétrique, nous supposons que les durées de survie suivent une certaine distribution. Habituellement, dans l'étude de données réelles, des informations préalables sur les événements que l'on considère sont prises en compte lors du choix d'une structure paramétrique pour les temps de survie. La fonction de risque de base h_0 prend alors une forme totalement paramétrique et les paramètres associés doivent être estimés. La méthode d'estimation classique dans ces modèles se fait par maximum de vraisemblance. Considérons quelques-unes des nombreuses applications des modèles de fragilité paramétrique dans la littérature. Dans de nombreux cas, l'hypothèse d'une distribution Weibull pour les temps d'événements est privilégiée (cf. Kuhn et al. (2016), Kong et al. (2010)). Dans certaines situations, la modélisation des temps d'événements est choisie en fonction de la structure des données disponibles. Une fonction de risque de base h_0 constante par morceaux s'avère utile, en particulier lorsqu'il s'agit de modéliser des effets saisonniers ou d'autres effets liés au climat comme dans les travaux de Getachew et al. (2013).

Le package R *parfm* (Munda et al. (2012)) permet d'estimer les paramètres par maximum de vraisemblance dans les modèles à fragilités partagées. Il est possible de choisir parmi une large gamme de fonctions de risque de base (Weibull, Gompertz, log-normale, etc) et de lois de fragilités dont la loi log-normale et la loi gamma.

Nous considérons deux modèles de fragilité paramétriques et nous illustrons les méthodes d'estimation dans les deux cas.

Estimation paramétrique dans le modèle à fragilités partagées gamma

Nous considérons le modèle (1.14) avec $\mathbf{u} \sim g(\eta, \frac{1}{\eta})$ où g est la densité de probabilité d'une distribution gamma de moyenne égale à 1.

$$\forall t \geq 0 \;\; h_{ij}(t|b_i) = h_0(t) u_i \, \exp(Z_{ij}^t \beta)$$

Les paramètres du modèle sont $\theta_g = (\beta, h_0, \eta)$. La structure de h_0 est souvent choisie suite à des informations apriori sur les temps de survie. La vraisemblance complète dans ce modèle s'écrit :

$$L_{\text{comp}}(\theta_g; \mathbf{X}, \mathbf{\Delta}, \mathbf{u}) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} (h_0(X_{ij}) u_i \, \exp(Z_{ij}^t \beta))^{\delta_{ij}}$$
$$\exp(-H_0(X_{ij}) u_i \, \exp(Z_{ij}^t \beta)) \times \prod_{i=1}^{N} g_\eta(u_i)$$

La log-vraisemblance marginale est obtenue en intégrant la vraisemblance complète par rapport à la fragilité $\mathbf{u}$ et en calculant ensuite le logarithme de l'expression obtenue :

$$\begin{aligned}\log L_{\text{marg}}(\theta_g; \mathbf{X}, \mathbf{\Delta}) &= \log \int L_{\text{comp}}(\theta_g; \mathbf{X}, \mathbf{\Delta}, \mathbf{u}) d\mathbf{u} \\ &= \sum_{i=1}^{N} d_i \log(\eta) - \log \Gamma\Big(\frac{1}{\eta}\Big) + \log \Gamma\Big(\frac{1}{\eta} + d_i\Big) \\ &- \Big(\frac{1}{\eta} + d_i\Big) \log\Big(1 + \eta \sum_{j=1}^{n_i} H_0(X_{ij}) \exp(Z_{ij}^t \beta)\Big) \\ &+ \sum_{j=1}^{n_i} \delta_{ij} (Z_{ij}^t \beta + \log(h_0(X_{ij}))) \end{aligned} \tag{1.17}$$

où $d_i = \sum_{j=1}^{n_i} \delta_{ij}$.

On obtient donc une expression analytique de la vraisemblance marginale. L'estimation des paramètres θ_g se fait en maximisant (1.17). La mise à jours des paramètres se fait généralement à l'aide de méthodes de descente de gradient.

Estimation paramétrique dans le modèle à fragilités partagées log-normale

Nous considérons le modèle (1.13) avec $\mathbf{b} \sim g_\eta$ où g est la densité de probabilité d'une loi normale de moyenne égale à 0 et de variance η.

$$\forall t \geq 0 \quad h_{ij}(t|b_i) = h_0(t) \exp(Z_{ij}^t \beta + b_i)$$

Les paramètres du modèle sont $\theta = (\beta, h_0, \eta)$. La vraisemblance complète dans ce modèle s'écrit :

$$\begin{aligned} L_{\text{comp}}(\theta; \mathbf{X}, \mathbf{\Delta}, \mathbf{b}) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} &(h_0(X_{ij}) \exp(Z_{ij}^t \beta + b_i))^{\delta_{ij}} \\ &\exp(-H_0(X_{ij}) \exp(Z_{ij}^t \beta + b_i)) \times \prod_{i=1}^{N} g_\eta(b_i) \end{aligned}$$

La vraisemblance marginale est obtenue en intégrant la vraisemblance complète par rapport à la fragilité $\mathbf{b}$:

$$L_{\text{marg}}(\theta; \mathbf{X}, \mathbf{\Delta}) = \int L_{\text{comp}}(\theta; \mathbf{X}, \mathbf{\Delta}, \mathbf{b}) d\mathbf{b} \tag{1.18}$$

L'estimation des paramètres θ se fait en maximisant cette vraisemblance marginale. Cependant, il n'existe pas de forme analytique de l'intégrale quand on suppose que $\mathbf{b}$ suit une loi normale. Dans ce cas de figure, il existe des méthodes d'estimation telles que l'estimation via une version stochastique de l'algorithme Expectation Maximization

(EM) (cf. Kuhn and El-Nouty (2013)) et des approximations de Laplace (cf. Munda et al. (2012)). Les méthodes impliquant une approximation de Laplace ne sont pas détaillées car elles sortent du cadre de cette thèse. Nous décrivons les algorithmes de type EM dans la section 1.7.

Propriétés asymptotiques des estimateurs paramétriques

Les propriétés asymptotiques du maximum de vraisemblance (MLE) ont été largement étudiées et établies dans dans le cadre général des modèles linéaires et nonlinéaires généralisés à effets mixtes. Wald (1949) a prouvé la consistance du MLE lorsque les observations sont indépendamment et identiquement distribuées. Les travaux de Bradley and Gart (1962) montrent la consistance faible et la normalité asymptotique du MLE lorsque les observations sont indépendantes mais ne sont pas distribuées de manière identique. Cependant, les conditions d'application des théorèmes de Bradley and Gart (1962) ou Wald (1949) ne sont pas vérifiables dans certains cas. Si l'on considère par exemple le modèle à fragilités partagées paramétrique, l'intégration sur le vecteur de fragilité pour obtenir la vraisemblance marginale rend les conditions difficiles à vérifier. Dans le cas de mesures répétées, les travaux de Nie (2006) et Nie (2007) visent à fournir des conditions moins contraignantes de consistance et de normalité asymptotique du MLE dans certains modèles à effets mixtes linéaires et nonlinéaires généralisés. Les conditions de régularité requises pour garantir les propriétés théoriques du MLE sont vérifiées dans quelques exemples simples, à savoir les modèles de régression logistique à effets mixtes et les modèles de courbe de croissance. Cependant, les résultats de Nie (2006) et Nie (2007) ne peuvent s'étendre aux modèles de fragilité car les conditions requises ne sont pas vérifiées, même dans des cas simples. À notre connaissance, il n'existe pas de résultat théorique sur les propriétés asymptotiques du MLE défini dans les modèles de fragilité avec une structure paramétrique sur la fonction de risque de base h_0.

1.6.2 Estimation semi-paramétrique

La littérature sur les approches semi-paramétriques est plus riche que sur les approches paramétriques. Si nous ne supposons aucune distribution sur les temps d'événements, il existe alors différentes manières de gérer la fonction de risque de base h_0. Un exemple d'approche semi-paramétrique consiste à estimer la fonction de risque de base de manière plus flexible sur une base de splines comme décrit dans les travaux de Rondeau et al. (2003) dans le modèle à fragilités partagées gamma. L'estimation des paramètres est alors basée sur une approche de vraisemblance pénalisée. Une autre approche fréquemment utilisée concerne l'estimation basée sur une vraisemblance partielle analogue à celle définie dans le modèle de Cox (cf. Cox (1975)) et décrite par l'équation 1.7. Suite à la définition d'une vraisemblance partielle intégrée dans le cadre du modèle à fragilités partagées (cf. Therneau (2018a)), des estimations basées sur la maximisation de la vraisemblance partielle intégrée (cf. Ha et al. (2017) et Ripatti and Palmgren (2000)) ont été proposées. Dans le chapitre 2, nous proposons une méthode d'estimation

basée sur cette vraisemblance partielle intégrée en utilisant une version stochastique de l'algorithme EM. L'utilisation de l'algorithme EM pour l'estimation des paramètres dans les modèles de fragilités remontent aux travaux de Klein (1992) où un modèle à fragilités partagées gamma est considéré. Le package R *frailtyEM* (cf. Balan and Putter (2019)) fournit une estimation des paramètres par maximum de la vraisemblance marginale dans des modèles à fragilité partagées semi-paramétriques en utilisant l'algorithme EM.

Dans cette section, nous visons à fournir une description de quelques méthodes d'estimation mentionnées dans le paragraphe précédent. Nous présentons ensuite les résultats théoriques établis sur les estimateurs.

Estimation semi-paramétrique dans le modèle à fragilités partagées gamma

Nous considérons toujours le modèle (1.14) avec $\mathbf{u} \sim g(\eta, \frac{1}{\eta})$ où g est la densité de probabilité d'une distribution gamma de moyenne égale à 1. On garde les mêmes notations pour les paramètres du modèle; $\theta_g = (\beta, h_0, \eta)$. Contrairement à l'approche paramétrique où l'on fait une hypothèse sur la distribution des temps de survie, la méthode de Rondeau et al. (2003) consiste à estimer h_0 sur une base de splines et maximiser la vraisemblance marginale définie dans l'équation (1.17) plus un critère de pénalité qui dépend de h_0. La log-vraisemblance pénalisée s'écrit :

$$pl(\theta_g; \mathbf{X}, \mathbf{\Delta}) = \log L_{\text{marg}}(\theta_g; \mathbf{X}, \mathbf{\Delta}) - \kappa \int_0^\infty h_0''^2(t)dt \tag{1.19}$$

où le paramètre κ est un paramètre de lissage positif qui agit sur la norme quadratique de la dérivée seconde de h_0. L'équation (1.19) représente un compromis entre l'ajustement aux données (contribution du terme $\log L_{\text{marg}}(\theta_g; \mathbf{X}, \mathbf{\Delta})$, et la régularité de la solution (contribution du terme $\int_0^\infty h_0''^2(t)dt$). La fonction h_0 est approximée par une combinaison linéaire de M-splines ($h_0(.) = \sum_{i=1}^m \tilde{\alpha}_i M_i(.)$). Les paramètres sont estimés en maximisant la quantité $pl(\theta_g; \mathbf{X}, \mathbf{\Delta})$ par l'algorithme de Marquardt qui est une combinaison de l'algorithme de Newton-Raphson et un algorithme de descente de gradient.

Le package R *frailtypack* (cf. Rondeau et al. (2012)) a été développé pour implémenter cette méthode d'estimation. Le package est régulièrement mis à jour pour inclure davantage de modèles tels que le modèle conjoint et les modèles de fragilité hiérarchiques. Il est aussi possible d'estimer les paramètres suivant le modèle de fragilité log-normale dans les versions plus récentes.

Estimation semi-paramétrique basée sur des linéarisations dans le modèle à fragilités partagées log-normale

Parmi les nombreuses autres méthodes d'estimation semi-paramétriques, une autre approche consiste à définir une vraisemblance partielle suivant les travaux de Cox (1975) qui nous permet de nous affranchir de la fonction de risque de base h_0. Il n'est donc plus nécessaire de faire un choix de paramétrisation de la fonction h_0 et d'estimer uniquement les paramètres de régression et de la loi de fragilité. Nous présentons deux méthodes qui diffèrent dans la méthode de maximisation de cette vraisemblance partielle intégrée par rapport aux fragilités. Une comparaison de ces deux méthodes avec notre nouvelle méthode d'estimation sera présentée dans le Chapitre 2.

En gardant la même structure de groupe et les mêmes notations, nous rappelons le modèle de fragilité log-normale défini pour $1 \leq i \leq N, 1 \leq j \leq n_i$ par :

$$\forall t \geq 0 \quad h_{ij}(t|b_i) = h_0(t) \exp(Z_{ij}^t \beta + b_i) \tag{1.20}$$

où b_i est le vecteur de fragilité du groupe i et on suppose que $\mathbf{b} = (b_i)_{1 \leq i \leq N}$ suit une gaussienne centrée de variance γ. La vraisemblance partielle complète dans ce modèle s'écrit :

$$\begin{aligned} L^p(\theta; \mathbf{X}, \mathbf{\Delta}, \mathbf{b}) &= \prod_{i=1}^{N} \prod_{j=1}^{n_i} \left(\frac{\exp(Z_{ij}^t \beta + W_{ij}^t b_i)}{\sum_{(l,k) \in R(X_{(ij)})} \exp(Z_{lk}^t \beta + W_{lk}^t b_l)} \right)^{\Delta_{ij}} \\ &\quad \times \prod_{i=1}^{N} g_\gamma(b_i) \end{aligned} \tag{1.21}$$

Cette vraisemblance partielle ne fait plus intervenir la fonction de risque de base h_0 et les paramètres à estimer sont $\theta = (\beta, \gamma)$. Suivant les travaux de Therneau (2018a), nous définissons la vraisemblance partielle intégrée qui s'obtient en intégrant la vraisemblance partielle complète calculée dans l'équation (1.21) par rapport au vecteur de fragilité $\mathbf{b}$.

$$L^p_{\text{marg}}(\theta; \mathbf{X}, \mathbf{\Delta}) = \int L^p(\theta; \mathbf{X}, \mathbf{\Delta}, \mathbf{b}) d\mathbf{b} \tag{1.22}$$

Les méthodes d'estimation implémentées dans le package R *frailtyHL* sont basées sur une approximation de Laplace de cette vraisemblance partielle intégrée qui est ensuite maximisée. Il est possible de spécifier l'ordre de l'approximation et le choix se porte donc entre coût computationnel et qualité d'estimation. Nous renvoyons à Do Ha et al. (2012) pour plus de détails sur les approximations.

L'approche utilisée dans le package R *coxme* repose sur la maximisation de la vraisemblance partielle intégrée en remplaçant le corps de l'intégrale de l'équation (1.22) par une série de Taylor du deuxième ordre. Cette approximation donne alors une intégrale qui peut être résolue analytiquement.

Nous comparons ces deux méthodes à une nouvelle méthode d'estimation basée sur un algorithme SAEM-MCMC présentée dans le Chapitre 2.

Propriétés asymptotiques des estimateurs semi-paramétriques

Nous avons présenté dans la Section 1.4.3 les propriétés asymptotiques de l'estimateur du maximum de vraisemblance partielle dans le modèle de Cox. Cependant, les estimateurs définis dans les différents modèles de fragilités présentés et les extensions de ces modèles n'ont pas tous fait l'objet de travaux théoriques. Nous mentionnons dans cette section les modèles de fragilités et les estimateurs correspondant pour lesquels des propriétés asymptotiques ont été établies. Nous précisons que l'égalité des estimateurs du paramètre de régression β par le NPMLE et

la vraisemblance partielle détaillée dans la section 1.4.4 dans le modèle de Cox s'étend naturellement aux modèles de fragilité (cf. Zeng and Lin (2007)).

En s'inspirant de l'approche par processus de comptage des travaux de Nielsen et al. (1992), un estimateur non paramétrique du maximum de vraisemblance (NPMLE) dans le modèle à fragilités partagées gamma est étudié par Murphy (1994). Le modèle défini dans l'article ne comprend pas de covariables, mais il est mentionné que la preuve devrait s'étendre à un modèle incorporant des covariables tant qu'il n'y a pas de colinéarité dans la matrice des covariables. Sous de faibles hypothèses de régularité, la consistance de l'estimateur du paramètre de la loi de fragilité gamma ainsi que la consistance de l'estimateur du risque cumulé sont démontrées. Suite à ce travail, la normalité asymptotique des estimateurs est établie (cf. Murphy (1995)).

Dans Parner (1998), les résultats asymptotiques sur le NPMLE sont étendus au modèle de fragilité gamma corrélée et des covariables sont incluses dans le modèle. Des hypothèses supplémentaires sont faites notamment sur les covariables et sur les paramètres de la loi de fragilité gamma. Une approche par processus de comptage est également utilisée dans l'élaboration de la preuve.

Le modèle de fragilité normale avec des covariables sur les fragilités est considéré par Gamst et al. (2009). En incorporant des covariables sur le vecteur de fragilité qu'on note W_{ij} dans le modèle (1.13), on obtient pour $1 \leq i \leq N, 1 \leq j \leq n_i$:

$$\forall t \geq 0 \quad h_{ij}(t|b_i) = h_0(t) \exp(Z_{ij}^t \beta + W_{ij} b_i) \tag{1.23}$$

Une autre différence notable par rapport au modèle (1.13) est la nature multivariée de $\mathbf{b}$. On considère $\mathbf{b} \sim \mathcal{N}(0_N, \Sigma)$. La log-vraisemblance pour le groupe i s'écrit :

$$\begin{aligned} l_i(\theta; X_i, b_i) = \sum_{j=1}^{n_i} \big\{ \delta_{ij} [\log(H\{X_{ij}\} + Z_{ij}^t \beta + W_{ij}^t b_i] - \\ H(X_{ij}) \exp(Z_{ij}^t \beta + W_{ij}^t b_i) \big\} + g_\Sigma(b_i) \end{aligned} \tag{1.24}$$

où $\theta = (\beta, \Sigma, h)$, $H\{t\}$ est la taille du saut en H à l'instant t et g est la densité d'une loi normale multivariée. Soit $\hat{\theta}_n = (\hat{\beta}_n, \hat{\Sigma}_n, \hat{H}_n)$ défini comme le NPMLE de θ. On note ici que la fonction $h_0(.)$ est remplacée par la fonction $H\{.\}$ dans la log-vraisemblance. Cette légère modification dans la fonction de vraisemblance est nécessaire dans l'élaboration de la preuve. Nous énonçons les hypothèses suivantes :

(L1) Conditionnellement aux covariables Z_{ij} et W_{ij}, les temps de censure C_{ij} sont indépendants des temps d'événements T_{ij} et des fragilités b_i.

(L2) Il existe $\epsilon > 0$ tel que $\mathbb{P}(C_{ij} \geq \tau | Z_{ij}, W_{ij}) \geq \epsilon$ p.s.

(L3) La fonction de risque de base h_0 est positive et continue sur l'intervalle de temps considéré $[0, \tau]$.

(L4) Les covariables Z_{ij} et W_{ij} sont bornées.

(L5) Les vrais paramètres β et Σ sont des éléments de l'intérieur d'un ensemble compact connu noté $\mathcal{K} = \{(\beta, \Sigma) : |\beta| \leq B,$ pour une constante B, et Σ est symétrique et définie positive, avec des valeurs propres bornées loin de 0 et $\infty\}$.

(L6) Les variables $(X_i, \Delta_i, Z_i, W_i)_{1 \leq i \leq N}$ sont i.i.d et $\mathbb{P}(n_i \geq 2) > 0$.

(L7) S'il existe un vecteur $\mathbf{c}$ et une matrice symétrique $\mathbf{S}$, tel que pour $k \neq j = 1, \ldots, n_i,\ \mathbf{c}^t[1, Z_{ij}^t]^t + W_{ij}^t \mathbf{S} W_{ik} = 0$ p.s, alors $\mathbf{c} = 0$ et $\mathbf{S} = 0$.

L'hypothèse **(L1)** est classique et faite dans tous les modèles considérés et l'hypothèse **(L2)** implique l'observation de suffisamment de temps d'événement dans l'intervalle $[0, \tau]$ considéré. Les hypothèses **(L4)** et **(L5)** sont plus techniques et sont nécessaires afin de garantir que certaines quantités restent positives. Si suffisamment de groupes ont des tailles d'effectifs plus grand que 1, on peut différencier entre b_i et β, ce qui permet de construire des estimations de Σ. Finalement, l'hypothèse **(L7)** est une extension de la condition 2(g) de Parner (1998), qui est nécessaire pour garantir la non-colinéarité des covariables Z_{ij} et W_{ij}. Ces hypothèses sont exactement les hypothèses **(C1)-(C7)** de Gamst et al. (2009) et suivant ces hypothèses, nous énonçons les théorèmes suivant (cf. Gamst et al. (2009)) :

Theorem 1 *Supposons vérifiées les hypothèses **(L1)-(L7)**, alors* $\left\|\hat{\beta}_n - \beta_n\right\| \to 0$, $\left\|\hat{\Sigma}_n - \Sigma_n\right\| \to 0$, *et* $\sup_{t \in [0,\tau]} \left\|\hat{H}_n(t) - H_0(t)\right\| \to 0$ *presque sûrement quand* $n \to \infty$, *où* $\|.\|$ *est la norme euclidienne.*

Theorem 2 *Supposons vérifiées les hypothèses **(L1)-(L7)**, alors* $\sqrt{n}(\hat{\theta}_n - \theta_0)$ *converge en loi vers un processus gaussien de moyenne nulle. Par ailleurs,* $\hat{\theta}_n$ *est asymptotiquement efficace.*

1.7 L'algorithme Expectation Maximization et ses variantes

Dans cette section, nous détaillons l'algorithme Expectation Maximization (EM) développé par Dempster et al. (1977) et ses différentes extensions. L'algorithme est fréquemment utilisé pour l'estimation des paramètres dans les modèles à variables latentes. Les modèles de fragilités sont des modèles à variables latentes et les algorithmes de type EM peuvent donc être utilisés pour estimer les paramètres dans ces modèles. Notamment, le package R *frailtyEM* (cf. Balan and Putter (2019)) implémente l'algorithme EM pour estimer les paramètres des modèles de fragilités semi-paramétriques par maximum de vraisemblance. Une version par approximation stochastique de l'EM est mise en œuvre par Kuhn and El-Nouty (2013) pour estimer les paramètres dans un modèle à fragilités partagées paramétrique.

1.7.1 L'algorithme Expectation Maximization

L'algorithme EM proposé par Dempster et al. (1977) est un algorithme itératif proposé pour estimer les paramètres du maximum de vraisemblance d'un modèle statistique qui dépend de variables latentes non observées. En d'autres termes, soit des valeurs manquantes existent dans les données, soit le modèle peut être formulé plus simplement en supposant l'existence de points de données supplémentaires non observés. On commence par décrire l'EM proposé par Dempster et al. (1977) dans un cadre général. Soit un modèle statistique qui génère un ensemble de données observées qu'on note par la variable $\mathbf{X}$, un ensemble de données non-observées ou latentes qu'on note par la variable $\mathbf{b}$ et un vecteur de paramètres inconnu qu'on note par $\theta \in \Theta$. L'estimateur du maximum de vraisemblance (MLE) des paramètres inconnus est déterminé en maximisant la vraisemblance marginale des données observées :

$$L_{\text{marg}}(\theta; \mathbf{X}) = \int L(\theta; \mathbf{X}, \mathbf{b})\, d\mathbf{b}$$

où $L(\theta; \mathbf{X}, \mathbf{b})$ est la vraisemblance complète des données. Cependant, cette vraisembalnce marginale n'a souvent pas de forme analytique et s'avère compliqué à calculer directement. L'algorithme EM consiste à estimer le maximum de la vraisemblance marginale en appliquant itérativement deux étapes; une étape dite Expectation (E) et une étape dite Maximization (M). À l'itération k de l'algorithme :

Etape E: Calcul de $Q(\theta|\theta_k) = \mathbb{E}_{\mathbf{b}|\mathbf{X},\theta_k}\left[\log L(\theta; \mathbf{X}, \mathbf{b})\right]$

où $Q(\theta|\theta_k)$ est l'espérance de la log-vraisemblance du paramètre θ, par rapport à la distribution conditionnelle de $\mathbf{b}$ sachant $\mathbf{X}$ et l'estimation courante des paramètres notée θ_k

Etape M: Maximization de la quantité $Q(\theta|\theta_k)$

$\theta_k = \arg\max_{\theta} Q(\theta|\theta_k)$

L'algorithme est initialisé avec des valeurs θ_0 et Q_0 arbitraires.

Résultat de convergence de l'algorithme EM

Dans le cadre des modèles appartenant à la famille exponentielle, Delyon et al. (1999) établissent un résultat de convergence de l'algorithme EM. On définit $\mathcal{L}$ l'ensemble des points stationnaires de la log vraisemblance $l = \log L_{\text{marg}}$ tel que $\mathcal{L} = \{\theta \in \Theta, \partial_\theta l(\theta) = 0\}$ et $d(x, \mathcal{A})$ la distance euclidienne entre un point x et l'ensemble $\mathcal{A}$. Le résultat de convergence de la suite $(\theta_k)_k$ générée par l'algorithme est donné par le théorème suivant :

Theorem 3 *Supposons vérifiées les hypothèses **(M1)-(M5)** (cf. Appendix C) de Delyon et al. (1999). Supposons aussi que l'adhérence de l'ensemble $\mathcal{L}$ soit un sous-ensemble compact de Θ. Alors pour tout point initial θ_0 dans Θ, la suite $(l(\theta_k))_k$ est croissante et $\lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0$*

1.7.2 L'algorithme Stochastic Approximation Expectation Maximization

L'algorithme EM décrit précédemment suppose une expression analytique de la quantité $\mathbb{E}_{\mathbf{b}|\mathbf{X},\theta_k}[\log L(\theta;\mathbf{X},\mathbf{b})]$ dans l'étape E, ce qui n'est pas forcément le cas dans certains modèles. La version d'approximation stochastique de l'EM (SAEM) proposée par Delyon et al. (1999) est une alternative puissante à l'EM lorsque l'étape E est difficile à calculer directement. C'est une extension du Monte Carlo EM introduit par Wei and Tanner (1990) qui propose de calculer l'intégrale à l'étape E à l'aide d'une méthode Monte Carlo. L'algorithme SAEM consiste à décomposer l'étape E en une étape de simulation et une étape d'intégration. Dans l'étape de simulation (étape S), on génère des réalisations de la variable latente selon la distribution a posteriori $p(\mathbf{b}|\theta_k)$ et l'étape d'approximation stochastique consiste à calculer $Q_k(\theta)$ par une approximation stochastique tout en approchant $\mathbb{E}_{\mathbf{b}|\mathbf{X},\theta_k}[\log L(\theta;\mathbf{X},\mathbf{b})]$ par une intégration de Monte Carlo suivant les travaux de Wei and Tanner (1990). L'étape M reste inchangée. L'algorithme à l'itération k s'écrit ainsi :

Etape S: Simulation de $M(k)$ réalisations $(\mathbf{b}_k(m), 1 \leq m \leq M(k))$ selon la distribution a posteriori $p(\mathbf{b}|\theta_k)$

Etape A: Calcul de la quantité $Q_k(\theta)$ par l'approximation stochastique suivante :

$Q_k(\theta) = Q_{k-1}(\theta) + \mu_k\Big(\frac{1}{M(k)}\sum_{m=1}^{M(k)} \log L(\theta;\mathbf{X},\mathbf{b}_k(m)) - Q_{k-1}(\theta)\Big)$

où $(\mu_k)_k$ est une suite de pas positifs, décroissante et convergeant vers 0.

Etape M: Maximisation de la quantité $Q_k(\theta)$

$\theta_k = \arg\max_{\theta} Q_k(\theta)$

Résultat de convergence de l'algorithme SAEM

Suivant les hypothèses de régularités **(M1)-(M5)** et des conditions liées à l'étape d'approximation stochastique (cf. Appendix C), la convergence presque sûre de la suite $(\theta_k)_k$ générée par l'algorithme converge vers un point critique de la vraisemblance marginale est établie par le Théorème 5 de Delyon et al. (1999).

1.7.3 Couplage d'une méthode de Monte Carlo Markov Chain avec l'algorithme SAEM

Cependant, pour beaucoup de modèles non linéaires, les données non-observées $\mathbf{b}$ ne peuvent pas être simulées exactement selon la distribution conditionnelle. Une alternative consiste à utiliser une méthode de Monte Carlo Markov Chain (cf. Robert and Casella (2013)), l'idée étant d'introduire une probabilité de transition Π_θ qui a comme unique distribution stationnaire la distribution conditionnelle selon laquelle nous voulons simuler (cf. Kuhn and Lavielle (2004)). On utilise alors le noyau de transition Π_{θ_k} pour simuler les données non-observées à l'itération k. L'algorithme de Hastings Metropolis (cf. Robert and Casella (2013)) permet de construire de tels noyaux de transition dans des cas très généraux, permettant ainsi une application de cet algorithme dans une grande variétés de modèles. L'étape M reste inchangée. Nous fixons $M(k) = 1$ dans la suite pour plus de simplicité. L'algorithme à l'itération k s'écrit ainsi :

Etape S: Simulation d'une réalisation $\mathbf{b}_k$ suivant la probabilité de transition d'une chaîne de Markov convergente Π_{θ_k} ayant la distribution conditionnelle aux observations $p(.|\theta_k)$ comme unique distribution stationnaire.

Etape A: Calcul de la quantité $Q_k(\theta)$ par l'approximation stochastique suivante :

$$Q_k(\theta) = Q_{k-1}(\theta) + \mu_k\Big(\log L(\theta; \mathbf{X}, \mathbf{b}_k) - Q_{k-1}(\theta)\Big)$$

Etape M: Maximisation de la quantité $Q_k(\theta)$

$$\theta_k = \arg\max_{\theta} Q_k(\theta)$$

Résultat de convergence de l'algorithme SAEM-MCMC

Theorem 4 *Supposons vérifiées les hypothèses* ***(M1)-(M5)****,* ***(SAEM1')****,****(SAEM2)****,****(SAEM3)*** *et* ***(C)*** *(cf. Appendix C) de Kuhn and Lavielle (2004). Alors,* $\lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0$ *avec probabilité 1.*

Ainsi, la suite $(\theta_k)_k$ générée par l'algorithme SAEM-MCMC converge presque sûrement vers un point critique de la log-vraisemblance marginale.

1.8 Les contributions de la thèse

Dans cette thèse, nous nous intéressons à trois aspects des modèles de fragilité à savoir la modélisation, les méthodes d'estimation et les propriétés de convergence des estimateurs. Nous proposons une contribution dans chacun des trois domaines mentionnés ci-dessus. Dans un premier temps, nous proposons une nouvelle méthode d'estimation basée sur la vraisemblance partielle intégrée sans faire d'approximation sur cette vraisemblance. Nous mettons en œuvre une variante de l'algorithme SAEM-MCMC et nous démontrons également les propriétés théoriques de convergence de l'algorithme. La deuxième contribution concerne l'étude des propriétés de convergence, en particulier les vitesses de convergence, de l'estimateur du maximum de vraisemblance dans les modèles à fragilités partagées paramétriques. Finalement, la troisième contribution présente une nouvelle modélisation qui permet de prendre en compte les corrélations spatiales au niveau individuel à travers un modèle de fragilité multivarié à corrélations spatiales. Cette nouvelle modélisation spatiale a été motivée par des données de malaria en Éthiopie. Une méthode d'estimation adaptée à ce contexte a été proposée.

1.8.1 Algorithme convergent pour l'estimation dans des modèles de fragilité multivariés par maximum de vraisemblance partielle intégrée

Dans le Chapitre 1, une nouvelle méthode d'estimation est proposée pour estimer les paramètres en maximisant la vraisemblance partielle intégrée via un algorithme SAEM-MCMC (cf. Kuhn and Lavielle (2004)). La procédure d'estimation stochastique proposée peut s'appliquer à des modèles de fragilité avec un large choix de distributions pour les fragilités et s'adapte à tout type de structure de corrélation entre les composantes de fragilité. Elle permet

aussi d'inclure des termes d'interaction aléatoire entre les covariables et les composantes de fragilité. La convergence presque sûre de l'algorithme d'estimation stochastique vers un point critique de la vraisemblance partielle intégrée est démontrée. Les propriétés de convergence numérique sont évaluées par des études de simulation et une comparaison avec deux méthodes existantes (*coxme* (cf. Therneau (2018a)) et *frailtyHL* (cf. Ha et al. (2017))) est effectuée. En particulier, la robustesse de la méthode par rapport à une mauvaise spécification de la loi de fragilité est illustrée à travers une étude de simulation. L'avantage de ne pas avoir à faire de choix de modélisation sur le risque de base h_0 est mis en évidence en comparant cet estimateur à un estimateur paramétrique par une étude de simulation. Finalement, la méthode est mise en oeuvre pour l'estimation des paramètres à partir de deux jeux de données réelles; un jeu de données d'infection de pis des vaches (mastitis) (cf. Kuhn et al. (2016)) et des données de cancer de la prostate (cf. Sylvester et al. (2006)). Ces travaux ont fait l'objet d'un article actuellement soumis. Une version en ligne de l'article est disponible sur arXiv (cf. Oodally et al. (2019)).

1.8.2 Etude des propriétés de convergence des estimateurs du maximum de vraisemblance dans le modèle paramétrique à fragilités partagées

Dans la section 1.6.2, nous décrivons les résultats théoriques obtenus sur les estimateurs dans différents types de modèles de fragilité. Cependant, à notre connaissance, il n'existe aucune étude sur les propriétés de convergence des estimateurs des paramètres dans les modèles de fragilité paramétriques. En s'inspirant des travaux de Nie (2007) dans les modèles à effets mixtes linéaires généralisés et non-linéaires, nous illustrons à travers une étude de simulation la consistance et les différentes vitesses de convergence des estimateurs par maximum de vraisemblance (MLE) des paramètres dans un modèle à fragilités partagées. Nous mettons en évidence à travers une étude de simulation que les estimateurs du maximum de vraisemblance dans les modèles de fragilité ont des comportements asymptotiques possiblement différents. Dans les cas où le nombre de groupes N et le nombre d'observations par groupe J tendent vers l'infini, nous distinguons notamment des vitesses différentes selon le paramétrage de la vraisemblance conditionnelle, plus précisément selon la présence d'une fragilité additive sur les paramètres ou pas. Nous nous intéressons aussi aux effets liés à la structure des covariables. Notamment, nous distinguons différentes vitesses de convergence du MLE des paramètres de régression β selon la nature des covariables associées, en particulier si la covariable est au niveau de l'observation ou au niveau du groupe. Considérons une covariable Z_{ij} pour des données structurées en groupes, $i = 1, \ldots, N$ et $j = 1, \ldots, J$ avec J le nombre d'observations dans chaque groupe. Considérons tout d'abord une covariable au niveau du groupe, i.e $Z_{ij} = \bar{Z}_i$ pour tout j. Ce type de covariable apporte a priori moins d'information par rapport à une covariable qui varie au niveau de l'observation. Intuitivement, les estimateurs des paramètres de régression associés à ces deux types de covariable ne convergent pas forcément à la même vitesse. On met dans un premier temps en évidence de façon théorique ce phénomène dans un modèle linéaire à effets mixtes. Finalement, nous illustrons ce phénomène par une étude de simulation

intensive sur des modèles paramétriques à fragilités partagées.

1.8.3 Estimation dans un modèle de fragilité à corrélations spatiales : application pour l'analyse de données de malaria

Faisant suite aux travaux de Li and Ryan (2002) sur un modèle de fragilité spatial, nous proposons un modèle de fragilités à corrélations spatiales général défini par l'équation (1.16) de la section 1.5.2 pour analyser des données de malaria. Les données de malaria sont intrinsèquement spatiales car collectées dans différents lieux. De plus, la maladie se propage d'hôte à hôte. La transmission se fait via un moustique, dont le développement et la reproduction sont favorisés par la présence de plans d'eau. Afin d'étudier l'influence des plans d'eau sur le taux de transmission de la malaria, un riche jeu de données a été constitué dans le secteur du barrage hydroélectrique de Gilgel Gibe dans le sud-ouest de l'Éthiopie (cf. Yewhalaw et al. (2013)). Il est constitué de temps d'infection par la malaria d'enfants répartis en villages, ainsi que de nombreuses covariables. Ces données ont déjà été analysées par un modèle de fragilité structuré selon les villages, incluant la distance entre l'enfant et le barrage comme covariable. Cependant, la proximité entre les enfants n'est pas prise en compte. Le modèle de fragilité à corrélations spatiales proposé permet de prendre en compte cette spécificité importante des données. Les paramètres du modèle sont estimés en maximisant la vraisemblance marginale via l'algorithme SAEM-MCMC. La performance de l'estimateur est évaluée sur des données simulées. La méthode est ensuite mise en œuvre pour analyser les données de malaria. Différents modèles incluant différentes modélisations du risque de base et de la structure de corrélation spatiale sont comparés. Ce travail sera soumis pour publication sous peu.

1.9 Résultats et conclusion de la thèse

Une nouvelle méthode d'estimation est proposée pour modéliser les données de survie corrélées. Cette méthode possède plusieurs atouts. Premièrement, on s'affranchit d'un choix de modélisation du risque de base qui s'avère pertinent comme démontré dans l'étude de simulation. Deuxièmement, plusieurs structures de corrélation peuvent être prisent en compte, que ce soit dans les modèles à fragilités partagées, les modèles de fragilité corrélés ou même des modèles où les effets aléatoires interagissent avec une covariable. La méthode prend également en compte les distributions de fragilité les plus fréquemment utilisées telles que la loi Gamma, la loi stable, et la loi gaussienne. Troisièmement, nous prouvons la convergence presque sûre de l'algorithme d'estimation sous des hypothèses classiques, ce qui n'est pas le cas pour la plupart des méthodes de fragilité existantes. En effet, nous définissons un modèle étendu en supposant que le paramètre de régression est une variable aléatoire et suit une distribution gaussienne afin d'être dans le cadre des modèles de la famille exponentielle dans le but de prouver la convergence presque sûre de l'algorithme. La maximisation de la vraisemblance partielle intégrée dans le modèle

étendu est équivalent à la maximisation de la vraisemblance partielle intégrée dans le modèle de fragilité classique, conduisant à de bonnes estimations des paramètres.
Une suite logique de ce travail consisterait à développer un package R permettant d'implémenter cette nouvelle méthode d'estimation.

La deuxième contribution de cette thèse porte sur l'étude des taux de convergence du maximum de vraisemblance (MLE) des paramètres dans les modèles à fragilités partagées paramétriques. Nous proposons une conjecture basée sur la paramétrisation de la vraisemblance conditionnelle ainsi que sur la structure des covariables. Premièrement, nous établissons théoriquement les différents taux de convergence des MLE dans un modèle linéaire à effets mixtes en fonction de la structure des covariables. Cependant, ce type de calcul ne peut pas être effectué dans des modèles de fragilité, dont les vraisemblances ont des expressions analytiques complexes impliquant en particulier des intégrales. Par conséquent, nous menons une étude de simulation intensive sur un modèle à fragilités partagées Weibull. Le cadre de simulation est mis en place de manière à tester différents scénarios notamment le paramétrage de la vraisemblance conditionnelle et la structure des covariables. Dans le cadre des modèles à fragilités partagées paramétriques, la conjecture suivante peut être formulée: (1) Les MLE des paramètres de la loi de fragilité sont $\sqrt{N}$-consistant. (2) Les MLE des paramètres du risque de base et les paramètres de régression impliqués dans la vraisemblance conditionnelle avec un terme de fragilité additif sont $\sqrt{N}$-consistant. (3) Les MLE des paramètres de régression qui ne sont pas associés à un terme de fragilité additif dans la vraisemblance conditionnelle et avec des covariables associées au niveau du groupe sont $\sqrt{N}$-consistant. (4) Tous les autres paramètres sont $\sqrt{NJ}$-consistant.
Il existe de nombreuses perspectives à ce travail. En particulier, suivant l'étude de simulation prometteuse sur le modèle à fragilités partagées Weibull, il serait intéressant d'étendre l'étude de simulation à d'autres modèles de fragilité. Ensuite, la prochaine étape logique consisterait à établir une preuve théorique de la conjecture proposée.

On propose une troisième contribution sous la forme d'un modèle de fragilité à corrélations spatiales. Le modèle développé est inspiré des données de malaria en Ethiopie. Les temps d'infection par la malaria de 2037 enfants vivant dans des villages situés à proximité d'un barrage hydroélectrique sont analysés. Le caractère spatial des données découle du fait que les enfants vivent à proximité les uns des autres. Par conséquent, nous considérons des corrélations au niveau de l'enfant plutôt qu'au niveau du village comme fait dans les analyses précédentes. Les paramètres sont estimés via un algorithme Expectation Maximization stochastique et les propriétés théoriques de l'algorithme sont établies. Nous considérons quatre modèles suivant des structures de corrélation et des fonctions de risque de base différentes tout en tenant compte des effets saisonniers. Les modèles sont comparés via un critère basé sur la vraisemblance pour sélectionner celui qui s'ajuste le mieux aux données. Nous considérons des

covariables à savoir le sexe, l'âge, la structure du toit et la distance au barrage et nous quantifions leurs effets sur l'incidence de la malaria. Cependant, nous concluons que les paramètres ne sont pas significatifs suite à des tests de rapport de vraisemblance. Par ailleurs, l'interprétation des estimations des paramètres associés à la structure de corrélation semble être conforme à la réalité biologique. En particulier, les estimations qui caractérisent la structure de corrélation indiquent une faible corrélation entre enfants au-delà d'une distance de 2 km. La distance de 2 km est d'ailleurs la distance maximale théorique parcourue par le vecteur de la maladie, le moustique.
Le jeu de données combiné au modèle de fragilité à corrélations spatiales que nous avons développé offre de nombreuses possibilités pour de futurs travaux de recherche. Dans notre étude, nous avons considéré quatre modèles avec deux fonctions de risque de base et deux structures de corrélation. Nous pourrions envisager davantage de modèles dans le but de mieux ajuster les données et de proposer des nouvelles informations sur l'incidence de la malaria dans la région. En particulier, il serait intéressant de prendre en compte la densité des moustiques dans le modèle. Au-delà de la problématique de la malaria, le modèle de fragilité spatiale que nous avons développé a le potentiel pour des applications plus larges. Avec la popularité croissante des données géographiques dans le domaine de l'épidémiologie en général, les modèles qui tiennent compte des corrélations spatiales peuvent offrir des informations plus approfondies sur la modélisation et l'interprétation du caractère spatial des maladies. Le développement d'applications mobiles pour limiter le risque de propagation de la récente épidémie de COVID-19 est un exemple d'une telle application.

Enfin, il serait intéressant d'étendre l'ensemble des contributions de cette thèse au contexte des modèles de fragilité impliquant des covariables de grande dimension. En effet, de nos jours, davantage d'informations sur les données sont collectées, en particulier des informations génétiques qui impliquent généralement des covariables de grande dimension. Il est essentiel d'inclure ces informations dans la tâche de modélisation pour effectuer des analyses de données pertinentes. Cependant, cela nécessite le développement de nouveaux outils d'estimation adaptés, tels que l'estimation pénalisée et la sélection de variables, conduisant à de nouveaux défis tant du point de vue numérique que théorique.

Chapter 2

Convergent stochastic algorithm for estimation in general multivariate correlated frailty models using integrated partial likelihood

2.1 Introduction

Survival analysis deals with time to event data. The Cox model introduced by Cox (1972) is most frequently used. It allows to model the instantaneous event rate, also called hazard, as the product of a baseline hazard function and a function of the covariates. The regression coefficients are estimated by maximisation of the partial likelihood function which does not depend on the baseline hazard function. The good asymptotic properties of the estimator, namely the consistency, asymptotic normality and efficiency, based on partial likelihood are detailed and proved in Andersen and Gill (1982). However, the Cox model assumes independence between event times, which is often not the case, e.g., in a clinical trial, the event times are clustered in hospitals. Frailty models introduced by Vaupel et al. (1979) accommodate for such correlation through non observed random effects. For more details on frailty models, we refer to Duchateau and Janssen (2008), Wienke (2010).

The literature on parameter estimation in frailty models is quite rich. Maximum likelihood estimation based on an Expectation Maximization (EM) algorithm has been studied by Nielsen et al. (1992) with frailties following a Gamma distribution in both non and semi-parametric models. The asymptotic properties of the maximum likelihood estimates with a plug-in estimator for the baseline hazard for a Gamma frailty model without covariates have been studied by

Murphy (1994), Murphy (1995) and for a correlated Gamma frailty distribution by Parner (1998). The choice of the Gamma distribution is often motivated by its mathematical convenience, as a closed form of the marginal likelihood can be calculated when the frailties are assumed to follow a Gamma distribution.

An approach based on the maximization of a penalized partial likelihood with a Laplace approximation of the marginal likelihood has been proposed by Ripatti and Palmgren (2000). Also, Duchateau and Janssen (2004) implemented an approach using an iterative algorithm based on the marginal and penalized likelihoods. A semi-parametric approach where the baseline hazard is estimated with a splines basis in the Gamma frailty model is implemented in the R package *frailtypack* developed by Rondeau et al. (2012). An estimation method based on the first and second order Laplace approximations of the complete partial likelihood has been proposed by Ha et al. (2017) and implemented in the R package *frailtyHL*. However to the best of our knowledge, none of these existing algorithms have been proven to be convergent theoretically.

The aim of this chapter is to propose an efficient convergent EM stochastic algorithm which maximizes the integrated partial likelihood in a large variety of frailty models. In order to prove the convergence of the proposed numerical method, we make use of an extended model which includes the regression parameter as a random variable with Gaussian distribution.This extended model belongs to the exponential family, allowing us to prove the almost sure convergence of the proposed algorithm toward a critical point of the integrated partial likelihood. Furthermore, as shown, maximizing the integrated partial likelihood in this extended model is approximatively equivalent to maximizing the integrated partial likelihood in the frailty model. Moreover, we show through simulation studies the performance of the proposed algorithm in various settings compared with existing methods and highlight the benefit of using the integrated partial likelihood. We also apply the proposed algorithm to analyze mastitis and bladder cancer datasets.

The chapter is organized as follows. Section 2 deals with the frailty model. The integrated partial likelihood function is detailed in Section 3. The extended frailty model and the estimator are presented in Section 4. The stochastic estimation procedure and its convergence property are detailed in Section 5. The simulation and real data studies are presented in Section 6 and Section 7 respectively. The chapter ends with a conclusion and a discussion.

2.2 The Frailty Model

2.2.1 Description of the model

We consider a population of individuals clustered into N groups. We denote by n_i the size of the i-th group for $1 \leq i \leq N$. We denote the event time and censoring time of the individual j in group i by T_{ij} and C_{ij} respectively for $1 \leq i \leq N$ and $1 \leq j \leq n_i$. We observe the variable $X_{ij} = \min(T_{ij}, C_{ij})$ and the censoring indicator defined as

$\Delta_{ij} = \mathbb{1}_{\{T_{ij} \leq C_{ij}\}}$. We denote by $\mathbf{X} = (X_{ij})_{1 \leq i \leq N, 1 \leq j \leq n_i}$ and by $\mathbf{\Delta} = (\Delta_{ij})_{1 \leq i \leq N, 1 \leq j \leq n_i}$ the observations.

We consider the following frailty model where the hazard for the individual j of group i is expressed as follows :

$$\forall t \geq 0 \; h_{ij}(t|b_i) = h_0(t) \;\; \exp(Z_{ij}^t \beta + W_{ij}^t b_i),$$

where $h_0(t)$ denotes the baseline hazard function at time t, Z_{ij} and W_{ij} the covariates of individual j of group i, $\beta \in \mathbb{R}^b$ the unknown regression parameter vector and $b_i \in \mathbb{R}^f$ the frailty vector of individuals of group i. We assume that the probability density function of the unobserved frailty is parametric and denote by γ its parameter taking values in $\mathbb{R}^c$.

Therefore the model parameters are h_0, β and γ. The parameter of interest is usually β, enabling the quantification of the effects of the covariates which is often the main objective of real data analysis.

2.2.2 Assumptions on the model

We introduce the following usual assumptions on the frailty model:

(F1) The censoring times (C_{ij}) are independent of the event times (T_{ij}) and of the frailties (b_i).

(F2) Conditional on the frailties (b_i), the event times (T_{ij}) are independent.

(F3) The frailty vectors $(b_i)_{1 \leq i \leq N}$ are identically and independently distributed having common density g_γ.

(F4) The function h_0 belongs to the set of continuously differentiable functions defined on $\mathbb{R}^+$ taking values in $\mathbb{R}^+$.

(F5) The probability density function of the frailties denoted by g_γ belongs to the set of exponential family of probability density functions where γ takes values in $\mathbb{R}^c$.

Remark 1 *We emphasize that no parametric assumption on the baseline hazard function h_0 is made in this paper. We note here that **(F4)** is required only for the construction of the partial likelihood. The regularity condition is therefore weaker than the one in Kuhn and El-Nouty (2013) where a choice of parametric structure is made on the baseline hazard function.*

2.3 Integrated partial likelihood for the frailty model

We consider the criteria defined by the integrated partial likelihood for the frailty model following the idea of Cox (1972). We consider the conditional partial likelihood defined as follows:

$$L_{\text{cond}}^p(\theta; \mathbf{X}, \mathbf{\Delta} \mid \mathbf{b}) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \left(\frac{\exp(Z_{ij}^t \beta + W_{ij}^t b_i)}{\sum_{(l,k) \in R(X_{(ij)})} \exp(Z_{lk}^t \beta + W_{lk}^t b_l)} \right)^{\Delta_{ij}}$$

where $\theta = (\beta, \gamma) \in \mathbb{R}^b \times \mathbb{R}^c$, $R(X_{(ij)}) = \left\{ 1 \leq l \leq N, 1 \leq k \leq n_l : X_{lk} \geq X_{(ij)} \right\}$ is the set of individuals still at risk at time $X_{(ij)}$ and $\mathbf{b} = (b_i)_{1 \leq i \leq N}$.

We then easily deduce the complete partial likelihood expression:

$$L^p(\theta;\mathbf{X},\mathbf{\Delta},\mathbf{b}) = \prod_{i=1}^{N} g_\gamma(b_i) \times \prod_{i=1}^{N}\prod_{j=1}^{n_i}\left(\frac{\exp(Z_{ij}^t\beta + W_{ij}^t b_i)}{\sum_{(l,k)\in R(X_{(ij)})}\exp(Z_{lk}^t\beta + W_{lk}^t b_l)}\right)^{\Delta_{ij}} \quad (2.1)$$

We emphasize that this partial likelihood no longer involves the baseline h_0 as the partial likelihood in the Cox model.

Finally we define the integrated partial likelihood as defined in Therneau (2018a), also called marginal partial likelihood, obtained by integrating the complete partial likelihood over the unobserved frailties $\mathbf{b}$:

$$L^p_{\text{marg}}(\theta;\mathbf{X},\mathbf{\Delta}) = \int L^p(\theta;\mathbf{X},\mathbf{\Delta},\mathbf{b})d\mathbf{b}$$

Remark 2 *We recall that as in the Cox model, this integrated partial likelihood is not a likelihood function, but acts as one as explained in Ripatti and Palmgren (2000).*

To perform parameter estimation based on this integrated partial likelihood, we will consider maximum likelihood estimation. Since we are in a latent variable model, we will apply a well-chosen stochastic version of the Expectation Maximization (EM) algorithm. Most of the theoretical convergence properties of stochastic EM like algorithms have been established within the exponential family as for example in Delyon et al. (1999), Kuhn and Lavielle (2004), Allassonnière et al. (2010). Since the complete partial likelihood defined in (2.1) does not belong to the exponential family of probability density functions, we introduce in the next section an extended frailty model, which falls within the exponential family.

2.4 Extended frailty model

2.4.1 Description of the extended frailty model

We consider an extended frailty model where the regression parameter β is considered as a population random variable. Therefore we assume in the following that the population variable β follows a Gaussian distribution with unknown expectation $\bar{\beta}$ and fixed variance $\sigma^2_{\bar{\beta}}$.

We denote the extended latent variables by $\xi = (b_i, i = 1,\ldots,n, \beta)$ and the new parameters to be estimated by $\eta = (\bar{\beta}, \gamma)$. The complete likelihood corresponding to the model can be written as follows:

$$L^e(\eta;\mathbf{X},\mathbf{\Delta},\xi) = \prod_{i=1}^{N} g_\gamma(b_i) f_{\bar{\beta}}(\beta) \times \prod_{i=1}^{N}\prod_{j=1}^{n_i}\left(\frac{\exp(Z_{ij}^t\beta + W_{ij}^t b_i)}{\sum_{(l,k)\in R(X_{(ij)})}\exp(Z_{lk}^t\beta + W_{lk}^t b_l)}\right)^{\Delta_{ij}}$$

where f stands for the Gaussian probability density function. This likelihood function belongs to the exponential family whenever the frailty probability density function g_γ belongs to the exponential family. Sufficient statistics are explicit and can be expressed as $S(\xi) = \left(\sum_{i=1}^{N} S_f(b_i), \beta \right)$ where $S_f(b_i)$ are sufficient statistics corresponding to the frailties (b_i).

By assumption **(F5)**, the complete partial likelihood defined in (2.2) belongs to the exponential family since it can be written as follows:

$$L^e(\eta; \mathbf{X}, \mathbf{\Delta}, \xi) = \exp(-\Psi(\eta) + \langle S(\xi), \Phi(\eta) \rangle)$$

where S, Ψ and Φ are Borel functions.

2.4.2 Definition of the maximum integrated partial likelihood estimate in the extended model

Following Therneau (2018a), we define the estimator $\hat{\eta}$ for the vector of parameters as the value that maximizes the integrated partial likelihood of the extended model:

$$\hat{\eta} = \underset{\eta}{\text{argmax}}\ L^e_{\text{marg}}(\eta; \mathbf{X}, \mathbf{\Delta}) \tag{2.2}$$

where the marginal partial likelihood in the extended model is obtained by integrating the complete extended partial likelihood over all the unobserved variables ξ:

$$L^e_{\text{marg}}(\eta; \mathbf{X}, \mathbf{\Delta}) = \int L^e(\eta; \mathbf{X}, \mathbf{\Delta}, \xi) d\xi \tag{2.3}$$

Since the computation of the integrated partial likelihood cannot be perfomed analytically, an EM type algorithm can be implemented for the maximization procedure. Therefore, we propose to calculate $\hat{\eta}$ by using a stochastic version of the EM algorithm following in the footsteps of Kuhn and El-Nouty (2013).

2.4.3 Comparison between maximum integrated partial likelihood estimators in the frailty model and in the extended frailty model

We highlight in this section the link between the extended frailty model and the frailty model introduced in Section 2.2. Indeed, we can write the marginal extended partial likelihood as follows:

$$\begin{aligned}
L^e_{\text{marg}}(\eta; \mathbf{X}, \mathbf{\Delta}) &= \int L^e(\eta; \mathbf{X}, \mathbf{\Delta}, \xi) d\xi \\
&= \int \prod_{i=1}^{N} \prod_{j=1}^{n_i} \left(\frac{\exp(Z_{ij}^t \beta + W_{ij}^t b_i)}{\sum_{(l,k) \in R(X_{(ij)})} \exp(Z_{lk}^t \beta + W_{lk}^t b_l)} \right)^{\Delta_{ij}} g_\gamma(b_i) f_{\bar{\beta}}(\beta) db_i d\beta \\
&= \int L^p_{\text{marg}}(\beta, \gamma; \mathbf{X}, \mathbf{\Delta}) f_{\bar{\beta}}(\beta) d\beta
\end{aligned}$$

Thus the marginal extended partial likelihood is equal to the expectation of the marginal partial likelihood in the

frailty model with respect to the Gaussian distribution chosen for the population variable β in the extended model. Therefore, applying Laplace approximation, we get that

$$L^e_{\text{marg}}(\bar{\beta}, \gamma; \mathbf{X}, \mathbf{\Delta}) \quad \approx \quad L^p_{\text{marg}}(\bar{\beta}, \gamma; \mathbf{X}, \mathbf{\Delta}) \tag{2.4}$$

Thus the maximum likelihood estimate in the extended model will be close to the maximum likelihood estimate in the frailty model.

2.5 Algorithmic method for inference in the extended frailty model

2.5.1 Description of the stochastic EM algorithm with truncation on random boundaries

We consider the stochastic EM algorithm introduced by Kuhn and Lavielle (2004) and extended by Allassonnière et al. (2010) to evaluate the estimator of the parameters defined in (2.2). The proposed algorithm, denoted by Algorithm $\mathcal{A}$ later, is an extension of the stochastic approximation EM algorithm developed by Delyon et al. (1999) where the EM algorithm is coupled with a Markov Chain Monte Carlo (MCMC) procedure to simulate the unobserved frailties.

Let $(\mathcal{K}_q)_q \geq 0$ be a sequence of increasing compact subsets of S such that $\bigcup_{q\geq 0} \mathcal{K}_q = S$ and $\mathcal{K}_q \subset \text{int}(\mathcal{K}_{q+1})$, for all $q \geq 0$.

Initialize η_0 in Θ, ξ_0 and s_0 in two fixed compact sets **K** and $\mathcal{K}_0$ respectively.

Each iteration of the algorithm is composed of four steps detailed below.

Repeat until convergence for $k \geq 1$:

1. **Simulation step:** Draw $\bar{\xi}$ of the unobserved variables from a transition probability $\Pi_{\eta_{k-1}}$ of a convergent Markov chain having as stationary distribution the conditional distribution $\pi^e_{\eta_{k-1}}(.|\mathbf{X}, \mathbf{\Delta})$ defined by

$$\pi^e_\eta(\xi|\mathbf{X}, \mathbf{\Delta}) = \frac{L^e(\eta; \mathbf{X}, \mathbf{\Delta}, \xi)}{\int L^e(\eta; \mathbf{X}, \mathbf{\Delta}, \xi) d\xi}$$

with the current parameters $\bar{\xi} \sim \Pi_{\eta_{k-1}}(\xi_{k-1}, .)$

2. **Stochastic approximation step:** Compute $\bar{s} = s_{k-1} + \mu_k(S(\bar{\xi}) - s_{k-1})$

3. **Truncation step:** If $\bar{s}$ is outside the current compact set $\mathcal{K}_{\kappa_{k-1}}$, where κ is the index of the current active truncation set, or too far from the previous value s_{k-1} then restart the stochastic approximation in the initial compact set, extend the truncation boundary to $\mathcal{K}_{\kappa_k}$ and start again with a bounded value of the missing vari-

able. Otherwise, set $(\xi_k, s_k) = (\bar{\xi}, \bar{s})$ and keep the truncation boundary to $\mathcal{K}_{\kappa_{k-1}}$.

4. **Maximization step:**

$$\eta_k = \underset{\eta}{\text{argmax}}\{-\Psi(\eta) + \langle s_k, \Phi(\eta)\rangle\}$$

Note that the sequence (ξ_k, s_k) generated by this algorithm satisfies two conditions at each iteration k. Namely we check whether the stochastic approximation wanders outside the current compact set and whether the current value is not too far from the previous value. The latter can be expressed as follows:

$$\|s_k - s_{k-1}\| \leq \epsilon_k$$

where $\epsilon = (\epsilon_k)_{k\geq 0}$ is a monotone non-increasing sequence of positive numbers. A more detailed description of the truncation step can be found in Andrieu et al. (2005).

2.5.2 Practical details on the implementation of the algorithm

1. We use as initial values for the regression parameter β the estimated values obtained when fitting the data by the Cox model. For the covariance matrix parameters of the frailty distribution, we do not use too small values in order to avoid a loss of elasticity of the stochastic trajectories of the algorithm.

2. The decreasing positive step size (μ_k) is taken as follows for all $0 \leq k \leq K_0,\ \mu_k = 1$ and for all $k > K_0,\ \mu_k = \frac{1}{(k-K_0)}$ where K_0 is a number to be specified. The algorithm is said to have no memory during the first K_0 iterations. After this burn-in time which allows for the algorithm to visit widely the parameter space, the sequence $(\mu_k)_k$ decreases and converges to zero as $k \to \infty$.

3. The transition kernel used for simulating the unobserved frailty is usually chosen as a transition kernel of a Metropolis Hastings algorithm with proposal distribution q equal to a Gaussian distribution centered at the current value ξ_{k-1} at the k^{th} iteration (cf. Appendix A for more details).

4. We define a stopping criterion based on the relative difference between two consecutive values of the parameters. Let us fix a positive threshold $\epsilon > 0$. If for some $k > 1$:

$$\frac{\|\eta_k - \eta_{k-1}\|}{\|\eta_{k-1}\|} < \epsilon$$

holds true for three consecutive iterations, the algorithm is stopped. We set $\epsilon = 10^{-4}$ in the simulation study.

2.5.3 Convergence property of the algorithm in the extended frailty model

We consider classical assumptions required to prove the convergence of EM like algorithms following those of Delyon et al. (1999).

(M3) The function $\bar{s} : \Theta \to S$ defined as:

$$\bar{s}(\eta) = \int_{\mathbb{R}^l} S(\xi)\pi_\eta^e(\xi \mid \mathbf{X}, \mathbf{\Delta})d\xi$$

is continuously differentiable on Θ.

(M4) The function $l^e : \Theta \to \mathbb{R}$ defined as the marginal extended log-likelihood

$$l^e(\eta) = \log \int_{\mathbb{R}^l} L^e(\eta; \mathbf{X}, \mathbf{\Delta}, \xi)d\xi$$

is continuously differentiable on Θ and

$$\partial_\eta \int_{\mathbb{R}^l} L^e(\eta; \mathbf{X}, \mathbf{\Delta}, \xi)d\xi = \int_{\mathbb{R}^l} \partial_\eta \, L^e(\eta; \mathbf{X}, \mathbf{\Delta}, \xi)d\xi$$

(M5) There exists a function $\hat{\eta} : S \to \Theta$ such that:

$$\forall \, s \in S, \forall \, \eta \in \Theta, \; L(\hat{\eta}(s), s) \geq L(\eta, s)$$

where $L : S \times \Theta \to \mathbb{R}$ is defined as

$$L(\eta, s) = \exp(-\Psi(\eta) + \langle s, \Phi(\eta) \rangle) \tag{2.5}$$

Moreover, the function $\hat{\eta}$ is continuously differentiable on S.

Following Andrieu et al. (2005), we state a first assumption **(A1')** that guarantees the existence of a global Lyapunov function denoted by w defined as:

$$w(s) = -\log \int L^e(\hat{\eta}(s); \mathbf{X}, \mathbf{\Delta}, \xi)d\xi \tag{2.6}$$

for the mean field h defined as:

$$h(s) = \int (S(\xi) - s)\pi_\eta^e(\xi \mid \mathbf{X}, \mathbf{\Delta})d\xi \tag{2.7}$$

(A1') The functions w and h are such that

(i) there exists an $M_0 > 0$ such that

$$\mathcal{S} = \{s \in S, \langle \nabla w(s), h(s) \rangle = 0\} \subset \{s \in S, w(s) < M_0\}$$

where w is defined in (2.6) and h is defined in (2.7).

(ii) there exists $M_1 \in]M_0, \infty]$ such that $\{s \in S, w(s) < M_1\}$ is a compact set.

(iii) the closure of $w(\mathcal{L})$ has an empty interior.

(A4) The sequences $\mu = (\mu_k)_{k\geq 0}$ and $\epsilon = (\epsilon_k)_{k\geq 0}$ are non-increasing, positive and satisfy $\sum_{k=0}^{\infty} \mu_k = \infty$, $\lim_{k\to\infty} \epsilon_k = 0$ and $\sum_{k=1}^{\infty} \{\mu_k^2 + \mu_k \epsilon_k^a + (\mu_k \epsilon_k^{-1})^p\} < \infty$, where $a \in]0, 1]$ and $p \geq 2$.

Finally we consider the usual drift assumption **(DRI)** which is detailed in Andrieu et al. (2005).

Theorem 5 *Assume that **(F1-F5)**, **(M3-M5)**, **(A1')**, **(A4)** and **(DRI)** are fulfilled. Then we have with probability 1*

$$\lim_{k\to\infty} d(\eta_k, \mathcal{L}) = 0$$

where $(\eta_k)_k$ is generated by Algorithm $\mathcal{A}$, $d(x, A)$ denotes the distance from x to any closed subset A and $\mathcal{L} = \{\eta \in \Theta, \partial_\eta \log L^e_{marg}(\eta; \mathbf{X}, \mathbf{\Delta}) = 0\}$.

The assumption **(A1')** corresponds to the assumptions **(A1)** (i), (ii), (iv) of Andrieu et al. (2005) respectively. Assumption **(A4)** deals with the conditions on the step-size sequences involved in the stochastic approximation and truncation steps of the algorithm.

Proof of Theorem 5: We will first apply Theorem 5.5 of Andrieu et al. (2005) to prove the convergence of the sequence (s_k) and checked therefore the required assumptions. To prove that assumption (**A1)(iii)** of Andrieu et al. (2005) is fulfilled in our case, we establish the following lemma following the lines of the proof of Lemma 2 of Delyon et al. (1999) using in our case the partial likelihood instead of the likelihood:

Lemma 6 *Assuming **(F1-F5)** and **(M3-M5)**, for any $s \in S \setminus \mathcal{S}$ $\langle \nabla w(s), h(s) \rangle < 0$*

Proof of Lemma 2 Assumption **(M1)** of Delyon et al. (1999) is implied by **(F5)**. To fulfil assumption **(M2)** of Delyon et al. (1999), it suffices to show that Ψ and Φ are twice continuously differentiable. This is a straight consequence of assumptions **(F1-F5)**. The end of the proof follows the same lines as Lemma 2 of Delyon et al. (1999).

Thereby assumption (**A1)(iii)** of Andrieu et al. (2005) is fulfilled in our case. As detailed in Andrieu et al. (2005), assumptions **(DRI)** imply assumptions **(A2-A3)** by Proposition 6.1. Thus we can apply Theorem 5.5 of Andrieu et al. (2005). We get that the sequence (s_k) generated by the algorithm satisfies $\lim_k d(s_k, \mathcal{S}) = 0$. Following the lines of the proof of Lemma 2 of Delyon et al. (1999), we get that $\lim_k d(\eta_k, \mathcal{L}) = 0$. The proof of Theorem 5 is therefore complete.

2.5.4 Estimation of the Fisher Information Matrix

We consider the usual estimate of the Fisher Information Matrix, namely the observed Fisher information matrix $I_{obs}(\eta) = -\partial^2_\eta \log L^e_{\text{marg}}(\eta; \mathbf{X}, \mathbf{\Delta})$ (see Andersen et al. (1997)). Using Louis's missing information principle (see Louis (1982)), we express the matrix $I_{obs}(\eta)$ as:

$$I_{obs}(\eta) = -\mathbb{E}_\eta\big(\partial^2_\eta \log L^e(\eta; \mathbf{X}, \mathbf{\Delta}, \xi) \mid \mathbf{X}, \mathbf{\Delta}\big) - \text{Cov}_\eta\big(\partial_\eta \log L^e(\eta; \mathbf{X}, \mathbf{\Delta}, \xi) \mid \mathbf{X}, \mathbf{\Delta}\big)$$

where $\mathbb{E}_\eta$ and Cov_η denote respectively the expectation and the covariance under the posterior distribution π^e_η of the frailty.

We approximate the quantity $I_{obs}(\eta)$ by a Monte Carlo sum based on the realizations of the Markov chain generated in the algorithm having as stationary distribution the posterior distribution π^e_η. After a burn-in period, we use the remaining M realizations $(\xi_m)_{1 \le m \le M}$ of the Markov chain to compute the following quantity:

$$\begin{aligned}\hat{I}_M(\hat{\eta}) = &-\frac{1}{M}\sum_{m=1}^{M} \partial^2_\eta \log L^e(\hat{\eta}; \mathbf{X}, \mathbf{\Delta}, \xi_m) \\ &- \frac{1}{M}\sum_{m=1}^{M} \big(\partial_\eta \log L^e(\hat{\eta}; \mathbf{X}, \mathbf{\Delta}, \xi_m) \partial_\eta \log L^e(\hat{\eta}; \mathbf{X}, \mathbf{\Delta}, \xi_m)^t\big) \\ &+ \frac{1}{M^2}\Big(\sum_{m=1}^{M} \partial_\eta \log L^e(\hat{\eta}; \mathbf{X}, \mathbf{\Delta}, \xi_m)\Big)\Big(\sum_{m=1}^{M} \partial_\eta \log L^e(\hat{\eta}; \mathbf{X}, \mathbf{\Delta}, \xi_m)\Big)^t\end{aligned}$$

The ergodic theorem guarantees the convergence of the quantity $\hat{I}_M(\hat{\eta})$ to the observed Fisher information matrix $I_{obs}(\hat{\eta})$ as M goes to infinity (see Meyn and Tweedie (1993)).

2.6 Simulation studies

All numerical studies have been done using R version 3.3.1 on an Intel Core i7-8550U CPU @ 1.99 GHz, 16 GB RAM.

The aim of our numerical experiments is to evaluate the performance of the proposed algorithm for computing the estimator defined in (2.2), called Maximum Integrated Partial Likelihood (MIPL), and to compare it to those of other estimators existing in the literature.

We consider the following setting inspired by the mastitis dataset described in Kuhn et al. (2016):

$$\forall t \ge 0 \; h_{ij}(t|b_i) = h_0(t) \;\; \exp(Z^t_{ij}\beta + b_i), \quad \lambda > 0, \rho > 0 \tag{2.8}$$

We consider a varying number of groups N and set the number of observations per group to $n_i = 4$. The regression parameter β used to simulate the data is set equal to the vector $(2, 3)$ of size 2. The covariates (Z_{ij}) are generated

independently according to a Bernoulli distribution. The frailties (b_i) are drawn from a centered normal distribution with variance $\gamma = 0.7$. Thus the model parameters are $\theta = (2, 3, 0.7)$. We consider also three different censoring settings, namely no censoring, a low censoring setting with 20% of censoring and a moderate one with 40% censoring.

2.6.1 Study of the consistency property of the estimate

We begin by studying the consistency of the estimate $\hat{\eta}$ numerically. The Weibull baseline hazard defined as $h_0(t) = \lambda\rho t^{\rho-1}$ for $t > 0$ is considered in this section using the parameter values $\lambda = 0.01$ and $\rho = 1.5$. The estimate $\hat{\eta}$ is evaluated using Algorithm $\mathcal{A}$.

Table 2.1: Mean of parameter estimates $\hat{\eta}$ and standard deviation in parentheses obtained from 500 repetitions with different number of groups $(N = 10, 20, 50)$. The number of observations per group is fixed at $n_i = 4$.

Censoring	Parameters	True values	N=10	N=20	N=50
0 %	β_1	2	2.111 (0.462)	2.091 (0.459)	1.981 (0.279)
	β_2	3	3.539 (0.813)	3.198 (0.662)	3.010 (0.353)
	γ	0.7	0.543 (0.662)	0.746 (0.584)	0.698 (0.272)
20 %	β_1	2	2.641 (1.60)	2.274 (0.612)	1.972 (0.321)
	β_2	3	4.058 (2.25)	3.282 (0.891)	2.925 (0.372)
	γ	0.7	1.158 (1.89)	0.970 (0.987)	0.708 (0.346)
40 %	β_1	2	2.598 (1.72)	2.148 (0.743)	1.928 (0.318)
	β_2	3	4.191 (2.64)	3.142 (0.949)	2.939 (0.443)
	γ	0.7	1.408 (2.52)	0.943 (1.17)	0.651 (0.419)

The results supporting the numerical consistency of $\hat{\eta}$ are displayed in Table 2.1. In all censoring settings, as the number of groups N progressively increases, the corresponding estimate gets closer to the true values and the standard deviation decreases. We note that the chosen value of σ_β^2 does not influence the estimate of $\bar{\beta}$. Indeed, as long as the value chosen for σ_β^2 is not too small or not too large, the algorithm converges towards the same limit value. We run the algorithm with different values of σ_β^2, namely $\{0.1, 1, 10\}$, on the same dataset and observe that the estimated value of $\bar{\beta}$ is the same whatever the choice for the value of σ_β^2. For example, the three estimates obtained for the first component of $\bar{\beta}$ are 2.012, 2.011 and 2.011.

Figure 2.1 depicts the estimated posterior distribution of the first component of the random variable β. The values are obtained from one run of the algorithm on one dataset. The first iterations are discarded to make sure the Markov

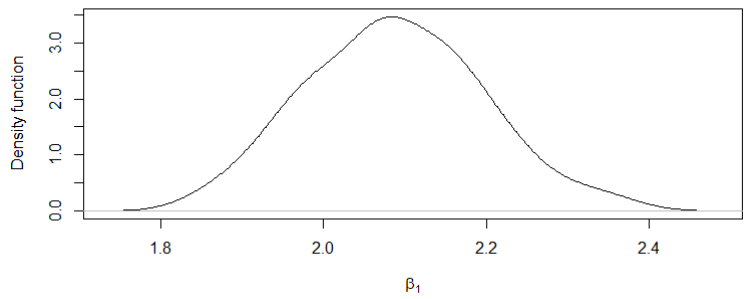

Figure 2.1: Posterior distribution of β_1

chain is in the stationary distribution and one value after every 10 iterations is taken to avoid autocorrelation of the chain. A Gaussian kernel is implemented to obtain a smooth graphical representation.

2.6.2 Comparing the maximum integrated partial likelihood estimate with a parametric estimate

We consider a parametric estimate defined in the model with a Weibull baseline hazard function defined as $h_0(t) = \lambda\rho t^{\rho-1},\ \lambda > 0, \rho > 0$ (cf. Appendix A for more details). We denote the vector of parameters by $\eta_{\text{weibull}} = (\lambda, \rho, \bar{\beta}, \gamma)$. The value of $\hat{\eta}_{\text{weibull}}$ is computed using the MCMC-SAEM algorithm proposed in Allassonnière et al. (2010). The event times are first simulated according to (2.8) with Weibull parameters $\lambda = 0.01$ and $\rho = 1.5$. The number of groups N is fixed at a value of 250. The results are presented in Table 2.2. We conclude that both methods give good estimates in this example where the model is well specified.

We then consider event times simulated using a Gompertz baseline hazard function defined as $h_0(t) = \lambda \exp(\alpha t), \lambda > 0, \alpha > 0$ with Gompertz parameters $\lambda = 0.08$ and $\alpha = 2$. The estimate $\hat{\eta}$ which does not require any modeling assumption of h_0 proves to be a good estimator whereas $\hat{\eta}_{\text{weibull}}$ does not give good results in this example as it can be seen in Table 2.3. The wrong specification of h_0 for the latter introduces bias in the estimation of the parameters. These results therefore highlight the advantages of not assuming a particular baseline hazard function h_0.

2.6.3 Comparing the maximum integrated partial likelihood estimate with other estimates

The aim of the following simulation study is to compare the performances of the maximum integrated partial likelihood estimate with those of other estimates. We consider the estimate based on penalized partial likelihood implemented in the R package *coxme* based on Ripatti and Palmgren (2000) and two estimates based on the h-likelihood implemented in the R package *frailtyHL* detailed in Ha et al. (2017). The estimate in the *coxme* package, denoted by $\hat{\theta}_{\text{coxme}}$, is based on the maximisation of a penalized partial likelihood. The h-likelihood methods implemented in

Table 2.2: Mean of parameter estimates and model-based standard error in parentheses obtained from 500 repetitions with the event times following a Weibull distribution to compare the parametric estimate to the integrated partial likelihood estimate. The number of groups is set to $N = 250$ and the number of observations per group is set to $n_i = 4$.

Censoring	Parameters	True Values	$\hat{\eta}$	$\hat{\eta}_{\text{weibull}}$
0 %	β_1	2	1.988 (0.116)	1.982 (0.133)
	β_2	3	2.984 (0.147)	2.944 (0.145)
	γ	0.7	0.674 (0.126)	0.701 (0.111)
20 %	β_1	2	1.942 (0.158)	1.931 (0.122)
	β_2	3	2.929 (0.198)	2.911 (0.140)
	γ	0.7	0.628 (0.141)	0.604 (0.114)
40 %	β_1	2	1.901 (0.179)	1.862 (0.141)
	β_2	3	2.849 (0.180)	2.806 (0.143)
	γ	0.7	0.631 (0.180)	0.569 (0.138)

frailtyHL are based on a Laplace approximation of the marginal partial likelihood which is then maximised. The two estimators based on h-likelihood differ in the order of the Laplace approximations. They are denoted by $\hat{\theta}_{\text{HL(0,1)}}$ and $\hat{\theta}_{\text{HL(1,2)}}$ with the first one based on the first order Laplace approximation and the second one based on the second order Laplace approximation.

Correct specification of the frailty distribution

We first investigate the effect of censoring when comparing the different estimation procedures. The event times were simulated with Weibull parameters $\lambda = 0.01$ and $\rho = 1.5$. The number of groups N is fixed at a value of 250.

In Table 2.4, in all censoring settings, we observe that the MIPL estimate $\hat{\eta}$ and the estimate $\hat{\theta}_{\text{HL(1,2)}}$ seem to be closer to the true values as opposed to the estimates $\hat{\theta}_{\text{coxme}}$ and $\hat{\theta}_{\text{HL(0,1)}}$.

In terms of computational time, the *coxme* package is the fastest when compared to the MIPL estimation procedure and *frailtyHL* package. We note however that the computational time of the maximum of the integrated partial likelihood estimate is about two times faster than the *frailtyHL* package which is the slowest one.

Table 2.3: Mean of parameter estimates and model-based standard error in parentheses obtained from 500 repetitions with the event times following a Gompertz distribution to compare the parametric estimate to the integrated partial likelihood estimate. The number of groups is set to $N = 250$ and the number of observations per group is set to $n_i = 4$.

Censoring	Parameters	True Values	$\hat{\eta}$	$\hat{\eta}_{\text{weibull}}$
0 %	β_1	2	1.974 (0.125)	1.385 (0.119)
	β_2	3	2.952 (0.166)	2.122 (0.139)
	γ	0.7	0.674 (0.145)	0.271 (0.122)
20 %	β_1	2	1.940 (0.137)	1.357 (0.082)
	β_2	3	2.919 (0.178)	2.000 (0.101)
	γ	0.7	0.618 (0.123)	0.241 (0.064)
40 %	β_1	2	1.884 (0.158)	1.315 (0.101)
	β_2	3	2.822 (0.172)	1.939 (0.112)
	γ	0.7	0.600 (0.159)	0.230 (0.090)

Table 2.4: Mean of parameter estimates and model-based standard error in parentheses obtained from 500 repetitions with the event times following a Weibull distribution. Comparison of MIPL estimate with *coxme* and *frailtyHL* estimates. The number of groups is set to $N = 250$ and the number of observations per group is set to $n_i = 4$.

Censoring	Parameters	True Values	$\hat{\eta}$	$\hat{\theta}_{\text{coxme}}$	$\hat{\theta}_{\text{HL(0,1)}}$	$\hat{\theta}_{\text{HL(1,2)}}$
0 %	β_1	2	1.992 (0.120)	1.981 (0.090)	1.981 (0.096)	2.001 (0.0974)
	β_2	3	2.989 (0.130)	2.962 (0.110)	2.952 (0.117)	2.981 (0.119)
	γ	0.7	0.682 (0.125)	0.665 (0.098)	0.666 (0.091)	0.703 (0.104)
20 %	β_1	2	1.947 (0.160)	1.922 (0.120)	1.930 (0.118)	1.954 (0.120)
	β_2	3	2.934 (0.204)	2.901 (0.151)	2.939 (0.155)	2.976 (0.158)
	γ	0.7	0.623 (0.138)	0.606 (0.107)	0.607 (0.107)	0.647 (0.117)
40 %	β_1	2	1.896 (0.182)	1.850 (0.125)	1.847 (0.125)	1.873 (0.126)
	β_2	3	2.854 (0.185)	2.791 (0.149)	2.808 (0.150)	2.846 (0.151)
	γ	0.7	0.628 (0.171)	0.575 (0.084)	0.576 (0.113)	0.615 (0.121)

Table 2.5: Mean of parameter estimates and model-based standard error in parentheses obtained from 500 repetitions with the event times following a Weibull distribution. Comparison of MIPL estimate with *coxme* and *frailtyHL* estimates when the frailty distribution is misspecified. A mixture of Gaussian frailties is used to simulate the dataset whereas a Gaussian frailty is assumed in the estimation procedure. The number of groups is set to $N = 250$ and the number of observations per group is set to $n_i = 4$.

Censoring	Parameters	True Values	$\hat{\eta}$	$\hat{\theta}_{\text{coxme}}$	$\hat{\theta}_{\text{HL(1,2)}}$
0 %	β_1	2	2.037 (0.150)	1.531 (0.124)	2.022 (0.110)
	β_2	3	3.058 (0.168)	2.304 (0.133)	3.019 (0.128)
	γ	(-)	27.9 (2.99)	6.079 (0.566)	23.0 (2.96)
20 %	β_1	2	1.907 (0.161)	1.468 (0.105)	1.878 (0.118)
	β_2	3	2.903 (0.203)	2.245 (0.115)	2.851 (0.138)
	γ	(-)	23.8 (3.12)	6.778 (3.06)	22.7 (2.96)
40 %	β_1	2	1.783 (0.191)	1.403 (0.120)	1.738 (0.132)
	β_2	3	2.656 (0.198)	2.129 (0.120)	2.581 (0.152)
	γ	(-)	18.9 (3.11)	7.110 (2.26)	17.4 (2.47)

Robustness to misspecification of the frailty distribution

We investigate in this section the case where the frailty distribution is misspecified. We first consider data simulated with a multiplicative Gamma frailty term. We observe that all estimating procedures give good estimations when a normal frailty is assumed for the estimation task (results non presented). Next we consider frailties drawn from a mixture of two normal distributions centered in -10 and 10, with common variances 2 and common weights 0.5.

In all estimating procedures, a normal frailty is assumed. The event times were simulated according to (2.8) with Weibull parameters $\lambda = 0.01$ and $\rho = 1.5$. The number of groups N is fixed at 250 and there are 4 observations per group. The results are presented in Table 2.5. We observe that the estimates obtained with our method and with *frailtyHL* are close to the true value whereas the one obtained by *coxme* does not adjust well to the misspecification of the frailty distribution leading to some bias in the estimation of β in this example.

2.7 Real data analysis

2.7.1 Mastitis dataset analysis

The mastitis dataset, which is the inspiration for the different simulation studies shown in the previous section, consists of 1196 cows that are tracked individually for the time to infection during a period of lactation. Each cow is considered as a group and the four udder quarters ($n_i = 4$) of each cow the observations. Udder quarters without infections are censored at the end of the period of lactation. We consider two covariates, namely parity and location. The parity is considered at two levels, primiparous and multiparous, and is a characteristic of the cow. The location can be front or rear and obviously evolves from one quarter to another in a cow. We use the model defined by equation (2.8) to analyze this dataset. The hazard ratio for parity (multiparous versus primiparous) equals 2.335 with 95 $\%$ confidence interval $[2.271; 2.399]$ and the hazard ratio for location (rear versus front) is equal to 0.774 with 95 $\%$ confidence interval $[0.712; 0.836]$. Finally the parameter γ is estimated as 4.79 with 95 $\%$ confidence interval $[4.594; 4.986]$. The estimates obtained are close to those obtained in Kuhn et al. (2016) where a Weibull baseline assumption is made. This similarity in the estimates can be explained by the fact that the Weibull baseline hazard is adequate for the mastitis dataset as it was shown in Geerdens et al. (2012).

2.7.2 Bladder cancer dataset analysis

We consider a bladder cancer dataset from the European Organisation for Research and Treatment of Cancer (EORTC). A combined analysis was carried out on individual patient data from 2596 superficial bladder cancer patients included in seven European Organization for Research and Treatment of Cancer trials 30781, 30782, 30791, 30831, 30832, 30845, and 30863 (Genito-Urinary tract cancer Group). Only the groups with more than 20 patients were included in the dataset to be analyzed. After data processing, we are left with 39 groups of patients of different sizes. The censoring level is about 51 $\%$ and about 80 $\%$ of the patients follow an intravesical treatment which is the only covariate considered (see Sylvester et al. (2006)). The studies conducted on this dataset suggest that the treatment effect might differ between centers. We therefore introduce, next to the center random effect b_{0i}, also a treatment by center random effect b_{1i}, and allow for correlation between these two random effects within center as done in Kuhn and El-Nouty (2013). This leads us to model the hazard function as follows:

$$h_{ij}(t|b_i) = h_0(t) \ \exp(b_{0i} + Z_{ij}^t(\beta + b_{1i})) \tag{2.9}$$

with $b = (b_0, b_1) \sim \mathcal{N}(0, \Gamma)$ where $\Gamma = \begin{pmatrix} \gamma_0^2 & \gamma_{01} \\ \gamma_{01} & \gamma_1^2 \end{pmatrix}$

We estimate the parameters $\eta = (\bar{\beta}, \gamma_0^2, \gamma_1^2, \gamma_{01})$ by maximising the integrated partial likelihood. We run the

algorithm 100 times using different initial values. We define the step-size sequence (μ_k) such that for all $0 \leq k \leq 350$, $\mu_k = 1$ and for all $k > 350$, $\mu_k = \frac{1}{(k-350)}$. The mean of the estimates obtained is equal to $\hat{\eta} = (-0.206, 0.0712, 0.0435, 0.0428)$ with respective model standard errors $(0.007, 0.0001, 0.0002, 0.000001)$. Some trajectories of parameters estimates are shown in Figure 2.2. We observe that whatever the initial conditions, all trajectories lead to more or less the same limits.

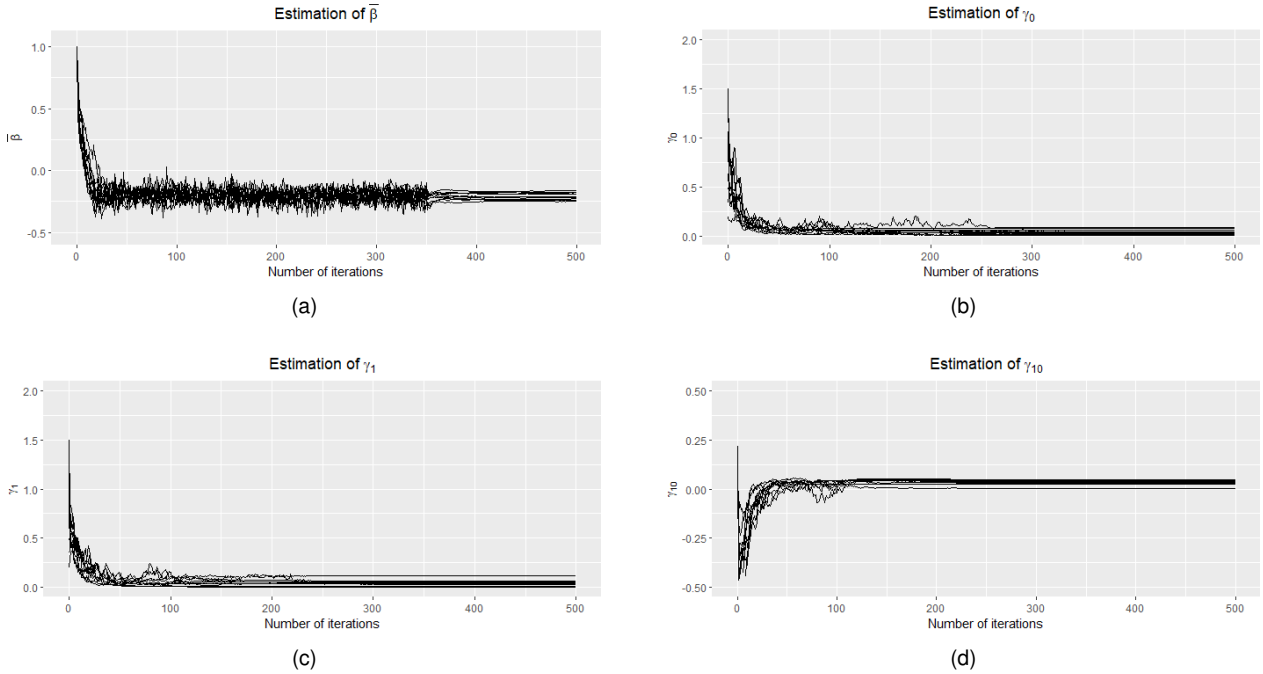

Figure 2.2: Representation of 100 runs of the algorithm for estimating parameters in the bladder cancer dataset.

Finally we perform a simulation study with the same data structure as the bladder cancer dataset. The event times are simulated following the model defined by equation (2.9) with a Weibull baseline hazard parametrized by $\lambda = 0.01$, $\rho = 1.5$ and the regression paramater $\beta = -0.2$. The frailty variances γ_0^2 and γ_1^2 are set to 0.08 and 0.05 respectively and the covariance term γ_{01} is set to 0.04. The group sizes are fixed to the same configuration of those in the real bladder cancer dataset. The mean of the estimates obtained on 500 repetitions equals $\hat{\eta} = (-0.211, 0.0729, 0.0455, 0.036)$ and is close to the true values for all the parameters. The model standard errors based on the estimation of the observed Fisher information matrix are also well estimated (results not presented).

2.8 Conclusion and discussion

We proposed a new method to model correlated survival data which has several strengths. First, it makes no assumption on the baseline hazard, which has proven to be important in the simulation study. Second, it can cope with diverse settings with respect to the correlation structure, from simple shared frailty models to correlated frailty

models and models where the random effects interact with the covariate. The method also accommodates the most frequently used frailty distributions such as Gamma, positive stable, inverse Gaussian and normal distributions. Third, we also proved the almost sure convergence of our method under classical assumptions, which is not the case for the alternative methods to fit frailty models. Indeed we extended the common frailty model specification assuming that the regression parameter is a random variable and has a Gaussian distribution in order to bring the model into the exponential family where we could establish the almost sure convergence of the algorithm. Maximizing the integrated partial likelihood in the extended model remains approximatively the same as maximizing the integrated partial likelihood in the common frailty model, leading to efficient parameter estimates.

Since we have proposed an efficient convergent algorithm to compute the maximum of the integrated partial likelihood estimate in various frailty models, it would be now of great interest to study theoretically its asymptotic properties such as consistency, asymptotic normality and efficiency.

Chapter 3

Convergence properties of maximum likelihood estimates in parametric shared frailty models

3.1 Introduction

In this chapter, we study the convergence properties of the maximum likelihood estimates (MLE) in parametric frailty models. As discussed in section 1.6.1 of Chapter 1, to the best of our knowledge, the convergence properties of the MLE in parametric frailty models have not yet been studied.

In regular statistical models, the asymptotic properties of the MLE have been widely studied and established in different settings. We refer to the works of Wald (1949) where he proved the consistency of the MLE for observations modeled by independent and identically distributed random variables. In Bradley and Gart (1962), the authors show the weak consistency and asymptotic normality of the MLE for observations modeled by independent but not identically distributed random variables. However, the assumptions required to apply the theorems of Bradley and Gart (1962) or Wald (1949) are not easily verifiable in several cases. Moreover, additional difficulties come up when considering non independent observations. Let us consider for example latent variable models which allow to take into account dependencies in data through non-observed variables also called latent variables. To carry out maximum likelihood estimation, those latent variables have to be integrated out to obtain the marginal likelihood. As a result, the analytical expression of the marginal likelihood involves an integral which considerably increases the difficulty in verifying the assumptions of the convergence results of Bradley and Gart (1962), Wald (1949). Bearing that in mind, the works of Nie (2006) and Nie (2007) aim to provide easier verifiable conditions for consistency and asymptotic normality of the MLE in generalized linear and nonlinear mixed-effects models. In particular, their

regularity assumptions required to guarantee the theoretical properties of the MLE are satisfied in some simple examples, namely mixed-effects logistic regression models and growth curve models. They also rightly point out the following statement : "Convergence rates of MLEs differ from parameter to parameter, which is not well explained in the literature".

To illustrate this statement, let us consider data clustered in N groups with J observations per group. This setting includes in particular the one of repeated measurements. We refer to Davidian and Giltinan (1995) for more details. With such a data structure, we may consider three possible scenarios for the asymptotic behavior: (A) $N \to \infty$ while J remains finite, (B) N remains finite while $J \to \infty$ and (C) both $N, J \to \infty$. We illustrate scenario (A) through the mastitis data presented in Laevens et al. (1997). A population of N cows is under study with each cow having a fixed number of udder quarters ($J = 4$). We can sample more and more cows so that $N \to \infty$ while $J = 4$ remains fixed. Consider the interviewer's variability data of Anderson and Aitkin (1985) which could be an example of scenario (B) or (C). The authors study data comprising interviewers and respondents. If a finite sample of interviewers is available while we can sample more and more respondents, then we are closer to scenario (B). And if we can take bigger samples of both interviewers and respondents, this classifies as scenario (C) with both $N, J \to \infty$.

3.1.1 Influence of the frailty terms on the convergence rates

We first investigate the influence of the frailty terms, namely the random effects of the model. Following the works of Nie (2007) in generalized linear and nonlinear mixed-effects models, we focus on the role played by J in particular when J goes to infinity. He formulated a condition based on a reparameterization of the model that allows for the determination of different convergence rates for MLEs. Indeed he established different convergence rates for the MLE depending on the expression of the conditional likelihood with respect to parameters and random effects. More precisely, under given regularity conditions (i) parameters of the random effects distribution are $\sqrt{N}$-consistent; (ii) parameters involved in the conditional likelihood with an additive random effect term are $\sqrt{N}$-consistent; (iii) other parameters involved in the conditional likelihood are $\sqrt{NJ}$-consistent. Moreover his results can be applied to a wide variety of latent variable models and are not limited to generalized linear and nonlinear mixed-effects models since their assumptions are based on quantities such as the distribution of the latent variables and on the distribution of the observations conditional to the latent variables. However, the assumptions set out by Nie (2006) and Nie (2007) are restrictive and are not satisfied in the framework of frailty models. Nonetheless, there is no reason to believe these results do not apply to frailty models. We conduct an intensive simulation study to highlight this phenomenon in a shared frailty model in section 3.4.

3.1.2 Influence of the structure of covariates on the convergence rates

We also investigate the effect of the structure of covariates on the convergence rates of MLEs in models with repeated measurements. With data clustered into groups as detailed above, covariates can take different forms with two special cases that are most common. Let us define the covariates $(Z_{ij})_{1\leq i\leq N, 1\leq j\leq J}$. The first case concerns covariates which we refer to as covariates at the group level. These covariates take the same value for all observations in the group, i.e $Z_{ij} = \bar{Z}_i$ for all j. The second case concerns covariates which we refer to as covariates at the observation level. These covariates vary across the observations within each group. Let us take the example of the mastitis dataset introduced in the previous chapter. The time to infection of the 4 udder quarters of 1196 cows are recorded. Each cow is considered as a group and the infection times of udder quarters are the observations. We therefore have four observations per group. We will focus on the covariates considered in this dataset which are parity and location. Parity is a binary covariate that indicates whether a cow has given birth once or more than once; primiparous or multiparous. This is therefore a characteristic of the cow. On the other hand, location is a covariate that indicates whether the udder quarter is located at the front or at the rear. It is therefore a covariate that evolves within the cow. A snippet of the mastitis data is shown in Table 3.1.

Table 3.1: Snippet of mastitis data

Cow	Parity	Location	Time to infection	status
1	0	0	0.73	1
1	0	0	0.73	1
1	0	1	1.30	1
1	0	1	0.73	1
2	1	0	0.73	1
2	1	0	2.22	1
2	1	1	2.22	1
2	1	1	0.73	1
3	1	0	0.82	1
3	1	0	1.75	0
3	1	1	1.61	1
3	1	1	1.75	0

The Fisher information provides a way of measuring the amount of information contained in the data. Intuitively, observations associated to a covariate at the observation level bring more information than observations associated to a covariate at the group level. Therefore, we conjecture a slower convergence rate for the MLE of the parameter associated to the covariate at the group level. Indeed, with covariates taking the same value within the group, there is added information only when the number of groups increase and not when the number of observations per group increase. On the other hand, covariates at the observation level carry more information with both increasing numbers of groups and observations per group.

The distinction between the two kinds of covariates has been made in Neuhaus and Kalbfleisch (1998) where they refer to covariates at the group level as cluster-constant or cluster-level and covariates at the observation level

as within-cluster covariates. However, the authors focus only on covariates at the observation level. We instead consider both types of covariates and study their influence on the convergence rates of MLEs in the two cases. The role played by the structure of the covariates is established theoretically in a linear mixed-effects model in section 3.3. We also demonstrate this phenomenon by means of an intensive simulation study for the parametric shared frailty model in section 3.5.

3.2 Convergence properties of maximum likelihood estimates in mixed-effects models

In this section, let us present the asymptotic results of Nie (2006) and Nie (2007) for generalized linear and nonlinear mixed-effects models and discuss their possible extensions to frailty models.

3.2.1 Consistency and asymptotic normality of the MLE in generalized linear and nonlinear mixed-effects models

We first recall the theorems established by Nie (2006) and Nie (2007) for the MLE in generalized linear and nonlinear mixed-effects models. Let us define a general model with observations denoted by $X_i = (X_{i1}, ..., X_{iJ})$ for groups $i = 1, \ldots, N$ consisting of J observations per group. Conditionally to the random effect $\mathbf{b} = (b_i)_{1 \leq i \leq N}$, the conditional probability density function of X_i is denoted by $\pi_i(\tau; X_i, b_i)$ where τ are the fixed-effects parameters of the model. The random effects b_i are assumed to follow a probability distribution of density g_η parameterized by η. We therefore denote by $\theta = (\tau, \eta) \in \Theta$ the parameters of the model. The marginal likelihood based on observation X_i is obtained by integrating the complete likelihood with respect to the random effect b_i :

$$M_i(\theta; X_i) = \int \pi_i(\tau; X_i, b_i) g_\eta(b_i) db_i$$

We first recall the theorem of Nie (2006) which establishes the consistency of MLEs in generalized linear and nonlinear mixed-effects models.

Theorem 7 *The maximum likelihood estimating equation*

$$\frac{\partial}{\partial \theta} \prod_{i=1}^{N} M_i(\theta; X_i) = 0 \tag{3.1}$$

has a root $\hat{\theta}_{MLE}$ *such that*

$$\mathbb{P}\Big(\lim_{N \to \infty} \hat{\theta}_{MLE} = \theta_0 \Big) = 1$$

if there is $C > 0$*, and an open subset* w *of* Θ *which contains the true parameter* $\theta_0 = (\tau_0, \eta_0)$*, such that the following conditions are true for all* $1 \leq i \leq N$*.*

1. *For almost all* X_i *the density* $\pi_i(\tau; X_i, b_i)$ *and* $g_\eta(b_i)$ *admit all first, second and third derivatives on* $\theta = (\tau, \eta) \in w$. *Conditions (e.g, uniformly integrability) are assumed to allow the change of order of integrations over* b_i *and all first, second and third differentiation of* $M_i(\theta; X_i)$ *on* Θ.

2. *There exist functions* $\phi(b_i)$ *for* $|\gamma| \leq 3$ *and* $\Phi(b_i)$, *such that* $\sup_{\theta\in\Theta}|D^\gamma g_\eta(b_i)| \leq \phi_\gamma(b_i)$ *and* $\inf_{\theta\in\Theta}|g_\eta(b_i)| \geq \Phi(b_i) > 0$, $\mathbb{E}_{b_i|\theta_0}[(g_{\eta_0}(b_i)^{-1}\phi_\gamma(b_i))^{32} \leq C$ *and* $\mathbb{E}_{b_i|\theta_0}[(\Phi(b_i)^{-1}g_{\eta_0}(b_i))^{32}] \leq C$.

3. *There exist functions* $F_{\alpha i}$*s for* $|\alpha| \leq 3$ *and* R_i*s such that* $\sup_{\theta\in\Theta}|D^\alpha\pi_i(\tau; X_i, b_i)| \leq F_{\alpha i}(X_i, b_i)$, $\inf_{\theta\in\Theta}\pi_i(\theta; X_i, b_i) \geq R_i(X_i, b_i) > 0$, $\mathbb{E}_{b_i|\theta_0}\mathbb{E}_{X_i|b_i,\tau_0}[(\pi_i(\tau_0; X_i, b_i)^{-1}F_{\alpha i}(X_i, b_i))^{32}] \leq C$, *and* $\mathbb{E}_{b_i|\theta_0}\mathbb{E}_{X_i|b_i,\tau_0}[(\pi_i(\tau_0; X_i, b_i)R_i(X_i, b_i))^{32}] \leq C$.

4. $\liminf_{N\to\infty} \lambda_N^f = \lambda^f > 0$, *where* λ_N^f *is the smallest eigenvalue of* $F_N(\theta_0) = -\frac{1}{N}\sum_{i=1}^N \mathbb{E}_{X_i|\theta_0}\left(\frac{\partial^2 ln\, M_i(\theta;X_i)}{\partial\theta\theta^t}\right)$

Theorem 7 has been established for generalized linear and nonlinear mixed-effects models. In particular, Nie (2006) proved that the assumptions required to apply Theorem 7 are satisfied by models commonly found in the literature such as a logistic mixed-effects model for interviewer variability data (cf. Anderson and Aitkin (1985)) and a growth curve model (cf. Davidian and Giltinan (1993)). Moreover, the theorem can be applied to latent variables models since the quantities involved in his assumptions are the random effects distribution and the conditional distribution.

Next, we recall the theorem of Nie (2007) which establishes the asymptotic normality and convergence rates of MLEs in generalized linear and nonlinear mixed-effects models. The distinction between the convergent rates of parameter estimates is achieved following conditions on a reparameterization of the conditional likelihood.

Condition 3. of Nie (2007). Assume that $\pi_i(\tau; X_i, b_i)$ has the form or can be reparameterized into the following form :

$$\pi_i(\tau; X_i, b_i) = \pi_i(X_i, \tau^{(1)} + b_i, \tau^{(2)})$$

where $\tau = (\tau^{(1)}, \tau^{(2)})$.

Theorem 8 *Under conditions 1, 3 and 4, and conditions in Lemma 1 or 2 of Nie (2007), as* $N \to \infty$ *and* $J \to \infty$,

$$\sqrt{NJ}(\tau^{(2)} - \tau_0) \sim \mathcal{N}(0, I^{-1}(\tau_0^{(2)})),$$

$$\sqrt{N}(\hat{\eta}_* - \eta_{*0}) \sim \mathcal{N}(0, \phi_*^{-1}),$$

where $\hat{\eta}_* = (\hat{\eta}, \hat{\tau}^{(1)})$ *and* $\eta_{*0} = (\eta_0, \tau_0^{(1)})$,

$$I^{-1}(\tau^{(2)}) \text{ is a Laplace approximation of } \lim_{N,J\to\infty}\frac{1}{NJ}\mathbb{E}_{X_i}\left(\left[-\frac{\partial^2 ln\, M_i(\theta; X_i)}{\partial\tau^{(2)}\partial\tau^{(2)^t}}\right]\right),$$

and

$$\phi_* = \mathbb{E}_{X_i}\left[-\frac{\partial^2 ln\, g(b_i, \eta_*)}{\partial\eta_*\partial\eta_*^t}\right]$$

The author conducted a simulation study based on existing models mentioned earlier such as the logistic mixed-effects model for interviewer variability data (cf. Anderson and Aitkin (1985)) and a growth curve model (cf. Davidian and Giltinan (1993)) to illustrate the convergence rates of the MLEs.

3.2.2 Extension of these results to frailty models and discussion

We set out to discuss the extensions of Theorem 7 and Theorem 8 to a parametric shared frailty model. Let us define a general parametric shared frailty model and the corresponding likelihood expressions. We consider N groups and J observations per group. We denote the event time and censoring time of the individual j in group i by T_{ij} and C_{ij} respectively for $1 \leq i \leq N$ and $1 \leq j \leq J$. We observe the variable $X_{ij} = \min(T_{ij}, C_{ij})$ and the censoring indicator defined as $\Delta_{ij} = \mathbb{1}_{\{T_{ij} \leq C_{ij}\}}$. The shared frailty model for $i = 1, \ldots, N$, $j = 1, \ldots, J$, is expressed as follows :

$$h_{ij}(X_{ij}|b_i) = h_\nu(X_{ij}) \exp(Z_{ij}^t \beta + W_{ij} b_i) \tag{3.2}$$

where $h_\nu(X_{ij})$ is the baseline hazard for observation j of group i at time X_{ij}, b_i the frailty vector of group i, β the unknown regression parameter, Z_{ij} and W_{ij} the covariates associated to the regression parameter and the frailty respectively. The parameters of the model are $\theta = (\nu, \beta, \eta)$ where ν are the parameters associated to the baseline hazard function and η is the parameter associated to the frailty distribution.

The conditional likelihood for the i^{th} group is therefore written as:

$$\begin{aligned} L_{\text{cond},i}(\theta; X_i, \Delta_i|b_i) = \prod_{j=1}^{J} \Big(h_\nu(X_{ij}) \exp(Z_{ij}^t \beta + W_{ij}^t b_i)\Big)^{\delta_{ij}} \\ \exp(-H_\nu(X_{ij}) \exp(Z_{ij}^t \beta + W_{ij}^t b_i)) \end{aligned} \tag{3.3}$$

where $X_i = (X_{i1}, \ldots, X_{iJ})$, $\Delta_i = (\Delta_{i1}, \ldots, \Delta_{iJ})$ and $H_\nu(X_{ij})$ is the cumulative baseline hazard for observation j of group i at time X_{ij}. The marginal likelihood for the i^{th} group is obtained by integrating the complete likelihood over the frailties b_i:

$$L_{\text{marg},i}(\theta; X_i, \Delta_i) = \int L_{\text{cond},i}(\theta; X_i, \Delta_i|b_i) g_\eta(b_i) db_i \tag{3.4}$$

where g_η is the probability density of the frailty distribution. The maximum likelihood estimator (MLE) for parameter θ is the value $\hat{\theta}_{\text{MLE}}$ of θ which maximises the marginal likelihood defined in equation (3.4) :

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\text{argmax}}\ L_{\text{marg}}(\theta; X, \Delta) \tag{3.5}$$

We now discuss the conditions of applications of Theorem 7 to parametric shared frailty models. Condition 1

is easily verifiable and we can follow along the lines of Nie (2006) to verify Condition 2 which is shown to be valid when the frailties b_i follow a normal distribution. Condition 4 only requires the Fisher information matrix to be positive definite and is therefore not restrictive. However, Condition 3 proves harder to verify even in usual parametric shared frailty models. Indeed using explicit expressions for the optimal quantities $R_i(X_i, b_i)$ and $F_{\alpha i}(X_i, b_i)$ result in making unrealistic assumptions on the distribution of event times.

Moreover, the possible influence of the structure of covariates has not been highlighted in Nie (2007). Therefore, within the framework of parametric shared frailty models, we conduct a simulation study in section 3.4. The simulation setting is set up with two objectives. First, we investigate the effect on the convergence rate of a parameter involved in the parametric shared frailty model with an additive frailty term. Second, we investigate the effect of the structure of covariates on the convergence rates of the MLE of associated regression parameters.

First, we highlight the effect of the structure of covariates analytically in the case of a simple linear mixed-effects model in the next section.

3.3 Case study of the convergence rates of maximum likelihood estimates in a linear mixed-effects model

3.3.1 Description of the model and likelihood expressions

Let us consider a population of N groups with each group consisting of J observations. The linear mixed-effects model (LMM) can therefore be defined for $i = 1, \dots, N$:

$$Y_i = Z_i\beta + W_i b_i + \epsilon_i \tag{3.6}$$

where $Y_i \in \mathbb{R}^J$ is the outcome variable, $Z_i \in \mathbb{R}^{J\times p}$ is a $J \times p$ design matrix of p predictor variables, $\beta \in \mathbb{R}^p$ are the fixed-effects regression coefficients, $W_i \in \mathbb{R}^{J\times q}$ is a $J \times q$ design matrix for the q random effects b_i and $\epsilon_i \in \mathbb{R}^J$ is the vector of the residuals. Also, it is assumed that b_i and ϵ_i follow independent and multivariate normal distributions such that $b_i \sim \mathcal{N}_q(0, D)$ and $\epsilon_i \sim \mathcal{N}_J(0, \Sigma)$

The LMM can be written as a two level hierarchical mixed-effects model :

$$\begin{cases} Y_i|b_i \sim \mathcal{N}_J(Z_i\beta + W_i b_i, \Sigma) \\ b_i \sim \mathcal{N}_q(0, D) \end{cases} \tag{3.7}$$

The outcome variable Y_i therefore follows the following marginal distribution :

$$Y_i \sim \mathcal{N}_J(Z_i\beta, \Gamma_i)$$

where $\Gamma_i = W_i D W_i^t + \Sigma$. The log density of the response variable Y_i can therefore be written as :

$$\log f(Y_i) = -\frac{J}{2}(\log(2\pi)) - \frac{1}{2}\log(\det(\Gamma_i)) - \frac{1}{2}(Y_i - Z_i\beta)^t \Gamma_i^{-1}(Y_i - Z_i\beta) \tag{3.8}$$

We sum over all the groups to obtain the log-likelihood :

$$\log f(Y) = -\frac{NJ}{2}(\log(2\pi)) - \frac{1}{2}\sum_{i=1}^{N}\log(\det(\Gamma_i)) - \frac{1}{2}\sum_{i=1}^{N}(Y_i - Z_i\beta)^t \Gamma_i^{-1}(Y_i - Z_i\beta) \tag{3.9}$$

In an effort to obtain analytical expressions of the estimates and their variances, let us set :

$$p = q = 1$$

$$D = \delta^2$$

$$\Sigma = \sigma^2 I_J$$

$$W_i = (W_{i1}, \ldots, W_{iJ})^t$$

After the above simplification, we can write Γ_i as $\Gamma_i = \sigma^2 I + \delta^2 W_i W_i^t$.

3.3.2 Maximum likelihood estimates of the parameters

In order to compute the maximum likelihood estimate for β, we differentiate the log-likelihood defined in equation (3.9) with respect to β,

$$\frac{\partial}{\partial\beta}\log f(Y) = \sum_{i=1}^{N}(Y_i - Z_i\beta)^t \Gamma_i^{-1} Z_i \tag{3.10}$$

The MLE $\hat{\beta}$ for parameter β is obtained by solving the score equation

$$\frac{\partial}{\partial\beta}\log f(Y) = 0 \tag{3.11}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{N} Y_i^t \Gamma_i^{-1} Z_i}{\sum_{i=1}^{N} Z_i^t \Gamma_i^{-1} Z_i} \tag{3.12}$$

We state the following Lemma that will be used to compute the matrix Γ_i.

Lemma 9 *If A and C are invertible, then*

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

Lemma 9 (cf. Henderson and Searle (1981)) is known as the Woodbury matrix identity. We verify the conditions to apply the Lemma as follows : $A = \sigma^2 I$ and is invertible. $B = \delta^2 W_i W_i^t$ where $C = \delta^2$ which is invertible. Now let

us compute Γ_i^{-1} by Lemma 9 :

$$\begin{aligned}\Gamma_i^{-1} &= (\sigma^2 I + \delta^2 W_i W_i^t)^{-1} \\ &= \sigma^{-2} I - \sigma^{-2} I W_i (\delta^{-2} + W_i^t W_i \sigma^{-2})^{-1} W_i^t \sigma^{-2} I \\ &= \sigma^{-2} I - \frac{\sigma^{-2} I (\delta^2 W_i W_i^t) \sigma^{-2} I}{\delta^2 (\delta^{-2} + W_i^t W_i \sigma^{-2})} \\ &= \sigma^{-2} I - \frac{\sigma^{-2} I (\delta^2 W_i W_i^t) \sigma^{-2} I}{1 + \frac{\delta^2}{\sigma^2} ||W_i||^2}\end{aligned}$$

Let us assume that $W_i = \mathbb{I}_J$, then

$$\Gamma_i^{-1} = \sigma^{-2} I - \sigma^{-2} \frac{(\delta^2 J_J) \sigma^{-2}}{\sigma^2 + J\delta^2} \tag{3.13}$$

where J_J is a matrix of size $J \times J$ with only ones. We now replace this expression in equation (3.12) :

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^N Y_i^t \Gamma_i^{-1} Z_i}{\sum_{i=1}^N Z_i^t \Gamma_i^{-1} Z_i} \\ &= \frac{\sigma^{-2} \sum_{i=1}^N \sum_{j=1}^J Y_{ij} Z_{ij} - \frac{\sigma^{-2}\delta^2}{\sigma^2 + J\delta^2} \sum_{i=1}^N (\sum_{j=1}^J Y_{ij})(\sum_{j=1}^J Z_{ij})}{\sigma^{-2} \sum_{i=1}^N \sum_{j=1}^J Z_{ij}^2 - \frac{\sigma^{-2}\delta^2}{\sigma^2 + J\delta^2} \sum_{i=1}^N (\sum_{j=1}^J Z_{ij})^2} \\ &= \frac{(\sigma^2 + J\delta^2) \sum_{i=1}^N \sum_{j=1}^J Y_{ij} Z_{ij} - \delta^2 \sum_{i=1}^N (\sum_{j=1}^J Y_{ij})(\sum_{j=1}^J Z_{ij})}{(\sigma^2 + J\delta^2) \sum_{i=1}^N \sum_{j=1}^J Z_{ij}^2 - \delta^2 \sum_{i=1}^N (\sum_{j=1}^J Z_{ij})^2}\end{aligned} \tag{3.14}$$

3.3.3 Influence of the structure of covariates on the convergence rates of the estimates

In order to study the convergence rate of the maximum likelihood estimate $\hat{\beta}$, we compute the variance of the estimate :

$$
\begin{aligned}
\mathsf{Var}(\hat{\beta}) &= \mathsf{Var}\left(\frac{\sum_{i=1}^{N} Y_i^t \Gamma_i^{-1} Z_i}{\sum_{i=1}^{N} Z_i^t \Gamma_i^{-1} Z_i}\right) \\
&= \frac{1}{(\sum_{i=1}^{N} Z_i^t \Gamma_i^{-1} Z_i)^2} \sum_{i=1}^{N} \mathsf{Var}(Y_i^t \Gamma_i^{-1} Z_i) \\
&= \frac{1}{(\sum_{i=1}^{N} Z_i^t \Gamma_i^{-1} Z_i)^2} \sum_{i=1}^{N} Z_i^t \Gamma_i^{-1} \mathsf{Var}(Y_i^t) \Gamma_i^{-1} Z_i \\
&= \frac{1}{(\sum_{i=1}^{N} Z_i^t \Gamma_i^{-1} Z_i)^2} \sum_{i=1}^{N} Z_i^t \Gamma_i^{-1} \Gamma_i \Gamma_i^{-1} Z_i \\
&= \frac{1}{(\sum_{i=1}^{N} Z_i^t \Gamma_i^{-1} Z_i)^2} \sum_{i=1}^{N} Z_i^t \Gamma_i^{-1} Z_i \\
&= \frac{1}{(\sum_{i=1}^{N} Z_i^t \Gamma_i^{-1} Z_i)} \\
&= \frac{1}{\sigma^{-2} \sum_{i=1}^{N} \sum_{j=1}^{J} Z_{ij}^2 - \frac{\sigma^{-2}\delta^2}{\sigma^2 + J\delta^2} \sum_{i=1}^{N} (\sum_{j=1}^{J} Z_{ij})^2}
\end{aligned} \tag{3.15}
$$

In this simple linear mixed-effects model, we aim to show that the convergence rate may differ depending on the structure of covariates.

Case of covariates at the group level

First, we consider covariates such that $Z_{i1}^{(1)} = Z_{i2}^{(1)} = \cdots = Z_{iJ}^{(1)}$ for all i. We now compute the inverse of the variance of the estimate to determine the rate of convergence in this particular case.

$$
\begin{aligned}
\mathsf{Var}(\hat{\beta})^{-1} &= \sigma^{-2} \sum_{i=1}^{N} \sum_{j=1}^{J} Z_{ij}^{(1)^2} - \frac{\sigma^{-2}\delta^2}{\sigma^2 + J\delta^2} \sum_{i=1}^{N} \Big(\sum_{j=1}^{J} Z_{ij}^{(1)} \Big)^2 \\
&= J\sigma^{-2} \sum_{i=1}^{N} Z_{i1}^{(1)^2} - \frac{\sigma^{-2}\delta^2}{\sigma^2 + J\delta^2} \sum_{i=1}^{N} (J Z_{i1}^{(1)})^2 \\
&= J\sigma^{-2} \left(1 - \frac{J\delta^2}{\sigma^2 + J\delta^2}\right) \sum_{i=1}^{N} Z_{i1}^{(1)^2} \\
&= \frac{J}{\sigma^2 + J\delta^2} \sum_{i=1}^{N} Z_{i1}^{(1)^2}
\end{aligned} \tag{3.16}
$$

$$
\mathsf{Var}(\hat{\beta})^{-1} = \frac{J}{\sigma^2 + J\delta^2} \sum_{i=1}^{N} Z_{i1}^{(1)^2}
$$

The MLE of parameter β is therefore $\sqrt{N}-$consistent. Next, let us consider the convergence rate of the MLE of the regression parameter associated to a covariate that varies at the observation level.

Case of covariates at the observation level

Let $Z_{ij}^{(2)}$ be defined such that $Z_{i,2j+1}^{(2)} = -Z_{i,2j}^{(2)}$ with $1 \leq j \leq 2J$. This covariate structure implies that :

$$\sum_{j=1}^{2J} Z_{ij}^{(2)} = 0$$

$$\sum_{i=1}^{N}\sum_{j=1}^{2J} {Z_{ij}^{(2)}}^2 = 2J \sum_{i=1}^{N} {Z_i^{(2)}}^2$$

We compute the inverse of the variance of the estimate

$$\begin{aligned} \text{Var}(\hat{\beta})^{-1} &= \sigma^{-2} \sum_{i=1}^{N}\sum_{j=1}^{2J} {Z_{ij}^{(2)}}^2 - \frac{\sigma^{-2}\delta^2}{\sigma^2 + J\delta^2} \sum_{i=1}^{N} \Big(\sum_{j=1}^{2J} Z_{ij}^{(2)} \Big)^2 \\ &= 2J\sigma^{-2} \sum_{i=1}^{N} {Z_i^{(2)}}^2 \end{aligned} \tag{3.17}$$

The MLE of parameter β is therefore $\sqrt{NJ}-$consistent. We show through this example that the structure of a covariate can influence the convergence rate of the associated parameter estimate. For instance, the MLE of a regression parameter for a covariate at the observation level has a faster convergence rate than for a covariate at the group level.

We conjecture that the same conclusion should apply to frailty models. However, proving the result in frailty models requires complex calculations. In particular there is no known analytic expression when solving the score equation for the regression parameter in those models. Therefore, we conduct a simulation study to investigate the convergence rates of MLEs in parametric shared frailty models presented in the next section.

3.4 Simulation study: Convergence properties of the MLE in parametric shared frailty models

The aim of this section is to conduct an intensive simulation study in the framework of the parametric shared frailty model. The simulation setting is detailed in the next section.

3.4.1 Description of the Weibull shared frailty model

We consider N groups and J observations per group. The event times and censoring times for observation j of group i are denoted by T_{ij} and C_{ij} respectively. We observe the variable $X_{ij} = \min(T_{ij}, C_{ij})$ and the censoring indicator defined as $\Delta_{ij} = \mathbb{1}_{\{T_{ij} \leq C_{ij}\}}$. The times to event are assumed to follow a Weibull distribution parametrized by λ and ρ. For $i = 1, \ldots, N$, $j = 1, \ldots, J$, we consider the shared frailty model defined in equation (3.2):

$$h_{ij}(X_{ij}|b_i) = \lambda\rho X_{ij}^{\rho-1} \exp(Z_{ij}^t\beta + W_{ij}b_i) \tag{3.18}$$

where $h_{ij}(X_{ij}|b_i)$ is the conditional hazard function for the j^{th} observation of group i. The parameter β is the regression parameter associated to the vector of covariates Z_{ij} and b_i is the random effect associated to group i with corresponding covariates W_{ij}. The frailty b_i is simulated following a centered normal distribution of variance σ^2. The parameters of the model are $\theta = (\lambda, \rho, \beta, \sigma^2)$.

3.4.2 Definition of the MLE for the Weibull shared frailty model

We estimate the parameters of the model via the maximum of the marginal likelihood. The conditional likelihood for the i^{th} group is expressed as follows:

$$L_{\text{cond},i}(\theta; X_i, \Delta_i|b_i) = \prod_{j=1}^{J} \left(\lambda\rho X_{ij}^{\rho-1}\exp(Z_{ij}^t\beta + W_{ij}^t b_i)\right)^{\delta_{ij}} \exp(-\lambda X_{ij}^{\rho}\exp(Z_{ij}^t\beta + W_{ij}^t b_i)) \tag{3.19}$$

The complete likelihood for i^{th} group is therefore written as:

$$L_{\text{comp},i}(\theta; X_i, \Delta_i, b_i) = \prod_{j=1}^{J} \left(\lambda\rho X_{ij}^{\rho-1}\exp(Z_{ij}^t\beta + W_{ij}^t b_i)\right)^{\delta_{ij}} \exp(-\lambda X_{ij}^{\rho}\exp(Z_{ij}^t\beta + W_{ij}^t b_i))g_{\sigma^2}(b_i) \tag{3.20}$$

where g is the probability density of a centered normal distribution of variance σ^2.

The marginal likelihood for the i^{th} group is obtained by integrating the complete likelihood over the frailties $\mathbf{b}$:

$$L_{\text{marg},i}(\theta; X_i, \Delta_i) = \int L_{\text{comp},i}(\theta; X_i, \Delta_i, b_i)db_i \tag{3.21}$$

The maximum likelihood estimate (MLE) for parameter θ is the value $\hat{\theta}_{\text{MLE}}$ of θ which maximises the marginal likelihood defined in equation (3.4).

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\text{argmax}}\ L_{\text{marg}}(\theta; X, \Delta) \tag{3.22}$$

However, this is not an analytically tractable integral when the frailties are assumed to follow a normal distribution as in our case. We therefore use a SAEM-MCMC algorithm (cf. Kuhn and El-Nouty (2013)) to estimate the parameters of the model.

3.4.3 Criteria to evaluate the convergence rate

In this section, we define a criteria to quantify the convergence rates of MLEs following the one defined in Nie (2007). Let us denote by $\hat{\theta}_{\text{MLE}}$ the MLE of the model under consideration. If the estimate $\hat{\theta}_{\text{MLE}}$ is said to be asymptotically

gaussian at a rate $\sqrt{r_{N,J}}$, then :

$$\sqrt{r_{N,J}}\frac{\hat{\theta}_{\text{MLE}}-\theta_0}{\sqrt{\text{Var}(\hat{\theta}_{\text{MLE}})}} \xrightarrow[N,J\to\infty]{} U$$

where U is a centered gaussian random variable and θ_0 is the true unknown parameter (cf. Van der Vaart (2000)).

Our aim is to highlight possible different convergence rates $r_{N,J}$ depending on the parameters considered in the model. We recall three asymptotic scenarios encountered in the literature :

1. The number of groups N goes to infinity while the number of observations per group J remains finite.
2. N is finite while J goes to infinity.
3. Both N and J go to infinity.

The first case is the most common case encountered in survival analysis. It is also the simplest one. In this scenario, we expect the MLE of all parameters to be $\sqrt{N}-$consistent. We instead study the convergence rates of the MLE in the second case where the number of groups N is fixed and the number of observations J per group goes to infinity.

Along the lines of Nie (2007), we define for a positive integer Q the reduction in variance which is computed as follows :

$$R_V(\hat{\theta}_{NJQ},\hat{\theta}_{NJ}) = 1-\frac{\text{Var}(\hat{\theta}_{NJQ})}{\text{Var}(\hat{\theta}_{NJ})} \tag{3.23}$$

In the event that the estimator is $\sqrt{N}-$consistent, $R_V(\hat{\theta}_{NJQ},\hat{\theta}_{NJ})$ tends to 0 which translates to no added information with more observations per group. If the estimator is $\sqrt{NJ}-$consistent, $R_V(\hat{\theta}_{NJQ},\hat{\theta}_{NJ}) = 1-\frac{1}{Q}$. For instance doubling the number of observations per group corresponds to $Q=2$ and would lead to 50% reduction in variance.

In the next section, we define parametric shared frailty models with different covariate structures and different status for the frailty terms.

3.4.4 Simulation setting with different covariate structures

We study the consistency and the convergence rates of the MLE of parameters in different modeling approaches. For $i=1,\dots,N$, $j=1,\dots,J$, we specify three different models $\boldsymbol{\mathcal{M}}_1$, $\boldsymbol{\mathcal{M}}_2$ and $\boldsymbol{\mathcal{M}}_3$ as follows :

$$h_{ij}(X_{ij}|b_i) = \lambda\rho X_{ij}^{\rho-1}\,\exp(Z_{ij,1}\beta_1+Z_{i,2}\beta_2+b_i) \tag{$\boldsymbol{\mathcal{M}}_1$}$$

$$h_{ij}(X_{ij}|b_i) = \lambda\rho X_{ij}^{\rho-1} \exp(Z_{i,1}(\beta_1 + b_i) + Z_{ij,2}\beta_2) \qquad (\boldsymbol{\mathcal{M}}_2)$$

$$h_{ij}(X_{ij}|b_i) = \lambda\rho X_{ij}^{\rho-1} \exp(Z_{ij,1}(\beta_1 + b_i) + Z_{ij,2}\beta_2) \qquad (\boldsymbol{\mathcal{M}}_3)$$

We recall that $b_i \sim \mathcal{N}(0, \sigma^2)$ in all cases. In model $\boldsymbol{\mathcal{M}}_1$, we consider two types of covariates; $Z_{ij,1}$ that varies at the observation level and $Z_{i,2}$ at the group level. In model $\boldsymbol{\mathcal{M}}_2$, the covariate $Z_{i,1}$ varies at the group level and acts on the frailty terms while $Z_{ij,2}$ varies at the observation level. In model $\boldsymbol{\mathcal{M}}_3$, both covariates vary at the observation level with $Z_{ij,1}$ also acting on the frailty terms.

3.5 Numerical experiments on the convergence rates of MLEs

The simulation setting is set up as follows: The number of groups denoted by N is fixed at 200. The number of observations per group, denoted by J, varies from 2 to 64. The event times are simulated following a Weibull distribution with parameters $\lambda = 0.01$ and $\rho = 1.5$. The frailties are simulated following a centered normal distribution of variance $\sigma^2 = 0.7$. The covariates are simulated following a Bernoulli distribution and are perfectly balanced within the group for covariates that vary at the observation level so as to avoid confounding. The regression parameters associated to the covariates are $(\beta_1, \beta_2) = (2, 2)$. The data is simulated under three different censoring settings (no censoring, 40 $\%$ censoring and 70 $\%$ censoring). The parameters are estimated by maximisation of the marginal likelihood constructed with a Weibull baseline assumption. A SAEM-MCMC algorithm is used to estimate the MLE of the parameters of the model. 200 repetitions of the datasets are generated in each case and the empirical mean and variance of the estimates are computed.

3.5.1 Effects of covariates varying at group and observation levels

We first simulate and estimate the MLE of parameters following model $\boldsymbol{\mathcal{M}}_1$. In the model, there are two types of covariates; $Z_{ij,1}$ that varies at the observation level and $Z_{i,2}$ at the group level. The boxplots of estimates in Figure 3.1 give a visual representation of the results obtained. All estimates seem to be consistent since it appears there is no bias. The dispersion in the estimations for $\hat{\beta}_{1,\text{MLE}}$ and $\hat{\rho}_{\text{MLE}}$ clearly decreases when the group sizes increase whereas the dispersion stays mostly the same for the other estimates.

The reductions in variance when the group sizes are doubled are presented in Table 3.2. It seems that $\hat{\beta}_{1,\text{MLE}}$ and $\hat{\rho}_{\text{MLE}}$ are $\sqrt{NJ}$-consistent whereas $\hat{\beta}_{2,\text{MLE}}$, $\hat{\sigma}^2_{\text{MLE}}$ and $\hat{\lambda}_{\text{MLE}}$ are $\sqrt{N}$-consistent. The results are consistent with what is observed on the boxplots in Figure 3.1. The different rates observed for $\hat{\beta}_{1,\text{MLE}}$ and $\hat{\beta}_{2,\text{MLE}}$ are expected following

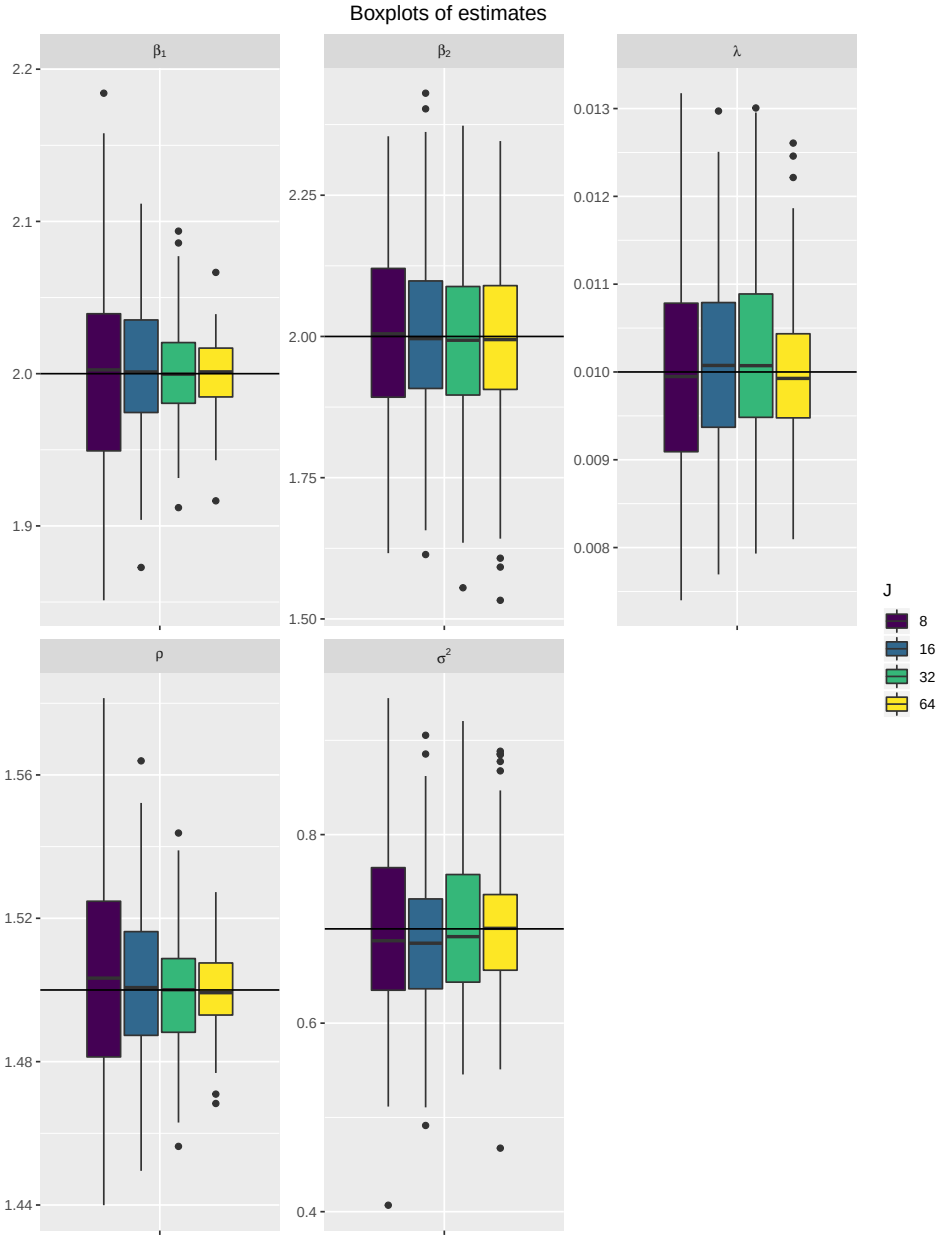

Figure 3.1: MLE of parameters from 200 repetitions of the datasets simulated following $h_{ij}(X_{ij}|b_i) = \lambda\rho X_{ij}^{\rho-1}\exp(Z_{ij,1}\beta_1 + Z_{i,2}\beta_2 + b_i)$. $N = 200, \beta_0 = (2,2), \sigma_0^2 = 0.7, \rho_0 = 1.5, \lambda_0 = 0.01$

Table 3.2: Reduction in variance based on empirical variances of the MLE of parameters from 200 repetitions of the datasets simulated following $h_{ij}(X_{ij}|b_i) = \lambda\rho X_{ij}^{\rho-1}\exp(Z_{ij,1}\beta_1 + Z_{i,2}\beta_2 + b_i)$. $N = 200, \beta_0 = (2,2), \sigma_0^2 = 0.7, \rho_0 = 1.5, \lambda_0 = 0.01$

Parameter	**Change in J**		
	J=8 $\rightarrow$ J=16	J=16 $\rightarrow$ J=32	J=32 $\rightarrow$ J=64
$\hat{\beta}_1$	0.557	0.520	0.471
$\hat{\beta}_2$	0.157	-0.0220	-0.102
$\hat{\sigma}^2$	0.302	0.0839	0.0903
$\hat{\rho}$	0.484	0.439	0.554
$\hat{\lambda}$	0.236	0.147	0.214

the proof established in the linear mixed-effects model. Regression parameter estimate $\hat{\beta}_{2,\text{MLE}}$ which is associated to a covariate at group level seems to converge at a slower rate than $\hat{\beta}_{1,\text{MLE}}$ which is associated to a covariate at observation level. Indeed, let us reparameterize model $\mathcal{M}_1$ with $\exp(\mu)$ instead of λ. The model can therefore be expressed as :

$$h_{ij}^{'}(X_{ij}|b_i) = \rho X_{ij}^{\rho-1} \exp(Z_{ij,1}\beta_1 + Z_{i,2}\beta_2 + b_i + \mu) \tag{3.24}$$

The conditional likelihood can therefore be written as :

$$L_{\text{cond},i}^{'}(\theta; X, \Delta \mid b_i) = \prod_{j=1}^{J} \left(\rho X_{ij}^{\rho-1} \exp(Z_{ij,1}\beta_1 + Z_{i,2}\beta_2 + (b_i + \mu))\right)^{\delta_{ij}}$$
$$\exp(-X_{ij}^{\rho} \exp(Z_{ij,1}\beta_1 + Z_{i,2}\beta_2 + (b_i + \mu)))$$

Following the reparameterization of the model, it is obvious that the frailty term b_i appears in the model only as an additive term with respect to the parameter μ.

Effect of censoring on MLEs in model $\mathcal{M}_1$

Table 3.3: Reduction in variance based on empirical variances of the MLE of parameters from 200 repetitions of the datasets simulated following $h_{ij}(X_{ij}|b_i) = \lambda\rho X_{ij}^{\rho-1} \exp(Z_{ij,1}\beta_1 + Z_{i,2}\beta_2 + b_i)$. $N = 200, \beta_0 = (2, 2), \sigma_0^2 = 0.7, \rho_0 = 1.5, \lambda_0 = 0.01$

Parameter	**Change in J**					
	Censoring	J=2 $\rightarrow$ J=4	J=4 $\rightarrow$ J=8	J=8 $\rightarrow$ J=16	J=16 $\rightarrow$ J=32	J=32 $\rightarrow$ J=64
$\hat{\beta}_1$		0.586	0.624	0.557	0.520	0.471
$\hat{\beta}_2$		0.238	0.0726	0.157	-0.0220	-0.102
$\hat{\sigma}^2$	0%	0.626	0.426	0.302	0.0839	0.0903
$\hat{\rho}$		0.564	0.700	0.484	0.439	0.554
$\hat{\lambda}$		0.640	0.537	0.236	0.147	0.214
$\hat{\beta}_1$		0.677	0.611	0.578	0.464	0.577
$\hat{\beta}_2$		0.479	0.361	-0.162	-0.0382	0.008
$\hat{\sigma}^2$	40%	0.774	0.527	0.280	0.276	0.0621
$\hat{\rho}$		0.721	0.601	0.657	0.522	0.482
$\hat{\lambda}$		0.728	0.606	0.494	0.192	0.277
$\hat{\beta}_1$		0.632	0.618	0.513	0.505	0.464
$\hat{\beta}_2$		0.547	0.375	0.318	0.102	0.217
$\hat{\sigma}^2$	70%	0.752	0.637	0.608	0.217	0.373
$\hat{\rho}$		0.690	0.585	0.555	0.563	0.497
$\hat{\lambda}$		0.626	0.580	0.455	0.336	0.260

We now investigate the effect of censoring on the MLEs calculated on model $\mathcal{M}_1$. From Table 3.3, it seems that $\hat{\beta}_{1,\text{MLE}}$ and $\hat{\rho}_{\text{MLE}}$ are $\sqrt{NJ}$-consistent whereas $\hat{\beta}_{2,\text{MLE}}$, $\hat{\sigma}^2_{\text{MLE}}$ and $\hat{\lambda}_{\text{MLE}}$ are $\sqrt{N}$-consistent. We observe the same convergence rates for the MLEs in different censoring settings. The effect of censoring is more pronounced in cases where the number of observations per group is small ($J = 2$ and $J = 4$). A higher censoring percentage leads to

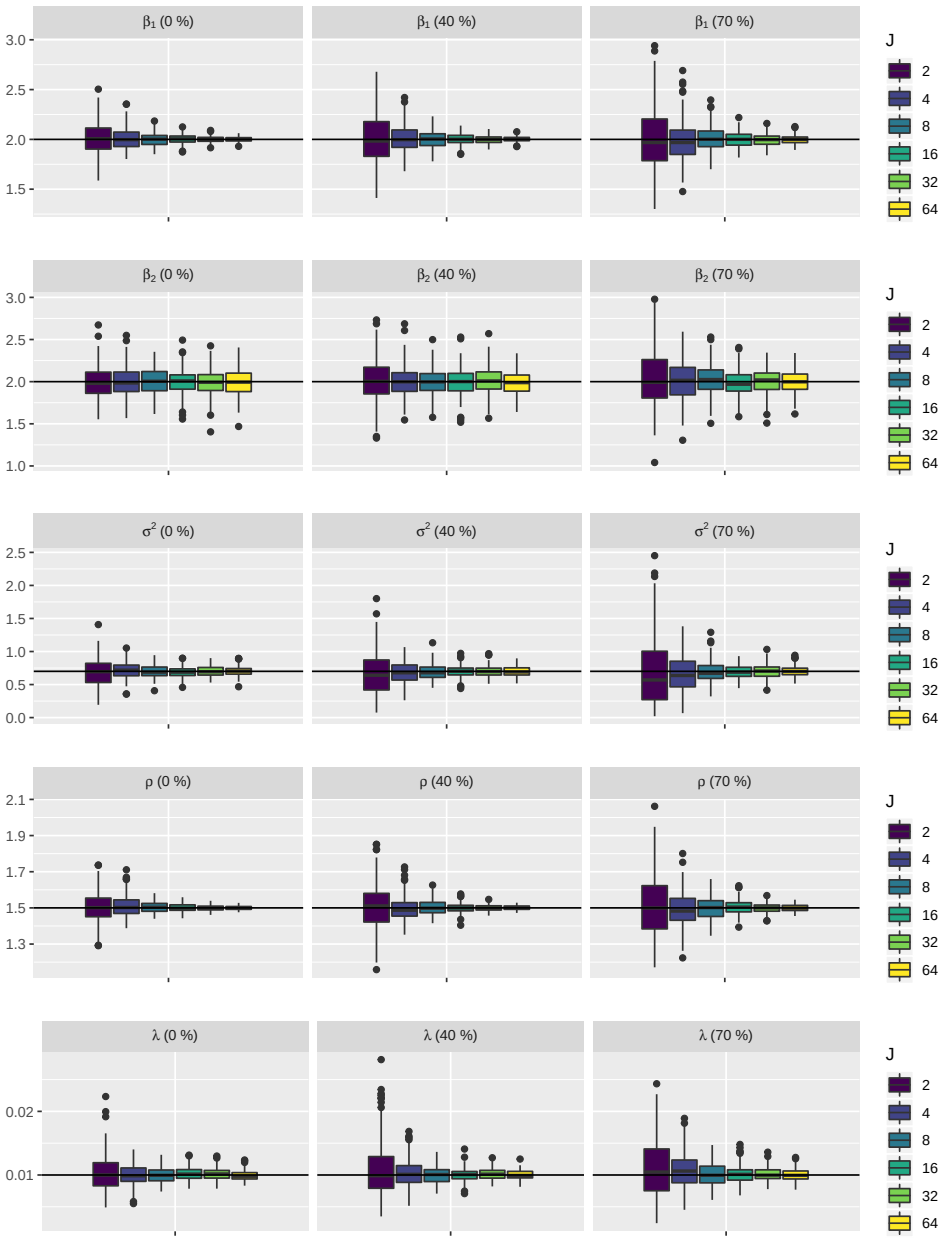

Figure 3.2: MLE of parameters from 200 repetitions of the datasets simulated following $h_{ij}(X_{ij}|b_i) = \lambda\rho X_{ij}^{\rho-1}\exp(Z_{ij,1}\beta_1 + Z_{i,2}\beta_2 + b_i)$. $N = 200, \beta_0 = (2,2), \sigma_0^2 = 0.7, \rho_0 = 1.5, \lambda_0 = 0.01$

Table 3.4: Reduction in variance based on empirical variances of the MLE of parameters from 200 repetitions of the datasets simulated following $h_{ij}(X_{ij}|b_i) = \lambda\rho X_{ij}^{\rho-1}\exp(Z_{i,1}(\beta_1+b_i)+Z_{ij,2}\beta_2)$. $N=200, \beta_0=(2,2), \sigma_0^2=0.7, \rho_0=1.5, \lambda_0=0.01$.

Parameter	**Change in J**		
	J=8 $\rightarrow$ J=16	J=16 $\rightarrow$ J=32	J=32 $\rightarrow$ J=64
$\hat{\beta}_1$	0.177	0.153	0.0183
$\hat{\beta}_2$	0.468	0.586	0.478
$\hat{\sigma}^2$	0.416	0.188	0.291
$\hat{\rho}$	0.485	0.578	0.475
$\hat{\lambda}$	0.442	0.429	0.351

higher standard deviations as indicated by the wider boxplots (cf. Figure 3.2). This is not very surprising as more censoring results in a lesser contribution to the likelihood and less precise estimations. We note that as the number of observations per group increases, the censoring effect is less pronounced.

3.5.2 Effect of a covariate at group level with an additive frailty term on the associated regression parameter

We consider Model ($\mathcal{M}_2$) :

$$h_{ij}(X_{ij}|b_i) = \lambda\rho X_{ij}^{\rho-1}\exp(Z_{i,1}(\beta_1+b_i)+Z_{ij,2}\beta_2)$$

The boxplots of the estimates obtained are shown in Figure 3.3. The reductions in variance are displayed in Table 3.4. We note that there is about 50 $\%$ reduction in variance for estimates $\hat{\beta}_{2,\text{MLE}}$, $\hat{\rho}_{\text{MLE}}$ and $\hat{\lambda}_{\text{MLE}}$ when the number of observations per group doubles. The reduction in variance is less consequent for $\hat{\beta}_{1,\text{MLE}}$ and $\hat{\sigma}^2_{\text{MLE}}$. This suggests that $\hat{\beta}_{2,\text{MLE}}$, $\hat{\rho}_{\text{MLE}}$ and $\hat{\lambda}_{\text{MLE}}$ are $\sqrt{NJ}$-consistent whereas $\hat{\beta}_{1,\text{MLE}}$ and $\hat{\sigma}^2_{\text{MLE}}$ are $\sqrt{N}$-consistent. We make the same observations in the two censoring settings of 40 $\%$ and 70 $\%$ (cf. Figure B.1 and Table B.1 in Appendix B). As was the case in the previous model analysed, the censoring effect is more pronounced when the number of observations per group is small.

We note here that the $\sqrt{N}$-consistency of $\hat{\beta}_{1,\text{MLE}}$ could be explained by both the fact that the covariate is at the group level and also by the frailty acting as an additive term on parameter β_1. The conditional likelihood can be factorized as follows :

$$L_{\text{cond},i}(\theta; X, \Delta|b_i) = \prod_{j=1}^{J}\left(\lambda\rho X_{ij}^{\rho-1}\exp(Z_{i,1}(\beta_1+b_i)+Z_{ij,2}\beta_2)\right)^{\delta_{ij}}$$
$$\exp(-\lambda X_{ij}^{\rho}\exp(Z_{i,1}(\beta_1+b_i)+Z_{ij,2}\beta_2))$$

Following the factorization, we conjecture that $\hat{\beta}_{1,\text{MLE}}$ is $\sqrt{N}$-consistent.

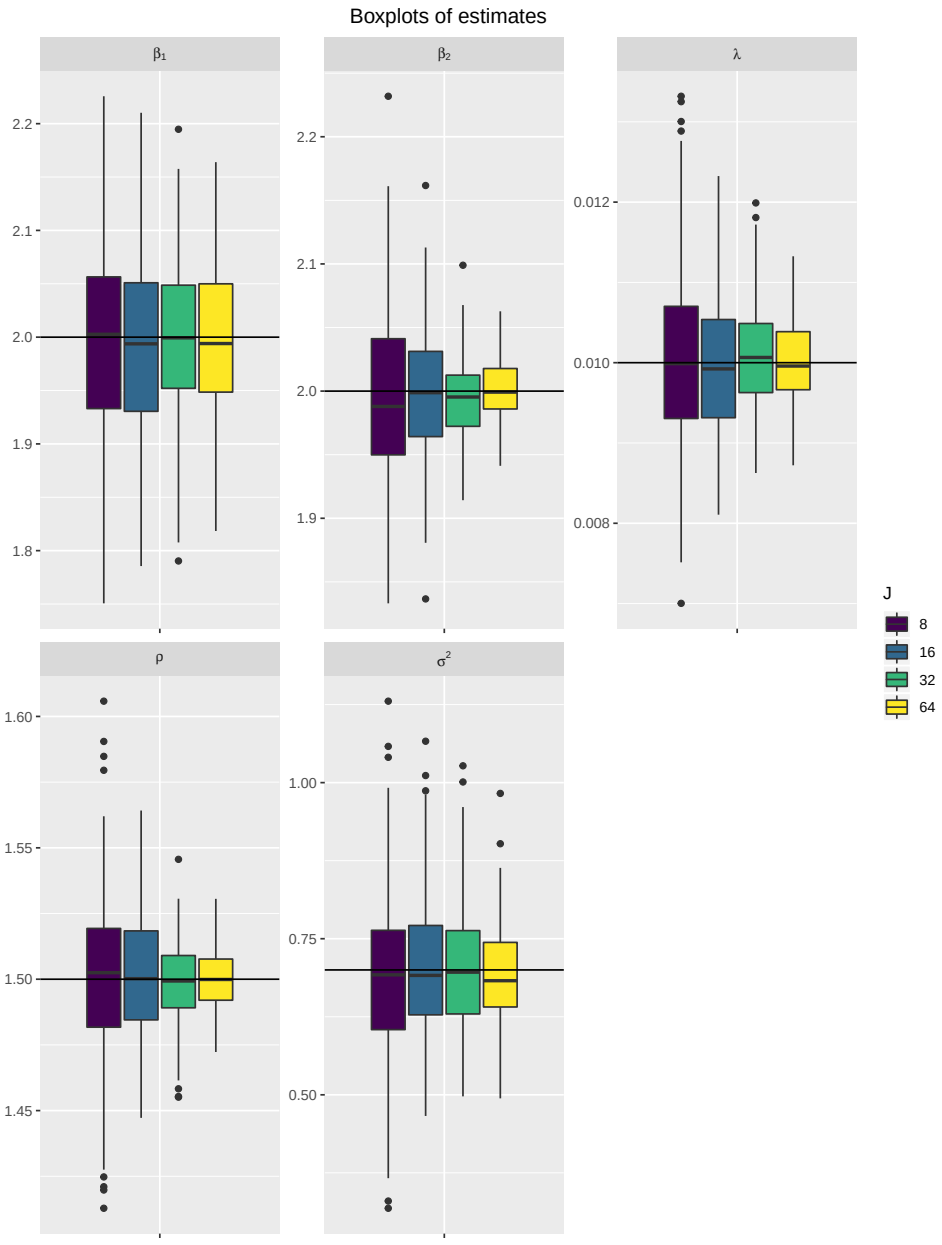

Figure 3.3: MLE of parameters from 200 repetitions of the datasets simulated following $h_{ij}(X_{ij}|b_i) = \lambda\rho X_{ij}^{\rho-1}\exp(Z_{i,1}(\beta_1+b_i)+Z_{ij,2}\beta_2)$. $N=200, \beta_0=(2,2), \sigma_0^2=0.7, \rho_0=1.5, \lambda_0=0.01$.

3.5.3 Effect of a covariate at observation level with an additive frailty term on the associated regression parameter

We consider Model $\mathcal{M}_3$.

$$h_{ij}(X_{ij}|b_i) = \lambda\rho X_{ij}^{\rho-1} \exp(Z_{ij,1}(\beta_1 + b_i) + Z_{ij,2}\beta_2)$$

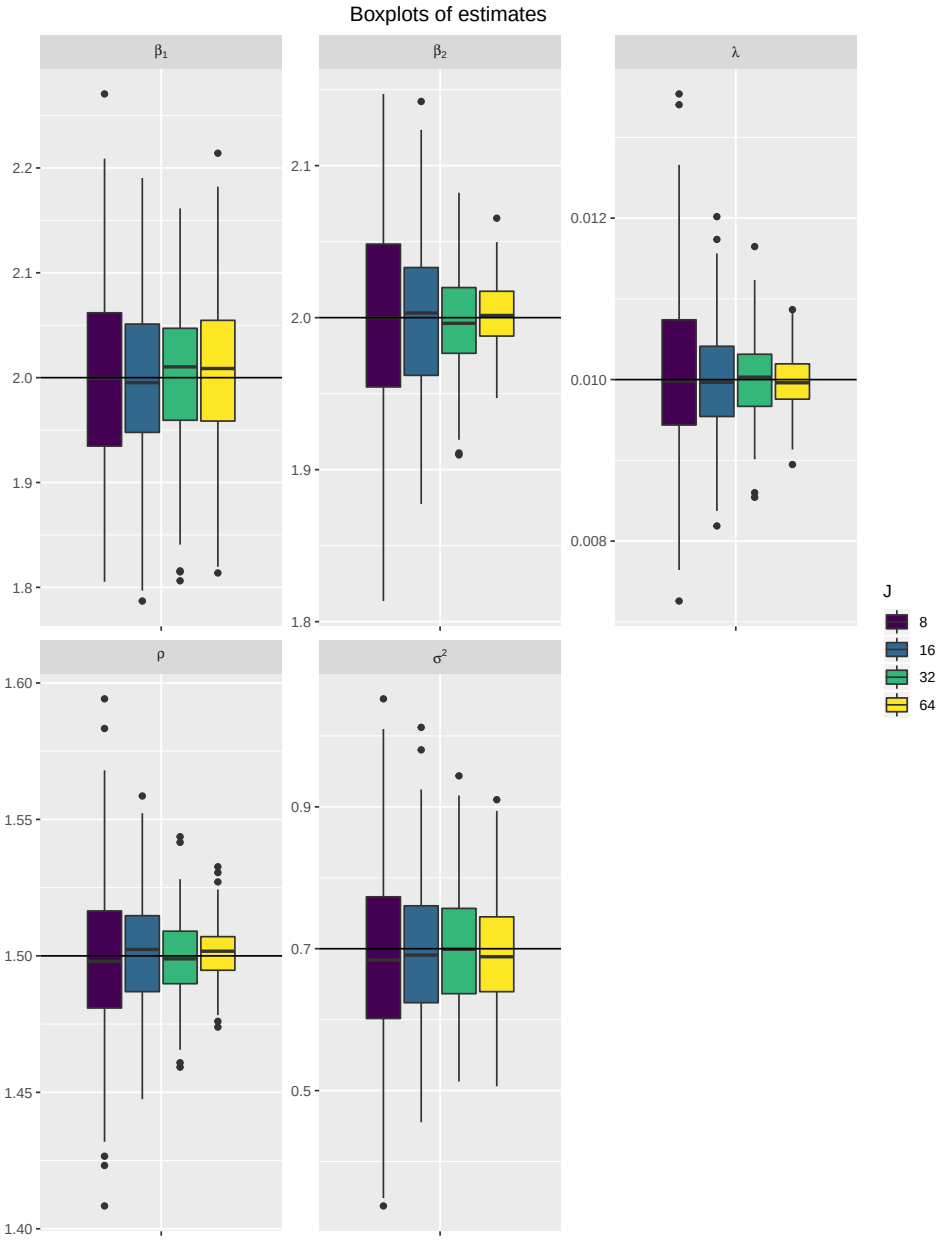

Figure 3.4: MLE of parameters from 200 repetitions of the datasets simulated following $h_{ij}(X_{ij}|b_i) = \lambda\rho X_{ij}^{\rho-1} \exp(Z_{ij,1}(\beta_1 + b_i) + Z_{ij,2}\beta_2)$. $N = 200, \beta_0 = (2, 2), \sigma_0^2 = 0.7, \rho_0 = 1.5, \lambda_0 = 0.01$.

The boxplots of the estimates obtained are shown in Figure 3.4. The reductions in variance are displayed in

Table 3.5. From Figure 3.4, it seems that the dispersion of the estimates decreases with an increasing number of observations per group for all parameters except for $\hat{\beta}_{1,\text{MLE}}$ and $\hat{\sigma}^2_{\text{MLE}}$. We come to the same conclusion under 40 % and 70 % censoring as shown in Figure B.2 of appendix B.

Table 3.5: Reduction in variance based on empirical variances of the MLE of parameters from 200 repetitions of the datasets simulated following $h_{ij}(X_{ij}|b_i) = \lambda\rho X_{ij}^{\rho-1}\exp(Z_{ij,1}(\beta_1 + b_i) + Z_{ij,2}\beta_2)$. $N = 200, \beta_0 = (2, 2), \sigma_0^2 = 0.7, \rho_0 = 1.5, \lambda_0 = 0.01$.

Parameter	**Change in J**		
	J=8 $\rightarrow$ J=16	J=16 $\rightarrow$ J=32	J=32 $\rightarrow$ J=64
$\hat{\beta}_1$	0.160	0.198	-0.109
$\hat{\beta}_2$	0.380	0.516	0.551
$\hat{\sigma}^2$	0.479	0.254	0.0727
$\hat{\rho}$	0.548	0.460	0.515
$\hat{\lambda}$	0.529	0.449	0.521

From Table 3.5, it seems that $\hat{\beta}_{2,\text{MLE}}$, $\hat{\rho}_{\text{MLE}}$ and $\hat{\lambda}_{\text{MLE}}$ are $\sqrt{NJ}$-consistent while $\hat{\beta}_{1,\text{MLE}}$ and $\hat{\sigma}^2_{\text{MLE}}$ are $\sqrt{N}$-consistent. It is interesting to note that $\hat{\beta}_{1,\text{MLE}}$ seems to be $\sqrt{N}$-consistent even if $Z_{ij,1}$ is at the observation level. The conditional likelihood can be expressed as :

$$L_{\text{cond},i}(\theta; X, \Delta|b_i) = \prod_{j=1}^{J}\Big(\lambda\rho X_{ij}^{\rho-1}\exp(Z_{ij,1}(\beta_1 + b_i) + Z_{ij,2}\beta_2)\Big)^{\delta_{ij}}$$
$$\exp(-\lambda X_{ij}^{\rho}\exp(Z_{ij,1}(\beta_1 + b_i) + Z_{ij,2}\beta_2))$$

Following the form of the conditional likelihood, it is clear that the frailty acts as an additive term on parameter β_1. Simulation results point to $\hat{\beta}_{1,\text{MLE}}$ being $\sqrt{N}$-consistent. The effect of the additive frailty term on β_1 therefore takes precedence over the structure of $Z_{ij,1}$ to determine the convergence rate.

3.5.4 Effect of the between-group heterogeneity on the estimates

In this section, we investigate the effect of the between-group heterogeneity on the MLEs. Considering the signal processing context, where the signal-to-noise ratio allows to compare the level of a signal to the level of the background noise, one can imagine a signal-to-between-group heterogeneity ratio in fraitly models which would compare the amplitude of the effects of covariates to the between-group heterogeneity level. Heterogeneity between groups may be here quantified by the variance σ^2 of the frailty distribution. In the previous simulation study, the variance of the frailty distribution was fixed such that $\sigma^2 = 0.7$. We propose to analyze the effect of a smaller between-group heterogeneity, say $\sigma^2 = 0.1$, on the MLEs.

Let us simulate data following model $\mathcal{M}_1$. We keep the same simulation setting as in the previous sections and simulate data for two different values of σ^2 namely 0.7 and 0.1. There is no censoring. We present the results of the parameter estimation in Figure 3.5. The left column and right column correspond to estimates when $\sigma^2 = 0.7$

and $\sigma^2 = 0.1$ respectively. The convergence rates of the MLEs appear similar in both cases. This agrees with the conjecture that the convergence rates of parameters are fully determined by the expression of the likelihood in particular by the way the frailty terms act on the parameters and by the structure of covariates. There seems to be no obvious differences in the standard deviations of the estimates of β_1 and ρ depending on the value of σ^2. However, for the estimates of β_2, σ^2 and λ, the standard deviation seems to be smaller in general when the between-group heterogeneity is smaller. Thus, it seems that the between-group heterogeneity influences the accuracy of the estimates of parameters that are $\sqrt{N}$-consistent. More precisely, the smaller the between-group heterogeneity is, the smaller the standard deviations. Thus, this effect of the between-group heterogeneity corresponds to the effect of the noise level in classical regression.

3.6 Conclusion and perspectives

In this chapter, we studied the consistency and convergence rates of MLEs in parametric shared frailty models. We recall that there are no asymptotic results established yet for MLEs in the framework of parametric shared frailty models. Based on existing results in other types of models, we studied the consistency and the convergence rates of MLEs in a Weibull shared frailty model. Through a simulation study, we demonstrated that the parameterization of the conditional likelihood plays a key role in determining the convergence rates of MLEs as well as the structure of covariates. Therefore, we proposed to compare two differently structured covariates; those varying at the observation level and those varying at the group level. Bearing that in mind, we established theoretically the different convergence rates of MLEs in a simple linear mixed-effects model depending on two structures of covariates. In particular, MLEs of parameters associated to covariates at the observation level and MLEs of parameters associated to covariates at the group level do not converge at the same rates. Establishing theoretically such kind of results in parametric shared frailty models requires carrying out complex calculations and handling of quantities expressed as integrates with respect to the frailty distribution. Therefore, we were not able to manage this for the moment. Thus, we conducted an intensive simulation study to extend the results to parametric shared frailty models.

On the basis of the conclusive simulation study, we state the following conjecture for parametric shared frailty models. The MLE of frailty distribution parameters are $\sqrt{N}$-consistent. The convergence rates of MLEs of baseline hazard parameters are determined with respect to the parameterization of the likelihood. If the baseline parameter is involved in the conditional likelihood with an additive frailty term, then the MLE is $\sqrt{N}$-consistent. Otherwise the MLEs of the baseline parameters are $\sqrt{NJ}$-consistent. Concerning the regression parameters, the convergence rates are determined with respect to the parameterization of the likelihood and the structure of covariates. If the regression parameter is involved in the conditional likelihood with an additive frailty term, then the MLE of this regression parameter is $\sqrt{N}$-consistent whatever the structure of the corresponding covariates. If the regression parameter is not subject to an additive frailty term, then the convergence rate is determined with respect to the

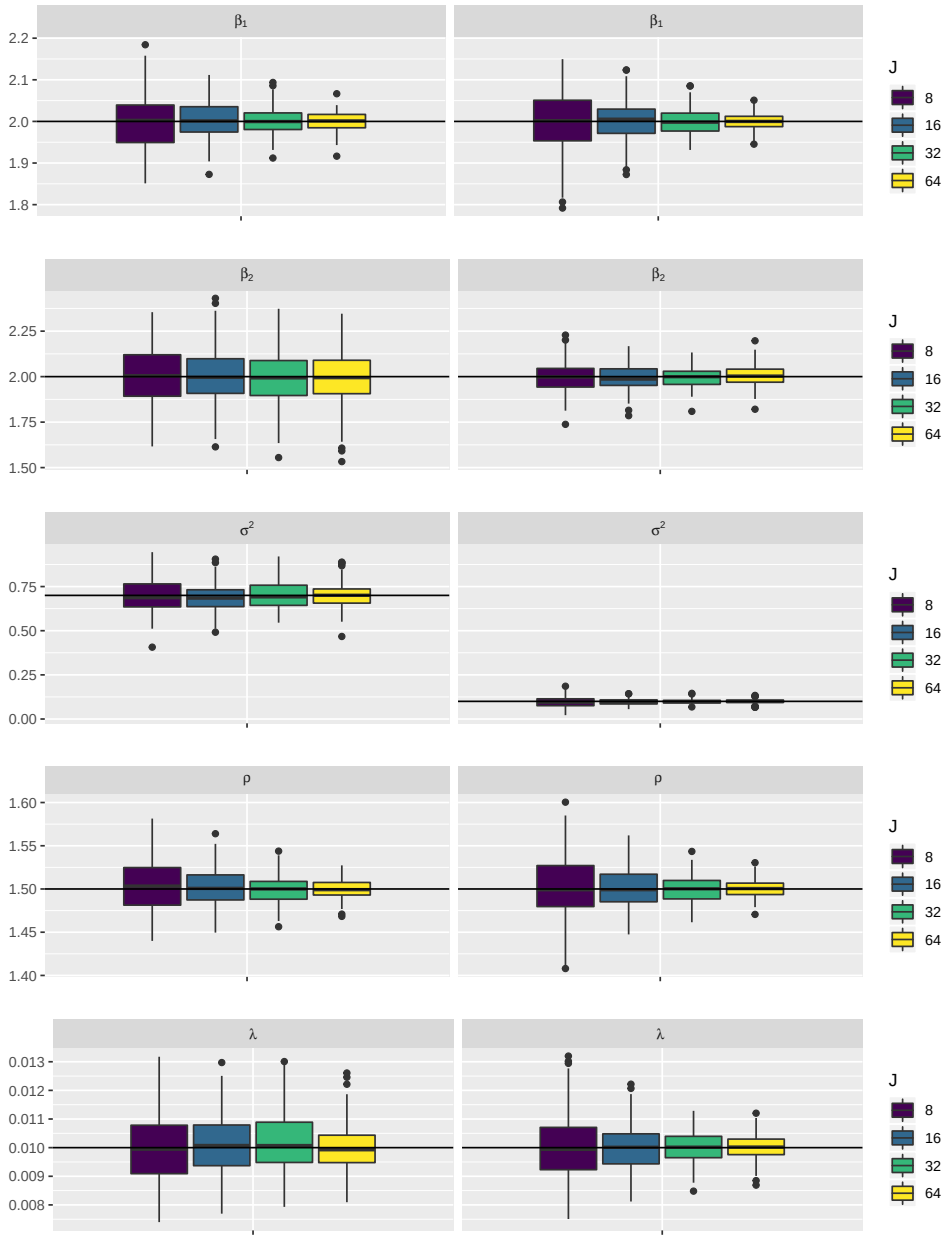

Figure 3.5: MLE of parameters from 200 repetitions of the datasets simulated following $h_{ij}(X_{ij}|b_i) = \lambda\rho X_{ij}^{\rho-1}\exp(Z_{ij,1}(\beta_1 + b_i) + Z_{ij,2}\beta_2)$. $N = 200, \beta_0 = (2,2), \rho_0 = 1.5, \lambda_0 = 0.01$. $\sigma_0^2 = 0.7$ (left column) and $\sigma_0^2 = 0.1$ (right column)

corresponding covariate structure. Namely, if the corresponding covariate is at the group level, the MLE is $\sqrt{N}$-consistent, whereas it is $\sqrt{NJ}$-consistent if the corresponding covariate is at the observation level. The conjecture is supported by simulation results based on a Weibull shared frailty model.

The encouraging simulation results motivate the development of a theoretical proof of this conjecture in general frailty models. Also, intensive simulation studies have to be conducted in other frailty models. Furthermore, considering the results obtained by Nie (2007) on the asymptotic normality of the MLE in generalized linear and nonlinear mixed-effects models, it would be interesting to investigate asymptotic normality of the MLE in frailty models, both by simulation study and also theoretically.

Concerning the mixed-effects models, we highlighted on a simple linear example the key role of the structure of covariates for determining the rate of convergence of the MLE of the parameters. It would be interesting to study more deeply through simulation and theoretically the effect of the structure of covariates on the convergence rate in general mixed-effects models.

Chapter 4

Estimation in a spatially correlated frailty model : application to malaria data

4.1 Introduction

This chapter deals with modeling time to malaria infection while accounting for spatial correlation present in the data. Malaria is a mosquito-borne infectious disease that affects humans. It remains a disease with high morbidity and mortality. Malaria dynamics are complex with a lot of factors influencing the transmission of the disease. With the mosquito acting as vector of the disease and the transmission occurring between infected and non-infected individuals depending on the proximity between them, the spatial correlation between individuals plays a crucial role. This highly motivates the development of spatially correlated frailty models to take into account the distance between individuals when studying the incidence of malaria and assessing other factors that may be of relevance such as age, distance to water bodies and seasons (rainy, moderate, dry). There is limited literature on spatial survival models. In addition to modeling the malaria data, we develop a model that is relevant in a wide range of applications.

In the first section, we give an overview of the malaria situation worldwide and more specifically in Ethiopia. In the second section, we describe the time to malaria data collected around the Gilgel Gibe hydroelectric dam and review in the third section the previous analyses done on the dataset that are already published. In the fourth section, we review existing models and estimation methods for spatially correlated survival data. In the fifth section, we propose to model survival data while taking into account the spatial correlation between all locations.

4.2 The malaria disease

4.2.1 Malaria as a worldwide phenomenon

Malaria is a disease that dates back to prehistoric times. Today, it is absent in a significant part of the developed world although the vector might still be prevalent. The transmission of the disease has been eliminated in the 1950s in America and in the 1970s in Europe. Since then, it has been mostly imported into the malaria-free regions by people coming from endemic regions and the efficient response to sporadic cases has prevented a resurgence. However, it is still a major health issue in the African continent and parts of Asia and South America. According to the World Malaria Report 2018 (cf. WHO (2018)) established by the World Health Organization (WHO), the estimated number of malaria cases worldwide was 219 million in 2017. Most cases (200 million or 92%) were recorded in Africa, far ahead of the Southeast Asia region (5%) and the Eastern Mediterranean region which accounts for only 2% of the cases. Globally, the number of deaths from malaria has been estimated at 435,000 compared to 451,000 in 2016 and 607,000 in 2010. Children under the age of 5 are the most vulnerable as shown in Figure 4.1. In 2017, they represented 61% (266,000) of deaths associated with malaria worldwide. The WHO Africa region alone recorded 93% of malaria-related deaths worldwide in 2017. Data for 2015-2017 shows no significant progress towards a decrease in the number of malaria cases worldwide.

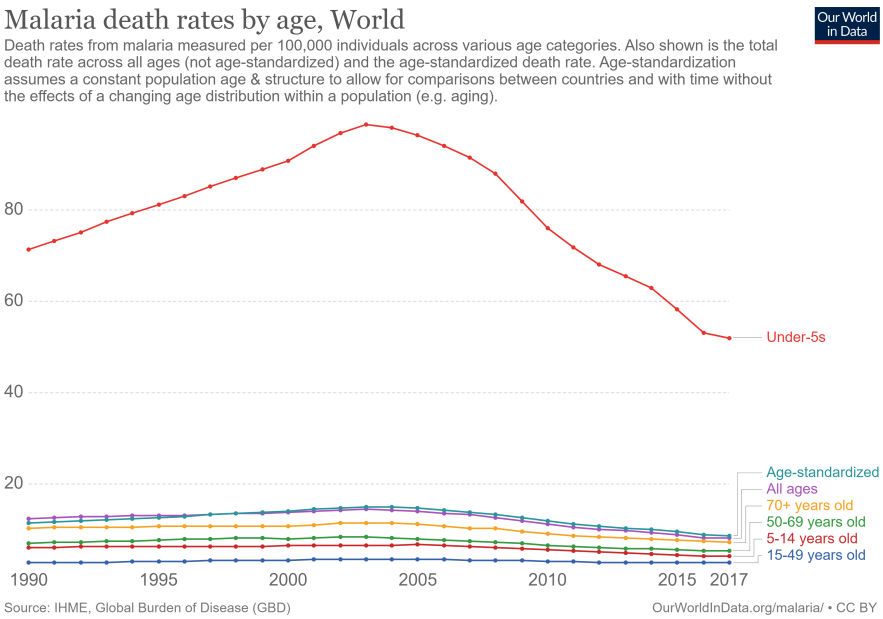

Figure 4.1: Malaria death rates by age (cf. https://ourworldindata.org/burden-of-disease)

Figure 4.2 clearly shows that most of the African continent, especially south of the Sahara, still has a malaria problem. There are many factors that contribute to the persistence of the disease in this region. A combination of insufficient health services, optimal climatic conditions for the breeding of the vector of the disease and a poorer

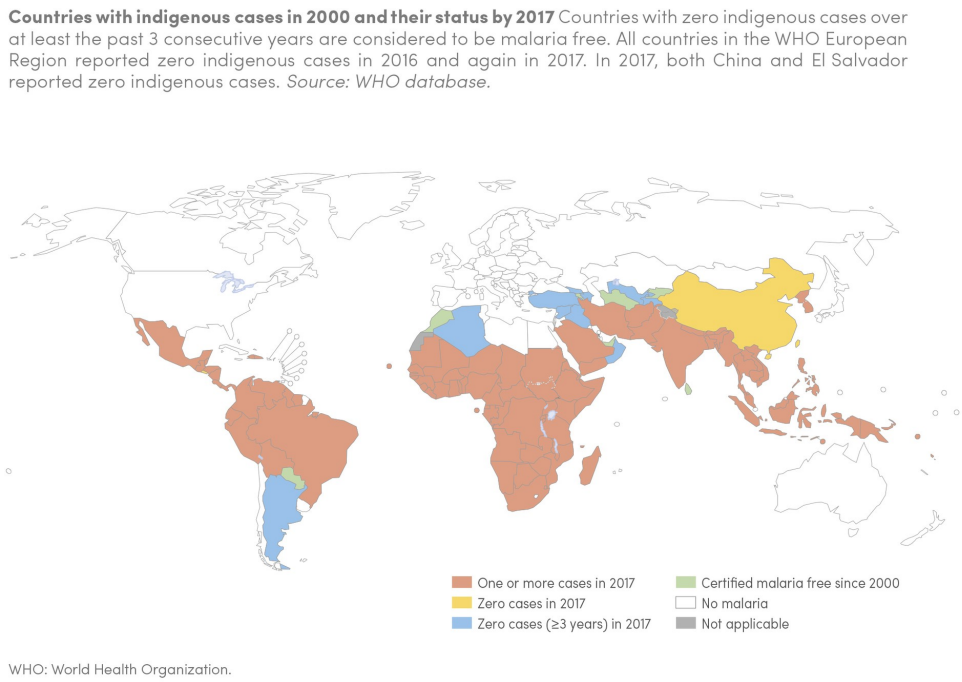

Figure 4.2: Malaria worldwide status from 2000 to 2017 (cf.World Health Organization (WHO) database)

population has contributed to the high prevalence of malaria in the region.

4.2.2 Malaria in Ethiopia

We will now focus on the situation of malaria in Ethiopia. According to the most recent information available on the WHO website, about 66 % of the population in Ethiopia is at risk of contracting the disease. For a country of 112 million people, this would put the number of people at risk at about 74 million. *Plasmodium falciparum* and *Plasmodium vivax* are the most dominant malaria parasites, distributed all over the country and account for 60 % and 40 % of cases respectively.

Our data have been collected in villages surrounding the Gilgel-Gibe hydroelectric dam reservoir in the Oromia region (cf. Yewhalaw et al. (2013)). In order to get a clearer picture of the malaria situation in that region, we demonstrate how the incidence of the disease has evolved in recent years. In Figure 4.3, the heat map of Ethiopia shows how the incidence varies across the country. The number of new cases is higher in the Oromia region when compared to the rest of the country. The graph in Figure 4.3 shows how the incidence varies from 2000 to 2017 in the region. There is a general decreasing trend from 2005 despite the slight increase in the number of new cases after 2015. There are various reasons as to why incidence is heterogeneous throughout the country. Altitude and climate both play a vital role in malaria transmission. Ethiopia is geographically diverse when it comes to elevation above sea level. Highland areas above 2500m altitude are malaria free. Areas in the 1500-2500m range are prone to frequent epidemics. Areas below 1500m are affected on a seasonal basis and the rest of the country faces a

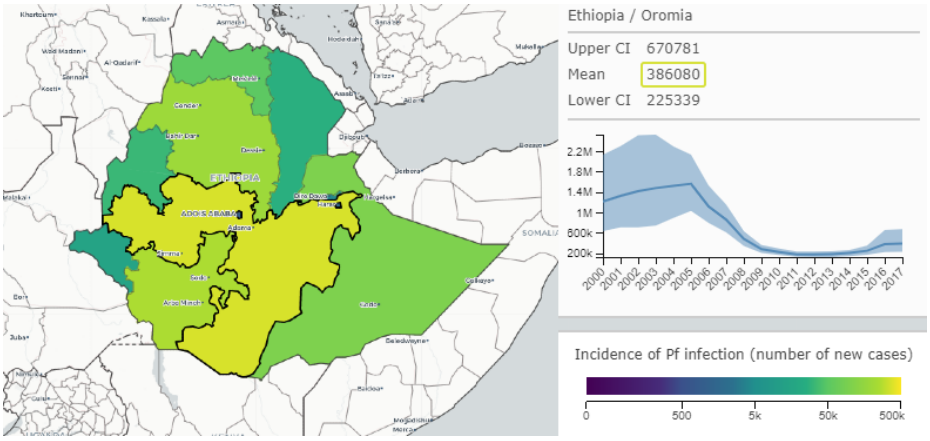

Figure 4.3: Malaria incidence (number of new cases) due to *Plasmodium falciparum* in 2017 in Ethiopia with summary statistics for the Oromia region

year round transmission of the disease. There are three main seasons namely the long rainy season from July to September, the dry season that runs from October to March and a moderate rainy season from April to June. For more information on the malaria problematic in Ehiopia, see Taffese et al. (2018) and Girum et al. (2019).

4.2.3 Transmission, diagnosis and treatment

Transmission

Malaria is caused by the protozoan parasite Plasmodium. Human malaria is caused by four different species of Plasmodium: *P. falciparum*, *P. malariae*, *P. ovale* and *P. vivax*. *Plasmodium falciparum* and *Plasmodium vivax* are the most common types of malaria parasites that infect humans and the former causes the most serious, life-threatening infections in humans. *Plasmodium falciparum* is the most prevalent malaria parasite in the WHO African Region. The malaria parasite is mainly transmitted at night when bitten by an infected female mosquito of the genus Anopheles. There are more than 400 different species of Anopheles mosquito; around 30 are malaria vectors of major importance. Malaria spreads when a mosquito becomes infected with the disease after biting an infected person, and the infected mosquito can then transmit the parasite to another person. The malaria parasites enter that person's bloodstream and travel to the liver. When the parasites mature, they leave the liver and infect red blood cells. A visual representation of how the disease is transmitted from mosquito to human to mosquito is given in Figure 4.4.

Malaria is generally not spread directly from person to person. However, in some rare cases malaria has been spread through blood transfusions and the sharing of needles. The main culprit is therefore the mosquito that acts as vector of the disease. Studying the mosquito population in affected areas can give great insights on the propagation of the disease. External factors such as the season (dry or rainy), presence of water bodies or humidity level play an important role on the incidence of the disease. Anopheles mosquitoes breed in natural water collections. During

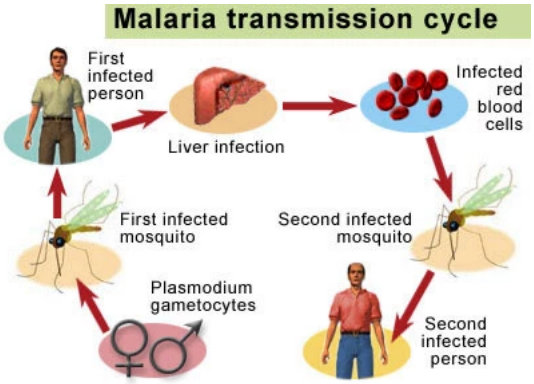

Figure 4.4: Malaria transmission schema (https://www.mayoclinic.org/)

the rainy season, breeding increases as water bodies act as prime breeding spots. Rooftops, ponds, wells, dams all play a part in increasing the mosquito population. It takes only about 1 week for the eggs to develop into adults. The anopheles mosquito has an average life span of 2-3 weeks. During that time, the mosquitoes can cover a region of radius 2 km.

The environment plays a crucial role in the transmission of the disease. Even in malaria-endemic countries, malaria transmission does not occur in all parts of the country. For example, the disease is not transmitted at very high altitudes, during colder seasons in some areas and in deserts (excluding the oases). Temperature is particularly important as below 20 degrees Celcius, *Plasmodium falciparum* (which causes severe malaria) cannot complete its growth cycle in the Anopheles mosquito, and thus cannot be transmitted.

Diagnosis

In this section, we briefly describe the different tests and treatments available for malaria. This information is based on publicly available data on the World Health Organization (WHO) and Centers for Disease Control and Prevention (CDC) websites. Most of the time malaria is diagnosed on the basis of the symptoms of the individual. For example, if a child living in a region with a high prevalence of malaria develops fever it is often assumed that they have malaria and they will be taken to a local hospital for treatment. Unfortunately, the regions with a high prevalence of malaria are often poor and access to proper health care is not often available. The infected children cannot get the required treatment in time and also pose a health risk to other children living in the vicinity.

Rapid diagnostic tests (RDTs)

RDTs are relatively simple to perform and interpret, they rapidly provide results and require limited training. Those characteristics make it a popular device for the diagnosis of malaria. There are more than 200 different variants but all of them follow more or less the same principles. A sample of blood usually obtained from a finger-prick is required and the results are available in about 30 minutes. In case of detection, some RDTs can even distinguish

between the types of parasites involved. Some can detect only a single species while some can detect multiple species. RDTs are very common especially in sub-Saharan Africa where a total of 276 million were sold in 2017 according to the world malaria report 2018. In the same year, 75% of malaria tests were conducted using RDTs.

Microscopy diagnosis

Microscopy diagnosis is performed in a laboratory by trained personnel. A sample of the patient's blood is spread out as a blood smear on a microscope slide. Two types of blood smears are analysed, thick and thin. The thick blood smear can tell whether the malaria parasite is present and the thin blood smear enables the laboratorian to identify the type of parasite involved. The test is therefore a visual one. The result of the test is obtained after viewing the blood smears in the laboratory and is therefore not readily available as in the case of RDTs.

Polymerase chain reaction (PCR)

Polymerase chain reaction (PCR) is a method widely used in molecular biology to make several copies of a specific DNA segment. DNA can be extracted from a patient's blood sample, copied and then analyzed to detect the presence or absence of malaria parasites. The PCR method is the most sensitive and specific but also the most expensive one.

As stated above, the most popular diagnosis method in the sub-Saharan region is the use of RDTs. Out of the three methods, it is the simplest, least expensive and requires only limited training. Microscopy often leads to over or under diagnosis in laboratories where the technicians don't have the proper training. For cost reasons, PCR is not common in poor regions even if it is the best of the three methods. More information and a detailed comparison between the methods can be found in Mfuh et al. (2019).

Treatment

Availability and type of treatment of malaria after a positive diagnosis varies widely depending on the country or region. Each country has specific national guidelines as to how to treat the disease. Common to most countries is the differentiation between uncomplicated malaria and severe malaria. As the name suggests, uncomplicated malaria is the milder one and can be treated on an outpatient basis. Symptoms include one or more of the following: fever, chills, sweats, headaches, muscle pains, nausea and vomiting. Severe malaria however is of a more serious nature and requires hospitalisation of the patient and immediate treatment. Symptoms in this case include confusion, coma, severe anemia, and respiratory difficulties. The parasite mostly responsible for severe malaria is *Plasmodium falciparum* making it the most dangerous parasite among those responsible for the disease. The WHO recommends that patients in malaria-endemic areas be treated within 24 hours after their first symptoms appear.

On top of national guidelines, treatment also depends on the type of parasite responsible for the disease in the patient and the clinical status of the patient (pregnancy, drug allergies, other medications being taken by the patient). The location where the infection occurred can be of particular importance as some regions present antimalarial drug resistance. Some of the most common drugs prescribed and approved by the WHO are artemesinin-based combination treatments, chloroquine, doxycycline, mefloquine and quinine. However, chloroquine and mefloquine, are no longer effective in many parts of the world. In cases of severe malaria, oral intake of the mentioned drugs might be complicated. In those cases, intravenous treatment is required and must be followed up with a complete course of oral antimalarial drugs.

4.3 The Gilgel Gibe malaria dataset

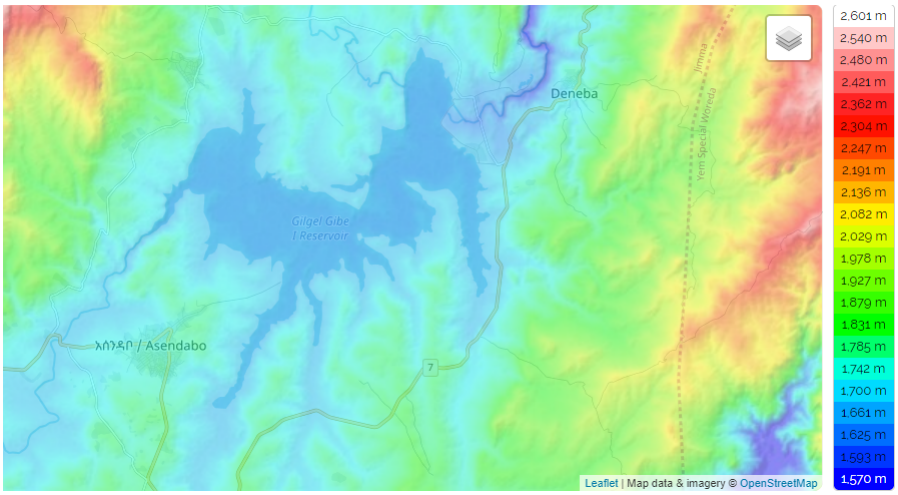

Figure 4.5: Elevation map of the the study area obtained using the website https://en-gb.topographic-map.com/

In 2003, the Gilgel Gibe hydroelectric dam in South West Ethiopia came into production. The dam covers an area of about 62 km^2 and is at an altitude of 1734-1864 metres above sea level as shown in Figure 4.5. The designated region is prone to frequent epidemics and seasonal patterns also play an important role in malaria transmission. In order to study the effect of the water reservoir on malaria, 16 villages at different distances (between 0.26 and 9.05 km) from the dam were randomly selected (cf. Yewhalaw et al. (2013)). The villages all have similar access to health facilities and are socioeconomically similar. Also, there were no major water bodies close to the villages except for the Gilgel Gibe hydroelectric dam. In total, 2080 children below 10 years were followed up on a weekly basis between July 2008 and June 2010 for the presence of malaria, resulting in a detailed time to malaria infection dataset. The location of the different households around the lake are shown in Figure 4.6. The number of households included in the study per village is about 130. In each house, only one child was randomly selected to be part of the study and therefore followed on a weekly basis. During each visit, the temperature of the child was taken and the parents were also asked whether the child had fever in the preceding week. For children with (a recall of) fever, a blood sample is

taken to assess whether the malaria parasite is present by microscopy. The event being considered here is the time to malaria infection. The data is interval censored in this case. If a child tests positive for malaria, the exact time of infection is some day between the current visit and the last one; we use the midpoint imputation. All the children who contracted the disease were treated accordingly and still followed after the event occurred. Information relative to the gender, age and location of each child was also recorded. The GPS coordinates of each household could therefore be used to calculate the distance between each house and the dam. We refer to Yewhalaw et al. (2013) for more details on the dataset.

4.4 Previous analyses of the Gilgel Gibe dataset

The first study in Ethiopia to investigate the effect of the dam on malaria incidence was conducted by Yewhalaw et al. (2013). After taking out the children who died or migrated before the end of the study, they were left with a cohort of 2040 children for the follow up. The deaths were due to various reasons and therefore not of particular interest in the study. Of this cohort, 951 children (48.09%) were females and 1059 (51.91%) males. Out of the 2040 children, 548 (26.86%) tested positive at least once for malaria infection due to *Plasmodium falciparum*. Adult mosquitoes were also collected monthly in all villages and counted. More details on the mosquito collection can be found in Yewhalaw et al. (2013). The covariates considered were: distance from the dam, season, sex, age, year, and mosquito density. Two models were used to analyze the data namely a piecewise constant baseline hazard frailty model with a gamma distributed random effect and a mixed-effects Poisson regression model. The village was treated as a random effect in both approaches. Different frailty models were considered to assess the impact of the covariates on the time to malaria infection. For instance, univariate models were fitted to evaluate the marginal effects of the covariates and a multivariate model to assess the impact of all the covariates simultaneously. Also, a mixed-effects Poisson regression model was used to explore the association between mosquito abundance and the following covariates: distance from the dam reservoir shore, year, climatic variables, and season. The study concluded that malaria transmission has a seasonal pattern (consistent over the two years) with incidence peaking just after the long rainy season and that there is no association between the distance to the dam and malaria incidence. Age was found to be a significant factor with children above three years of age having significantly higher *Plasmodium falciparum* risk than children less than 3 years of age.

In Getachew et al. (2013), the authors investigated the effect of distance of households to the dam on malaria incidence. The study was conducted on the same dataset as described above with some alterations. Malaria infection from both *Plasmodium falciparum* and *Plasmodium vivax* were considered. A survival and a count regression model were fitted and the log-likelihood expressions in both models were shown to be the same in the case the mean village distance is used for a child in the frailty model. In the count model, the data is aggregated. For example, the individual child distance to the dam is required for the survival model whereas the mean of the children distances is

taken for the count model. Two survival models are considered, a parametric proportional hazards model referred to as marginal model in the paper and a log-normal parametric frailty model referred to as conditional model in the paper. We recall that 16 villages are included in the study with different numbers of children considered per village. The marginal model is defined for $i = 1, \ldots, N$, $j = 1, \ldots, n_i$:

$$h_{ij}(t) = \sum_{m=1}^{M} \mathbb{1}_{[\tau_{m-1}, \tau_m[}(t) \exp(\lambda_{ijm}) \tag{4.1}$$

where N is the number of villages, n_i is the number of children in village i and

$$\lambda_{ijm} = \alpha_0 + \beta_y Z_{ym} + \beta_{s2} Z_{s2,m} + \beta_{s3} Z_{s3,m} + \beta_d Z_{ij,d}$$

and $t \in [\tau_{m-1}, \tau_m[$ for $m \in [1, M]$. The cut-off points $(\tau_m)_{1 \leq m \leq M}$ are fixed so as to split the time axis into 6 periods corresponding to three seasons per year and two study years. The first term α_0 is the baseline hazard and is constant within each of the M time periods. The regression parameters $\beta_y, \beta_{s2}, \beta_{s3}$ and β_d are associated to the effects of the second year, the second season, the third season and the distance respectively. The covariates are defined as follows :

$$Z_{ym} = \begin{cases} 1, & m > 3 \\ 0, & m \leq 3 \end{cases}$$

$$Z_{s2,m} = \begin{cases} 1, & m = 2, 5 \\ 0, & \text{otherwise} \end{cases}$$

$$Z_{s3,m} = \begin{cases} 1, & m = 3, 6 \\ 0, & \text{otherwise} \end{cases}$$

The covariate $Z_{ij,d}$ is quantitative and consists of the distances to the dam for all children. The conditional model is defined by allowing for a random effect at the village level in model 4.1 :

$$h_{ij}(t|b_i) = \sum_{m=1}^{M} \mathbb{1}_{[\tau_{m-1}, \tau_m[}(t) \exp(\lambda_{ijm} + b_i) \tag{4.2}$$

where b_i is assumed to follow a normal distribution with mean 0 and a variance parameter to be estimated. Both models found no significant association between the distance to the dam and malaria incidence.

The data has also been analyzed by Belay et al. (2017) where they explored the association between longitudinal measurement of mosquito abundance and time to malaria in the cohort of children through a joint Bayesian approach. The parameters associated to age and structure of roof were found to be non significant with the 95 $\%$ credible intervals containing zero in both cases. However, the parameter associated to the distance to the dam was

deemed significant as the 95 $\%$ credible interval did not contain zero. From the estimate obtained, they concluded that each additional km away from the dam results in an average reduction of malaria relative risk of 5.7%.

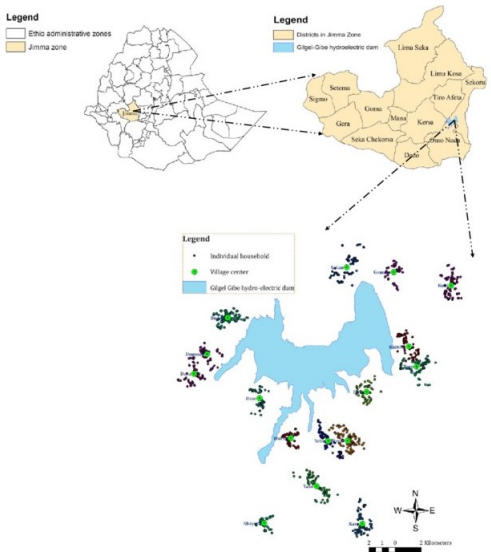

Figure 4.6: Map of Ethiopia showing districts in Jimma zone, Gilgel-Gibe hydroelectric dam and study villages. (cf. Belay et al. (2017))

Limitations of the previous studies

The village was considered as a group in Getachew et al. (2013) and Yewhalaw et al. (2013) on this dataset. However, the village clustering is hard to justify when displayed on a map (cf. Figure 4.6). The green points represent the village centers. Some households are physically closer to a village they have not been assigned to.

Frailty model estimations are not trustworthy when there exists a high correlation between the distance to the dam or the mean distance to the dam (in the case of aggregated data) and the group, i.e., the village.

Another direct consequence of the village clustering used in the different analyses is the assumption of shared frailty for every child in the same village. One child living merely a few hundred metres from another child living in a different village will not share the same frailty term. On the other hand, the same child will have the same frailty effect as a child living far away regardless of the distance between their households based only on the village assignment.

The model we propose aims at correcting for those limitations by taking into account the distance between each child and not considering the village assignation which seems to be serving only an administrative purpose without having any practical relevance.

4.5 Review of modeling and estimation methods for spatially correlated survival data

The modeling of spatial correlation in survival analysis is relatively new. Many fields have contributed to the theory and applications of spatial statistics in general such as astronomy, economics and epidemiology. The work of John Snow on the mapping of the cholera outbreak to study the spread of the disease is an example of the contribution of epidemiology to the field. Recently, geographical information systems (GIS) have been pretty useful in developing applications to track the spread of the novel COVID-19 outbreak and some applications went as far as to provide individual user information of the level of exposure in China (cf. Boulos and Geraghty (2020) for more information). GIS applications are tools that allow users to input queries, analyze spatial information, edit data in maps, and present the results of all these operations. The application of spatial statistics to survival analysis is relatively new and could improve data modelling. Spatial dependency in data is often not taken into account due to various reasons. Privacy concerns are non-negligible when dealing with geographical coordinates of patients in a clinical study for example. Aggregated spatial data at the district level of a malaria dataset in Afghanistan are modelled using generalized estimating equations (GEE) in Adegboye et al. (2018) to assess the impact of environmental factors such as rainfall, wind and vegetation on the incidence of malaria. The study also considers the anisotropic nature of the underlying spatial process, i.e orientation is also considered on top of distance. However, the correlation parameters are treated more as nuisance parameters and the aggregation of data at the district level are two limitations of the study. There are various works that deal with estimation in spatial frailty models. For instance, Banerjee et al. (2003) proposed a parametric frailty model to estimate parameters using a Bayesian approach to analyse an infant mortality dataset in Minnesota. The dataset considered is grouped into clusters and the spatial dependence between the clusters is modelled using two different approaches; a geostatistical approach where the exact locations are needed and a lattice approach where the relative distance between the groups is required. Li and Ryan (2002) developed a semi-parametric spatial frailty model with Monte Carlo simulations and Laplace approximation of a rank based marginal likelihood. Along the same lines, Lin (2012) estimates parameters of a log-normal spatial frailty model using a two-iteration approach based on an approximate likelihood function; alternating between the estimation of the regression parameter and the variance components. They establish the asymptotic properties of the estimate under some regularity conditions. A broad overview of spatial data analysis namely in disease mapping and survival analysis is detailed in Banerjee (2016). In general, the models proposed in the works mentioned above are adapted so as to answer relevant questions related to the dataset being analysed.

All of the above models introduce spatial correlation between groups of observations where the observations share different geographical coordinates. In our approach, presented in the next section, we go beyond that and model correlation at the location level.

4.6 Estimation in spatially correlated multivariate frailty models

4.6.1 Description of the spatially correlated multivariate frailty model

We propose a multivariate frailty model with spatial correlations at the location level. A location represents a child or group of children with the same geographical coordinates. We consider a sample of N locations. For $1 \leq i \leq N$, $1 \leq j \leq n_i$, the time of infection and the time of censoring for the observation j of location i are modeled by random variables denoted by T_{ij} and C_{ij} respectively. The number of days till first detection of infection is considered. Then, for $1 \leq i \leq N$, $1 \leq j \leq n_i$ the right censored time and the censoring indicator are denoted by X_{ij} and Δ_{ij} respectively and defined by $X_{ij} = \min(T_{ij}, C_{ij})$ and $\Delta_{ij} = \mathbb{1}_{T_{ij} \leq C_{ij}}$. For simplicity, we denote $\mathbf{X} = (X_{ij})_{1 \leq i \leq N, 1 \leq j \leq n_i}$ and $\mathbf{\Delta} = (\Delta_{ij})_{1 \leq i \leq N, 1 \leq j \leq n_i}$. The spatially correlated multivariate frailty model is defined for $1 \leq i \leq N$, $1 \leq j \leq n_i$ by:

$$h_{ij}(t|b_i) = h_0(t) \ \exp(Z_{ij}^t \beta + b_i) \tag{4.3}$$

where $h_{ij}(t|b_i)$ denotes the instantaneous hazard of occurrence of the event for the observation j of location i at time t, $h_0(t)$ the baseline hazard function (more details in the next section) at time t, b_i the frailty vector of the location i, β the vector of the unknown regression parameters, Z_{ij} the vector of covariates associated with the observation j of group i, namely the sex of the child, age and distance to the dam. Let us denote the frailties $\mathbf{b} = (b_i)_{1 \leq i \leq N}$.

Modeling the baseline hazard function

The baseline hazard function is parameterized by a piecewise constant function. The function h_0 is defined as $h_0(t) = h_m$ for $t \in [\tau_{m-1}, \tau_m[$ for $m \in [1, M]$ where $(\tau_m)_{m \in [1,M]}$ is a strictly increasing sequence and $\tau_0 = 0$. The model can therefore be written as:

$$h_{ij}(t|b_i) = \sum_{m=1}^{M} h_m \mathbb{1}_{[\tau_{m-1}, \tau_m[}(t) \exp(Z_{ij}^t \beta + b_i) \tag{4.4}$$

Modeling the frailty correlation structures

The frailty vector $\mathbf{b}$ is assumed to follow a multivariate normal distribution with mean 0 and covariance matrix $\Gamma = \sigma^2 \Sigma(\rho)$ where σ^2 is a scaling factor and $\Sigma(\rho)$ is the correlation matrix parameterized by $\rho > 0$. We consider two different correlation structures (cf. Li and Ryan (2002)). Let us specify the correlation matrices to be implemented :

$$\Sigma_1(\rho) = \exp(-\rho D) \tag{4.5}$$

$$\Sigma_2(\rho) = \frac{1}{1 + D^{\rho}} \tag{4.6}$$

where $D = (d_{ii'}) \in \mathbb{R}^{N \times N}$ such that $d_{ii'}$ corresponds to the distance between location i and location $i^{'}$ for $i \neq i^{'}$ and $d_{ii'} = 0$ by definition.

4.6.2 Methods for parameter estimation and model comparison

Definition of the estimator

We estimate the parameters of the model by maximising the marginal likelihood. We introduce the following assumptions on the spatial frailty model:

(F1) The censoring times (C_{ij}) are independent of the event times (T_{ij}) and of the frailties (b_i).

(F2') Conditional on the frailty vector $\mathbf{b}$, the event times (T_{ij}) are independent.

By assumptions **(F1)**, **(F2')** and the assumption that $\mathbf{b}$ follows a multivariate normal distribution, we express the complete likelihood as :

$$L_{\text{comp}}(\theta; \mathbf{X}, \mathbf{\Delta}, \mathbf{b}) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \left(\frac{\left(\sum_{m=1}^{M} h_m \mathbb{1}_{[\tau_{m-1}, \tau_m[}(X_{ij}) \exp(Z_{ij}^t \beta + b_i) \right)^{\Delta_{ij}}}{\exp(H_0(X_{ij}) \exp(Z_{ij}^t \beta + b_i))} \right) f_{\Gamma}(b) \tag{4.7}$$

where the cumulative hazard function $H_0(X_{ij}) = \sum_{m=1}^{M} h_m (\tau_m - \tau_{m-1}) \mathbb{1}_{[\tau_m, \infty[}(X_{ij}) + \sum_{m=1}^{M} (X_{ij} - \tau_{m-1}) h_m \mathbb{1}_{[\tau_{m-1}, \tau_m[}(X_{ij})$ and f_{Γ} is the density of a multivariate gaussian distribution parameterized by Γ. The model parameters to be estimated are $\theta = ((h_m)_{1 \leq m \leq M}, \beta, \sigma^2, \rho)$.

The marginal likelihood is obtained by integrating the complete likelihood with respect to the frailty vector $\mathbf{b}$:

$$L_{\text{marg}}(\theta; \mathbf{X}, \mathbf{\Delta}) = \int L_{\text{comp}}(\theta; \mathbf{X}, \mathbf{\Delta}, \mathbf{b}) d\mathbf{b}$$

We define the estimator of the maximum of the marginal likelihood $\hat{\theta}$ such that $\hat{\theta} = \text{argmax } L_{\text{marg}}(\theta; \mathbf{X}, \mathbf{\Delta})$. The estimator cannot be evaluated directly, in particular when the marginal likelihood does not admit an analytical form, which is the case in the frailty model we consider. In practice, we calculate the value of the estimator via a SAEM-MCMC algorithm (cf. Kuhn and Lavielle (2004), Allassonnière et al. (2010)) described in the next section.

Description of the SAEM-MCMC algorithm with truncation on random boundaries

The likelihood function defined in equation (4.7) belongs to the exponential family since it can be written as follows :

$$L_{\text{comp}}(\theta; \mathbf{X}, \mathbf{\Delta}, \mathbf{b}) = \exp(-\Psi(\theta) + \langle \mathcal{S}(\mathbf{b}), \Phi(\theta) \rangle) \tag{4.8}$$

where $\mathcal{S}$, Ψ and Φ are Borel functions. Sufficient statistics are explicit and can be expressed as $\mathcal{S}(\mathbf{b}) = \left(b_{ij}b_{i'j'}, i, i^{'} = 1,\ldots,N, j, j^{'} = 1,\ldots,n_i, \exp(b_{ij}), i = 1,\ldots,N, j = 1,\ldots,n_i\right)^t$. Let $(\mathcal{K}_q)_q \geq 0$ be a sequence of increasing compact subsets of S such that $\bigcup_{q\geq 0}\mathcal{K}_q = S$ and $\mathcal{K}_q \subset \text{int}(\mathcal{K}_{q+1})$, for all $q \geq 0$.

Initialize θ_0 in Θ, b_0 and s_0 in two fixed compact sets **K** and $\mathcal{K}_0$ respectively.

Each iteration of the algorithm, denoted by Algorithm $\mathcal{B}$ later, is composed of four steps detailed below.

Repeat until convergence for $k \geq 1$:

1. **Simulation step:** Draw $\bar{\mathbf{b}}$ of the unobserved frailty vector from a transition probability $\Pi_{\theta_{k-1}}$ of a convergent Markov chain having as stationary distribution the conditional distribution $\pi_{\theta_{k-1}}(.|\mathbf{X}, \mathbf{\Delta})$ where

$$\pi_\theta(\mathbf{b}|\mathbf{X}, \mathbf{\Delta}) = \frac{L_{\text{comp}}(\theta; \mathbf{X}, \mathbf{\Delta}, \mathbf{b})}{L_{\text{marg}}(\theta; \mathbf{X}, \mathbf{\Delta})}$$

with the current parameters $\bar{\mathbf{b}} \sim \Pi_{\theta_{k-1}}(\mathbf{b}_{k-1}, .)$

2. **Stochastic approximation step:** Compute $\bar{s} = s_{k-1} + \mu_k(S(\bar{\mathbf{b}}) - s_{k-1})$ where the sequence $(\mu_k)_k$ satisfies the following conditions:
$0 \leq \mu_k \leq 1, \sum \mu_k = +\infty, \sum \mu_k^2 < +\infty$.

3. **Truncation step:** If $\bar{s}$ is outside the current compact set $\mathcal{K}_{\kappa_{k-1}}$, where κ is the index of the current active truncation set, or too far from the previous value s_{k-1} then restart the stochastic approximation in the initial compact set, extend the truncation boundary to $\mathcal{K}_{\kappa_k}$ and start again with a bounded value of the missing variable. Otherwise, set $(\mathbf{b}_k, s_k) = (\bar{\mathbf{b}}, \bar{s})$ and keep the truncation boundary to $\mathcal{K}_{\kappa_{k-1}}$.

4. **Maximization step:**

$$\theta_k = \underset{s}{\text{argmax}}\ \hat{\theta}(s)$$

where the function $\hat{\theta}$: S $\rightarrow \Theta$ is defined such that:

$$\forall\ s \in \text{S}, \forall\ \theta \in \Theta,\ L(\hat{\theta}(s), s) \geq L(\theta, s)$$

where $L(\theta, s) = \exp(-\Psi(\theta) + \langle s, \Phi(\theta)\rangle)$ (cf. assumption (M5) in Appendix C)

Note that the sequence $(\mathbf{b}_k, s_k)$ generated by this algorithm satisfies two conditions at each iteration k. Namely we check whether the stochastic approximation wanders outside the current compact set and whether the current value is not too far from the previous value. The latter can be expressed as follows:

$$\|s_k - s_{k-1}\| \leq \epsilon_k$$

where $\epsilon = (\epsilon_k)_{k\geq 0}$ is a monotone non-increasing sequence of positive numbers. A more detailed description of the truncation step can be found in Andrieu et al. (2005).

Convergence property of algorithm $\mathcal{B}$

In this section, we prove the almost sure convergence of the sequence $(\theta_k)_k$ generated by algorithm $\mathcal{B}$ to a critical point of the marginal likelihood. Let us detail the assumptions we make on the model, the dynamic of the algorithm and the Markov kernel in the simulation step.

We make classical model assumptions **(M3)**-**(M5)** (detailed in Appendix C) to prove the convergence of EM like algorithms following those of Delyon et al. (1999). Following Andrieu et al. (2005), we state a first assumption **(A1')** that guarantees the existence of a global Lyapunov function denoted by w defined as:

$$w(s) = -\log \int L(\hat{\theta}(s); \mathbf{X}, \mathbf{\Delta}, \mathbf{b}) d\mathbf{b} \tag{4.9}$$

for the mean field h defined as:

$$h(s) = \int (\mathcal{S}(\mathbf{b}) - s) \pi_\theta(\mathbf{b}|\mathbf{X}, \mathbf{\Delta}) d\mathbf{b} \tag{4.10}$$

(A1') The functions w and h are such that

(i) there exists an $M_0 > 0$ such that

$$\mathcal{S} = \{s \in S, \langle \nabla w(s), h(s) \rangle = 0\} \subset \{s \in S, w(s) < M_0\}$$

where w is defined in (4.9) and h is defined in (4.10).

(ii) there exists $M_1 \in]M_0, \infty]$ such that $\{s \in \mathsf{S}, w(s) < M_1\}$ is a compact set.

(iii) the closure of $w(\mathcal{L})$ has an empty interior.

We now state a condition on the step-size sequence in the stochastic approximation step of the algorithm.

(A4) The sequences $\mu = (\mu_k)_{k\geq 0}$ and $\epsilon = (\epsilon_k)_{k\geq 0}$ are non-increasing, positive and satisfy $\sum_{k=0}^{\infty} \mu_k = \infty$, $\lim_{k\to\infty} \epsilon_k = 0$ and $\sum_{k=1}^{\infty} \{\mu_k^2 + \mu_k \epsilon_k^a + (\mu_k \epsilon_k^{-1})^p\} < \infty$, where $a \in]0, 1]$ and $p \geq 2$.

Finally we consider the usual drift assumption **(DRI)** (cf. Andrieu et al. (2005)). This set of assumptions is classical in Markov chain literature.

Theorem 10 *Assume that **(F1),(F2'), (M3)-(M5), (A1'), (A4)** and **(DRI)** are fulfilled. Then we have with probability 1*

$$\lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0$$

where $(\theta_k)_k$ is generated by algorithm $\mathcal{B}$, $d(x, A)$ denotes the distance from x to any closed subset A and $\mathcal{L} = \{\theta \in \Theta, \partial_\theta \log L_{marg}(\theta; \mathbf{X}, \mathbf{\Delta}) = 0\}$.

Proof of Theorem 10: The proof of Theorem 10 follows the lines of the proof of Theorem 5 established in Chapter 2. We first apply Theorem 5.5 of Andrieu et al. (2005) to prove the convergence of the sequence $(s_k)_k$ and check therefore the required assumptions.

As detailed in Andrieu et al. (2005), drift assumptions **(DRI)** imply assumptions **(A2-A3)** by Proposition 6.1. Thus we can apply Theorem 5.5 of Andrieu et al. (2005).

This results in the sequence $(s_k)_k$ generated by algorithm $\mathcal{B}$ satisfying $\lim_k d(s_k, \mathcal{S}) = 0$. We recall that we choose a piecewise baseline function h_0 and we assume that $\mathbf{b}$ follows a multivariate normal distribution. In addition to those model hypotheses and assumptions **(F1)**, **(F2')**, we have that assumptions **(M1)** and **(M2)** of Delyon et al. (1999) hold in our case. It suffices to show that Ψ and Φ are twice continuously differentiable to satisfy assumption **(M2)**. This is guaranteed by assumptions **(F1)**, **(F2')**, the choices of the frailty distribution and of the baseline hazard function h_0. As shown in equation (4.8), we can write the complete likelihood in exponential form which fulfils assumption **(M1)**. Following the lines of the proof of Lemma 2 of Delyon et al. (1999), we get that $\lim_k d(\theta_k, \mathcal{L}) = 0$. The proof of Theorem 10 is therefore complete.

Model comparison

There are different possibilities for the modeling of the quantities in the spatially correlated frailty model defined by equation (4.3). Indeed, in section 4.6.1, we defined a piecewise constant function for baseline hazard h_0. We can consider different cut-off points and different values of M to model h_0. There are also two correlation structures to choose from as shown in section 4.6.1. Models defined following different correlation structures are nonnested.

In addition to comparing different correlation structures and modeling of h_0, we can also determine the significance of regression parameter estimates by means of likelihood-ratio tests.

Comparing nonnested models

In the case of nonnested models, the model that best fits the data is chosen based on penalized likelihood based criteria such as Akaike information criterion (AIC) (cf. Akaike (1974)) and Bayesian information criterion (BIC) (cf. Schwarz et al. (1978)).

Let k be the number of estimated parameters in the model and N the number of data points. Let $\hat{L}$ be the maximum value of the likelihood function for the model. Then the AIC of the model is calculated as follows :

$$\text{AIC} = -2\log(\hat{L}) + 2k$$

The BIC of the model is calculated as follows :

$$\text{BIC} = -2\log(\hat{L}) + k\log(N)$$

The two criteria are similar but with different penalties for the number of parameters. They do not answer the same question though. The AIC tries to select the model that most adequately describes an unknown reality while the BIC tries to find the true model among the set of model candidates. The unrealistic assumption that the true model is among the candidate set is quite problematic for the BIC. A comparison of the two approaches is studied by Burnham et al. (2011). They present a few simulation studies that suggest AIC tends to have performance advantages over BIC. If we compare models with an equal number of parameters and built on the same number of data points, then AIC and BIC are equivalent.

Likelihood-ratio test

The likelihood-ratio test assesses the goodness of fit of two nested statistical models based on the ratio of their likelihoods. Suppose we have a statistical model with parameter space Θ. The null and alternative hypotheses are defined as follows: $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta \setminus \Theta_0$ where Θ_0 is a subset of Θ. The test statistic is then given by :

$$L_{\text{RT}} = -2 \log \left(\frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)} \right)$$

The quantity inside the parentheses is called the likelihood ratio. Since all likelihoods are positive and the value of the numerator cannot be bigger than that of the denominator in the likelihood ratio, the ratio is bounded between zero and one. Under the null hypothesis, the test statistic L_{RT} will be asymptotically chi-squared distributed with degrees of freedom equal to the difference in dimensionality of Θ and Θ_0 (cf. Wilks (1938)).

4.6.3 Implementation of the estimation algorithm

First, we detail the computation of the Maximization step (M step) of algorithm $\mathcal{B}$.

Equations for the Maximization step of Algorithm $\mathcal{B}$

Let us first define the complete log-likelihood of the data :

$$\begin{aligned} \log L_{\text{comp}}(\theta; \mathbf{X}, \mathbf{\Delta}, \mathbf{b}) \quad = \quad & \sum_{i=1}^{N} \sum_{j=1}^{n_i} \Bigg(\Delta_{ij} \Big(\log \big(\sum_{m=1}^{M} h_m \mathbb{1}_{[\tau_{m-1}, \tau_m[}(X_{ij}) \big) + \\ & Z_{ij}^t \beta + b_i \Big) - H_0(X_{ij}) \exp(Z_{ij}^t \beta + b_i) \Bigg) + \\ & \log f_\Gamma(b) \end{aligned}$$

where the cumulative hazard function $H_0(X_{ij}) = \sum_{m=1}^{M} h_m (\tau_m - \tau_{m-1}) \mathbb{1}_{[\tau_m, \infty[}(X_{ij}) + \sum_{m=1}^{M} (X_{ij} - \tau_{m-1}) h_m \mathbb{1}_{[\tau_{m-1}, \tau_m[}(X_{ij})$.

We differentiate the complete log-likelihood with respect to each parameter to obtain the necessary equations to update the parameter estimates in algorithm $\mathcal{B}$.

Updating parameter $(h_m)_{1\leq m\leq M}$

Differentiating the complete log-likelihood with respect to h_m, we obtain :

$$\frac{\partial \log L_{\text{comp}}(\theta;\mathbf{X},\mathbf{\Delta},\mathbf{b})}{\partial h_m} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i}\Delta_{ij}\mathbb{1}_{[\tau_{m-1},\tau_m[}(X_{ij})}{h_m} \tag{4.11}$$

$$-\sum_{i=1}^{N}\sum_{j=1}^{n_i}\exp(Z_{ij}^t\beta+b_{ij})\left((\tau_m-\tau_{m-1})\mathbb{1}_{[\tau_m,\infty[}(X_{ij})+(X_{ij}-\tau_{m-1})\mathbb{1}_{[\tau_{m-1},\tau_m[}(X_{ij})\right) \tag{4.12}$$

We obtain an analytic expression for the estimation of h_m :

$$\hat{h}_m = \frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i}\Delta_{ij}\mathbb{1}_{[\tau_{m-1},\tau_m[}(X_{ij})}{\sum_{i=1}^{N}\sum_{j=1}^{n_i}\exp(Z_{ij}^t\beta+b_{ij})\left((\tau_m-\tau_{m-1})\mathbb{1}_{[\tau_m,\infty[}(X_{ij})+(X_{ij}-\tau_{m-1})\mathbb{1}_{[\tau_{m-1},\tau_m[}(X_{ij})\right)} \tag{4.13}$$

Updating parameter β

Differentiating the complete log-likelihood with respect to β, we obtain :

$$\frac{\partial \log L_{\text{comp}}(\theta;\mathbf{X},\mathbf{\Delta},\mathbf{b})}{\partial \beta} = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\left(\Delta_{ij}Z_{ij}^t - Z_{ij}^t H_0(X_{ij})\exp\left(Z_{ij}^t\beta+b_i\right)\right) \tag{4.14}$$

$$\frac{\partial^2 \log L_{\text{comp}}(\theta;\mathbf{X},\mathbf{\Delta},\mathbf{b})}{\partial \beta^2} = -\sum_{i=1}^{N}\sum_{j=1}^{n_i} Z_{ij}Z_{ij}^t H_0(X_{ij})\exp\left(Z_{ij}^t\beta+b_i\right) \tag{4.15}$$

The Newton Raphson method is used to update the values of parameter β.

Updating parameter σ^2

Only the last term of the log-likelihood depends on the parameter σ^2. We start by expressing the log density in a simpler form:

$$\begin{aligned}\log f_\Sigma(b) &= -N\log(\sqrt{2\pi}) - \frac{1}{2}\log(\det(\sigma^2\Sigma(\rho))) - \frac{1}{2\sigma^2}b^t\Sigma(\rho)^{-1}b \\ &= -N\log(\sqrt{2\pi}) - \frac{1}{2}\log\Big((\sigma^2)^N\det(\Sigma(\rho))\Big) - \frac{1}{2\sigma^2}b^t\Sigma(\rho)^{-1}b\end{aligned} \tag{4.16}$$

We then differentiate the log-likelihood with respect to σ^2 to obtain:

$$\frac{\partial \log L_{\text{comp}}(\theta;\mathbf{X},\mathbf{\Delta},\mathbf{b})}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}b^t\Sigma(\rho)^{-1}b \tag{4.17}$$

We obtain an analytic expression to update the parameter σ^2:

$$\hat{\sigma}^2 = \frac{1}{N}b^t\Sigma(\hat{\rho})^{-1}b \tag{4.18}$$

Updating parameter ρ

We now differentiate the log-likelihood with respect to ρ.

$$\begin{aligned}\frac{\partial \log L_{\text{comp}}(\theta; \mathbf{X}, \mathbf{\Delta}, b)}{\partial \rho} &= -\frac{1}{2}\frac{(\sigma^2)^N \text{det}(\Sigma(\rho))\text{tr}(\Sigma(\rho)^{-1}\frac{\partial \Sigma(\rho)}{\partial \rho})}{(\sigma^2)^N \text{det}(\Sigma(\rho))} + \frac{1}{2} b^t (\sigma^2)^{-1}\Sigma(\rho)^{-1}\frac{\partial}{\partial \rho}\Sigma(\rho)\Sigma(\rho)^{-1} b \\ &= -\frac{1}{2}\text{tr}\Big(\Sigma(\rho)^{-1}\frac{\partial}{\partial \rho}\Sigma(\rho)\Big) + \frac{1}{2}(\sigma^2)^{-1} b^t \Sigma(\rho)^{-1}\frac{\partial}{\partial \rho}\Sigma(\rho)\Sigma(\rho)^{-1} b \end{aligned} \tag{4.19}$$

The values of the parameter ρ are updated with the help of a gradient descent method.

Computation of the marginal likelihood

We recall that the frailty $\mathbf{b}$ is assumed to follow a multivariate normal distribution parameterized by Γ. A numerical approximation of the marginal likelihood is computed based on the parameter estimates $\hat{\theta} = ((\hat{h}_m)_{1\leq 1\leq M}, \hat{\beta}, \hat{\sigma}^2, \hat{\rho})$. We simulate C independent realizations of $(\mathbf{b}_c)_{1\leq c\leq C}$ following a multivariate normal distribution with mean zero and covariance matrix $\hat{\Gamma} = \hat{\sigma}^2\Sigma(\hat{\rho})$. The marginal likelihood is then computed based on a Monte Carlo sum as follows :

$$\hat{L}_{\text{marg}}(\theta; \mathbf{X}, \mathbf{\Delta}) = \frac{1}{C}\sum_{c=1}^{C} L_{\text{cond}}(\hat{\theta}; \mathbf{X}, \mathbf{\Delta}|\mathbf{b}_c)$$

where

$$L_{\text{cond}}(\theta; \mathbf{X}, \mathbf{\Delta}|\mathbf{b}) = \prod_{i=1}^{N}\prod_{j=1}^{n_i}\left(\frac{\Big(\sum_{m=1}^{M} h_m \mathbb{1}_{[\tau_{m-1},\tau_m[}(X_{ij})\ \exp(Z_{ij}^t\beta + b_i)\Big)^{\Delta_{ij}}}{\exp(H_0(X_{ij})\ \exp(Z_{ij}^t\beta + b_i))}\right)$$

The law of large numbers (cf. Van der Vaart (2000)) ensures that the Monte Carlo sum converges to the marginal likelihood as $C \to \infty$. The bigger the number of realizations C, the better the quality of the approximation. We compute $\hat{L}_{\text{marg}}$ for increasing values of C. The values obtained are compared and once they are of the same order, the quality of the approximation is determined to be sufficient.

Computation of the Fisher information matrix

We obtain an estimate of the Fisher information matrix through the observed Fisher information matrix $I_{obs}(\theta) = -\partial^2_\theta \text{log} L_{\text{marg}}(\theta; \mathbf{X}, \mathbf{\Delta})$ (cf. Andersen et al. (1997)). Using Louis's missing information principle (cf. Louis (1982)), we express the matrix $I_{obs}(\theta)$ as :

$$I_{obs}(\theta) = -\mathbb{E}_\theta\big(\partial^2_\theta \log L_{\text{comp}}(\theta; \mathbf{X}, \mathbf{\Delta}, \mathbf{b})|\mathbf{X}, \mathbf{\Delta}\big) - \text{Cov}_\theta\big(\partial_\theta \log L_{\text{comp}}(\theta; \mathbf{X}, \mathbf{\Delta}, \mathbf{b})|\mathbf{X}, \mathbf{\Delta}\big)$$

where $\mathbb{E}_\theta$ and Cov_θ denote respectively the expectation and the covariance under the posterior distribution π_θ of the frailty.

We approximate the quantity $I_{obs}(\theta)$ by a Monte Carlo sum based on the realizations of the Markov chain generated in the algorithm having as stationary distribution the posterior distribution π_θ. After a burn-in period, we use

the remaining C realizations $(\mathbf{b}_c)_{1\leq c\leq C}$ of the Markov chain to compute the following quantity :

$$\begin{aligned}\hat{I}_C(\hat{\theta}) = &-\frac{1}{C}\sum_{c=1}^{C}\partial_\theta^2 \log L_{\text{comp}}(\hat{\theta};\mathbf{X},\mathbf{\Delta},\mathbf{b}_c)\\ &-\frac{1}{C}\sum_{c=1}^{C}\left(\partial_\theta \log L_{\text{comp}}(\hat{\theta};\mathbf{X},\mathbf{\Delta},\mathbf{b}_c)\partial_\theta \log L_{\text{comp}}(\hat{\theta};\mathbf{X},\mathbf{\Delta},\mathbf{b}_c)^t\right)\\ &+\frac{1}{C^2}\left(\sum_{c=1}^{C}\partial_\theta \log L_{\text{comp}}(\hat{\theta};\mathbf{X},\mathbf{\Delta},\mathbf{b}_c)\right)\left(\sum_{c=1}^{C}\partial_\theta \log L_{\text{comp}}(\hat{\theta};\mathbf{X},\mathbf{\Delta},\mathbf{b}_c)\right)^t\end{aligned}$$

The ergodic theorem guarantees the convergence of the quantity $\hat{I}_C(\hat{\theta})$ to the observed Fisher information matrix $I_{obs}(\hat{\theta})$ as C goes to infinity (cf. Meyn and Tweedie (1993)). In addition to the derivatives calculated to compute the M-step of algorithm $\mathcal{B}$ in section 4.6.3, we also compute the following second and cross derivatives :

$$\frac{\partial^2 \log L_{\text{comp}}(\theta;\mathbf{X},\mathbf{\Delta},\mathbf{b})}{\partial h_m^2} = -\frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i}\Delta_{ij}\mathbb{1}_{[\tau_{m-1},\tau_m[}(X_{ij})}{h_m^2}$$

$$\frac{\partial^2 \log L_{\text{comp}}(\theta;\mathbf{X},\mathbf{\Delta},\mathbf{b})}{\partial(\sigma^2)^2} = \frac{N}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3}\mathbf{b}^t\Sigma(\rho)^{-1}\mathbf{b}$$

$$\begin{aligned}\frac{\partial^2 \log L_{\text{comp}}(\theta;\mathbf{X},\mathbf{\Delta},\mathbf{b})}{\partial\rho^2} = &-\frac{1}{2}\text{tr}\Big(-A(\rho)+\Sigma(\rho)^{-1}\frac{\partial^2}{\partial\rho^2}\Sigma(\rho)\Big)\\ &+\frac{1}{2}(\sigma^2)^{-1}\mathbf{b}^t\Big(\{-A(\rho)\\ &+\Sigma(\rho)^{-1}\frac{\partial^2}{\partial\rho^2}\Sigma(\rho)\}\Sigma(\rho)^{-1}-A(\rho)\,\Sigma(\rho)^{-1}\Big)\mathbf{b}\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \log L_{\text{comp}}(\theta;\mathbf{X},\mathbf{\Delta},\mathbf{b})}{\partial\beta\partial h_m} = &-\sum_{i=1}^{N}\sum_{j=1}^{n_i}Z_{ij}^t\,\exp\,(Z_{ij}^t\beta+b_i)\\ &\times\left((\tau_m-\tau_{m-1})\mathbb{1}_{[\tau_m,\infty[}(X_{ij})+(X_{ij}-\tau_{m-1})\mathbb{1}_{[\tau_{m-1},\tau_m[}(X_{ij})\right)\end{aligned}$$

$$\frac{\partial^2 \log L_{\text{comp}}(\theta;\mathbf{X},\mathbf{\Delta},\mathbf{b})}{\partial\rho\partial\sigma^2} = -\frac{1}{2(\sigma^2)^2}\mathbf{b}^t\Sigma(\rho)^{-1}\frac{\partial}{\partial\rho}\Sigma(\rho)\Sigma(\rho)^{-1}\mathbf{b}$$

We give some practical details on the implementation of the algorithm.

1. When dealing with real data such as the malaria data which we will describe later, the high dimension of the frailty $\mathbf{b}$ can make the simulation step computationally intensive. We deal with the high dimensional nature of $\mathbf{b}$ using an adaptive version of the SAEM-MCMC algorithm (cf. Haario et al. (2001)). At iteration k of the algorithm, the transition kernel used for simulating the unobserved frailty is chosen as a transition kernel of a random walk Metropolis Hastings algorithm with proposal distribution q equal to a normal distribution centred at the current value $\mathbf{b}_{k-1}$. Usually, we update $\mathbf{b}_{k-1}$ coordinate-wise with each candidate accepted with a

probability that has to be computed N times for a frailty vector of size N. This step is computationally intensive with an increasing size N of the frailty vector. Instead, we update blocks of size K at a time. In doing so, the computational cost is reduced by a factor of K. The size of the block K is tuned to ensure the stationarity of the Markov Chain generated and the convergence of the parameters estimated. A subtle compromise has to be found between a too big value of K which translates to a change in too many directions at a time for $\mathbf{b}$ and a too small K that has a high computational cost.

Updating the frailty vector block-wise serves only a numerical purpose. With more efficient programming, notably using a low-level programming language such as C++ instead of R, the algorithm should be reasonably fast in handling a coordinate-wise update of the frailty vector.

2. In order to reduce the computational time to achieve convergence, it is crucial to start the algorithm with good initial estimates. The usual approach is to start with estimates obtained from simpler existing methods. For instance, we use as initial values for the regression parameter β the estimated values obtained when fitting the data by a lognormal shared frailty model. For the baseline components parameters, we use estimates obtained from a piecewise constant proportional hazards model as starting point.

3. The decreasing positive step size $(\mu_k)_k$ is taken as follows for all $0 \leq k \leq K_0,\ \mu_k = 1$ and for all $k > K_0,\ \mu_k = \frac{1}{(k-K_0)}$ where K_0 is a number to be specified. The step size $(\mu_k)_k$ verifies assumption **(A4)**. The algorithm is said to have no memory during the first K_0 iterations. After this burn-in time which allows for the algorithm to widely visit the parameter space, the sequence $(\mu_k)_k$ decreases and converges to zero as $k \to \infty$.

4. Following Ripatti et al. (2002), we consider a stopping criterion based on the relative difference between two consecutive values of the parameters. Let us fix a positive threshold $\epsilon > 0$. If for some $k > 1$:

$$\frac{\|\theta_k - \theta_{k-1}\|}{\|\theta_{k-1}\|} < \epsilon$$

holds true for three consecutive iterations, the algorithm is stopped.

4.6.4 Simulation study

Study of the consistency property of the estimate $\hat{\theta}$

We begin by studying the consistency of the estimate $\hat{\theta}$ numerically. The simulation setting is set up so as to be close to the malaria data. The data are generated as follows :

$$h_{ij}(t|b_i) = (h_1 \mathbb{1}_{[0,\tau_1[}(t) + h_2 \mathbb{1}_{[\tau_1,\tau_2[}(t) + h_3 \mathbb{1}_{[\tau_2,\infty[}(t))\ \exp(Z_{ij}^t \beta + b_i) \tag{4.20}$$

We compare the estimates in two different settings: $N = 100$ and $N = 300$ with location sizes varying from 1 to 4. The frailty vector is simulated following a multivariate normal distribution with covariance matrix $\sigma^2 \Sigma(\rho)$ where

the correlation structure is defined by $\Sigma(\rho) = \exp(-\rho D)$ with scaling parameter $\sigma^2 = 2$ and $\rho = 1$. We take subsets of the real malaria distance matrix to define the matrix D; the 100 first distances and 300 first distances. The baseline components are fixed at $(h_1, h_2, h_3) = (2, 0.5, 1)$ with change points $(\tau_1, \tau_2) = (0.2, 2)$, regression parameter $\beta = (2, 3)$ and the covariates Z_{ij} are simulated following a Bernoulli distribution. There is no censoring. The results are presented in Table 4.1.

Table 4.1: Mean of parameter estimates and empirical standard error in parentheses obtained from 100 repetitions with the event times simulated following model (4.20). Comparing estimates when $N = 100$ and $N = 300$ with location sizes varying from 1 to 4.

Parameters	True Values	$N = 100$	$N = 300$
(h_1, h_2, h_3)	(2, 0.5, 1)	(2.281, 0.591, 1.305) (1.217, 0.374, 0.889)	(1.964, 0.488, 0.968) (0.757, 0.193, 0.421)
(β_1, β_2)	(2, 3)	(2.017, 3.088) (0.355, 0.422)	(1.983, 2.967) (0.169, 0.208)
σ^2	2	2.267 (0.909)	2.002 (0.519)
ρ	1	1.258 (0.415)	1.022 (0.212)

We make several observations from Table 4.1. We observe an improvement in all estimates when the sample size increases from 100 locations to 300 locations. The estimates obtained when $N = 300$ are closer to the true values. We note significant improvements in the estimation of parameters (h_1, h_2, h_3) which go from $(2.281,\ 0.591,\ 1.305)$ for $N = 100$ to $(1.964,\ 0.488,\ 0.968)$ for $N = 300$. With a small sample size, there are fewer observations in each interval which results in less information for the estimation of the baseline components. The standard deviations are smaller when $N = 300$.

Robustness to misspecification of the correlation structure

The aim of this section is to assess the robustness of the model and parameter estimates when the correlation structure is misspecified. We keep the same simulation setting as in the previous section with $N = 300$ and different location sizes varying from 1 to 4. We simulate the event times under three different censoring settings (no censoring, moderate censoring and heavy censoring). The censoring times are simulated following an exponential distribution with the rate parameter adjusted so as to obtain the desired censoring level. We estimate the parameters in two cases; when the correlation structure is correctly specified ($\Sigma(\rho) = \exp(-\rho D)$) and when the correlation structure is misspecified ($\Sigma(\rho) = \frac{1}{1+D^\rho}$). We denote by $\hat{\theta}_{\text{c}}$ the estimates obtained when we estimate in the correct model, i.e assuming the correlation structure used in the simulation procedure and we denote by $\hat{\theta}_{\text{w}}$ the estimates obtained when we purposely misspecify the correlation structure in the estimation procedure. The estimates obtained from 100 repetitions are displayed in Table 4.2.

Table 4.2: Mean of parameter estimates and empirical standard error in parentheses obtained from 100 repetitions with the event times simulated following model (4.20). Comparison of estimates $\hat{\theta}_c$ when the correlation structure is correctly specified and estimates $\hat{\theta}_w$ when the correlation structure is misspecified. The number of locations is set to $N = 300$ with different location sizes varying from 1 to 4.

Censoring	Parameters	True Values	$\hat{\theta}_c$	$\hat{\theta}_w$
0 %	(h_1, h_2, h_3)	(2, 0.5, 1)	(1.964, 0.488, 0.968) (0.757, 0.193, 0.421)	(2.019, 0.522, 1.059) (1.60, 0.440, 0.839)
	(β_1, β_2)	(2, 3)	(1.983, 2.967) (0.169, 0.208)	(2.005, 3.003) (0.183, 0.213)
	σ^2	2	2.002 (0.519)	2.500 (0.612)
	ρ	1	1.022 (0.212)	0.846 (0.151)
40 %	(h_1, h_2, h_3)	(2, 0.5, 1)	(1.989, 0.506, 1.104) (0.991, 0.275, 0.667)	(2.308, 0.594, 1.219) (1.588, 0.431, 0.819)
	(β_1, β_2)	(2, 3)	(2.041, 3.043) (0.197, 0.242)	(2.045, 3.067) (0.218, 0.255)
	σ^2	2	2.059 (0.592)	2.465 (0.702)
	ρ	1	1.079 (0.112)	0.885 (0.119)
60 %	(h_1, h_2, h_3)	(2, 0.5, 1)	(2.050, 0.525, 1.215) (0.992, 0.304, 0.824)	(2.213, 0.582, 1.244) (1.53, 0.545, 1.060)
	(β_1, β_2)	(2, 3)	(2.013, 2.999) (0.240, 0.255)	(2.018, 2.978) (0.236, 0.244)
	σ^2	2	2.158 (0.667)	2.460 (0.795)
	ρ	1	1.290 (0.398)	0.888 (0.112)

From Table 4.2, we observe that the estimates of the regression parameters β and baseline parameters $(h_m)_{1\leq m\leq 3}$ are quite robust to the misspecification of the covariance structure. However, we observe a slight degradation of the estimate $\hat{h}_3$ under heavy censoring under both correct and misspecification of the correlation structure. The events in the last interval are disproportionately censored in a number of datasets generated which explains the effect on the quality of this particular estimate. The parameter estimates ρ and σ^2 are close to the true values when the correlation structure is correctly specified. When the correlation structure is misspecified, the scaling factor σ^2 seems to be compensating for the wrong assumption of correlation structure which leads to an overestimation of the estimate.

Comparing with models that do not take into account spatial correlation

In this section, we compare the estimations obtained by the spatially correlated frailty model with two models that do not take into account the spatial nature present in data. We keep the same simulation setting as in section 4.6.4. We compare our estimator with two existing methods namely a shared log-normal frailty model and a piecewise constant

proportional hazards method. Using the *coxme* package (cf. Therneau (2018b)), we implement a log-normal shared frailty model :

$$\forall t \geq 0 \quad h_{ij}(t|b_i) = h_0(t) \exp(Z_{ij}^t \beta + b_i) \tag{4.21}$$

We denote the estimator by $\hat{\theta}_{\text{sf}}$. We recall that *coxme* estimates parameters based on the maximization of the integrated partial likelihood with a Laplace approximation to deal with the intractable integral. The partial likelihood no longer involves the baseline h_0 which is therefore not estimated in this case. We also estimate the parameters using a piecewise proportional hazards model through the *eha* package (cf.Broström and Broström (2019)). The model is defined as follows :

$$h_{ij}(t) = (h_1 \mathbb{1}_{[0,\tau_1[}(t) + h_2 \mathbb{1}_{[\tau_1,\tau_2[}(t) + h_3 \mathbb{1}_{[\tau_2,\infty[}(t)) \exp(Z_{ij}^t \beta)$$

We denote this estimator by $\hat{\theta}_{\text{ph}}$ and correctly specify the cut-points τ_1 and τ_2 as those used in simulating the data. We insist on the fact that this is not a comparison of our method with the two packages mentioned. We aim to assess the impact of not taking into account spatial correlation in data and merely implement the packages to compare with estimation methods that do not account for spatial correlation.

From Table 4.3, we observe that estimates of the regression parameters and baseline components are close to the true values only when in the column corresponding to estimates $\hat{\theta}_{\text{c}}$. Otherwise, the regression parameter estimates are severely biased. The estimates corresponding to the baseline components in the case of the piecewise proportional hazards model are also biased even if we specified the correct cut-points. For example, when all events are non-censored, the baseline component estimates for the piecewise proportional hazards model are $(2.806,\ 0.302,\ 0.188)$ and are therefore far from the true values $(2,\ 0.5,\ 1)$. This modeling approach is further disadvantaged by the proportional hazards assumption and consequently not taking into account heterogeneity in the data whether spatially correlated or not. In the log-normal shared frailty, heterogeneity in the data is accounted for but the frailty terms associated to each location are assumed to be independent and the spatial correlation is not included in the model. The regression parameter estimates obtained are closer to the true values when compared to those obtained by the piecewise proportional hazards model but still far from the true values and those obtained when taking into account the spatial correlation. We compare for example, when all events are non-censored, β estimates $(1.692,\ 2.552)$ which are far from the true values $(2,\ 3)$ whereas the estimates when the spatial correlation is accounted for are $(1.983,\ 2.967)$ and are therefore much closer to the true values. The variance obtained with the log-normal shared frailty model is difficult to interpret in this case as the frailty distribution is misspecified. We make the same observations in the moderate and heavy censoring settings with bigger standard deviation values in general as more events are censored. The results show the importance of taking into account spatial correlations when present in data.

Table 4.3: Mean of parameter estimates and empirical standard error in parentheses obtained from 100 repetitions with the event times simulated following model (4.20). Comparison of estimates $\hat{\theta}_{\text{c}}$, $\hat{\theta}_{\text{ph}}$ and $\hat{\theta}_{\text{sf}}$. The number of locations is set to $N = 300$ with location sizes varying from 1 to 4.

Censoring	Parameters	True Values	$\hat{\theta}_{\text{c}}$	$\hat{\theta}_{\text{ph}}$	$\hat{\theta}_{\text{sf}}$
0 %	(h_1, h_2, h_3)	(2, 0.5, 1)	(1.964, 0.488, 0.968) (0.757, 0.193, 0.421)	(2.806, 0.302, 0.188) (0.605, 0.071, 0.072)	$\times$ $\times$
	(β_1, β_2)	(2, 3)	(1.983, 2.967) (0.169, 0.208)	(1.412, 2.003) (0.208, 0.216)	(1.692, 2.552) (0.212, 0.264)
	σ^2	2	2.002 (0.519)	$\times$ $\times$	1.074 (0.381)
	ρ	1	1.022 (0.212)	$\times$ $\times$	$\times$ $\times$
40 %	(h_1, h_2, h_3)	(2, 0.5, 1)	(1.989, 0.506, 1.104) (0.991, 0.275, 0.667)	(2.821, 0.300, 0.386) (0.606, 0.075, 0.160)	$\times$ $\times$
	(β_1, β_2)	(2, 3)	(2.041, 3.043) (0.197, 0.242)	(1.368, 2.041) (0.188, 0.200)	(1.694, 2.548) (0.222, 0.263)
	σ^2	2	2.059 (0.592)	$\times$ $\times$	0.997 (0.396)
	ρ	1	1.079 (0.112)	$\times$ $\times$	$\times$ $\times$
60 %	(h_1, h_2, h_3)	(2, 0.5, 1)	(2.050, 0.525, 1.215) (0.992, 0.304, 0.824)	(2.929, 0.354, 0.643) (0.704, 0.139, 0.343)	$\times$ $\times$
	(β_1, β_2)	(2, 3)	(2.013, 2.999) (0.240, 0.255)	(1.432, 2.140) (0.219, 0.219)	(1.666, 2.480) (0.250, 0.259)
	σ^2	2	2.158 (0.667)	$\times$ $\times$	0.973 (0.409)
	ρ	1	1.290 (0.398)	$\times$ $\times$	$\times$ $\times$

4.7 Gilgel Gibe malaria data analysis

4.7.1 Modeling of the malaria data

In this section, we analyze the malaria data collected in the surroundings of the Gilgel Gibe hydroelectric dam using our spatially correlated multivariate frailty model. We consider a population of 2037 children whose times to first malaria infection are studied. The geographical coordinates (longitude and latitude) of the children located in 16 villages around the dam are used to compute the inter-distances between all children. The village structure is not taken into account as the village boundaries are mostly of an administrative nature. For each child, on top of the geographical coordinates, we consider four covariates namely the sex of the child, age, the structure of the roof of the child's household and the distance to the dam. We consider the spatially correlated multivariate frailty described in the construction of model (4.3). As opposed to previous studies on the dataset and spatial frailty models in general, we consider a frailty term at the child level. Some children have the exact same geographical coordinates or live so close to each other that the locations recorded are exactly the same. Those children will share a common frailty term so that the distance matrix remains invertible and positive definite. We note however that this "grouping" structure is different from the usual grouping structure in spatial survival models in the sense that the members of the group have the same geographical coordinates. In the common grouping structure, the group normally refers to a region, state or country (cf. Li and Ryan (2002), Banerjee (2016)). By abuse of notation, we refer to the children with same geographical coordinates as groups of children. The number of children in a group varies from 1 to 11 with most groups consisting of 1, 2 or 3 children (760 groups of 1, 364 groups of 2 and 89 groups of 3).

4.7.2 Description of the spatially correlated frailty models

We recall the spatially correlated multivariate frailty model which is defined for $1 \leq i \leq N$, $1 \leq j \leq n_i$:

$$h_{ij}(t|b_i) = \sum_{m=1}^{M} h_m \mathbb{1}_{[\tau_{m-1},\tau_m[}(t) \ \exp(Z_{ij}^t \beta + b_i)$$

The times to first malaria infection are possibly right censored and denoted by $\mathbf{X} = (X_{ij})_{1\leq i\leq N, 1\leq j\leq n_i}$. We denote the censoring indicator by $\mathbf{\Delta} = (\Delta_{ij})_{1\leq i\leq N, 1\leq j\leq n_i}$. The covariates $\mathbf{Z}$ are sex, age, distance to the dam and nature of the roof with associated regression parameters $\beta = (\beta_{\text{sex}}, \beta_{\text{age}}, \beta_{\text{d}}, \beta_{\text{roof}})$. The frailties $(b_i)_{1\leq i\leq N}$ are assumed to follow a multivariate normal distribution with mean 0 and covariance matrix parameterized by σ^2 and ρ. The model parameters to be estimated are therefore $\theta = ((h_m)_{1\leq m\leq 6}, \beta, \sigma^2, \rho)$. The estimates are obtained using an adaptative SAEM-MCMC algorithm. The estimation procedure and algorithm are detailed in Section 4.6.2. The algorithm detailed in Section 4.6.2 is implemented with a Gibbs-block of size $K = 10$. Smaller block sizes were tested and gave similar results while the algorithm failed to converge for larger block sizes. Initial parameter estimates for β are obtained from a shared frailty model without spatial correlation using the R package *coxme* (cf.

Therneau (2018b)). The starting point for the baseline components are the estimates given by R package *eha* (cf. Broström and Broström (2019)) which allows for estimation in a piecewise baseline proportional hazards model. We consider four models following two specifications of the baseline hazard functions and two correlations structures.

Modeling the baseline following the two years and three seasons

The baseline hazard function is parameterized by a piecewise constant function to take into account the different years and seasons. The function h_0 is defined as $h_0(t) = h_m$ for $t \in [\tau_{m-1}, \tau_m[$ for $m \in [1, 6]$ where $(\tau_m)_{m\in[1,6]}$ is a strictly increasing sequence and $\tau_0 = 0$. The six baseline components correspond to the long rainy season, dry season and moderate rainy season of the first and second year as shown in Figure 4.7.

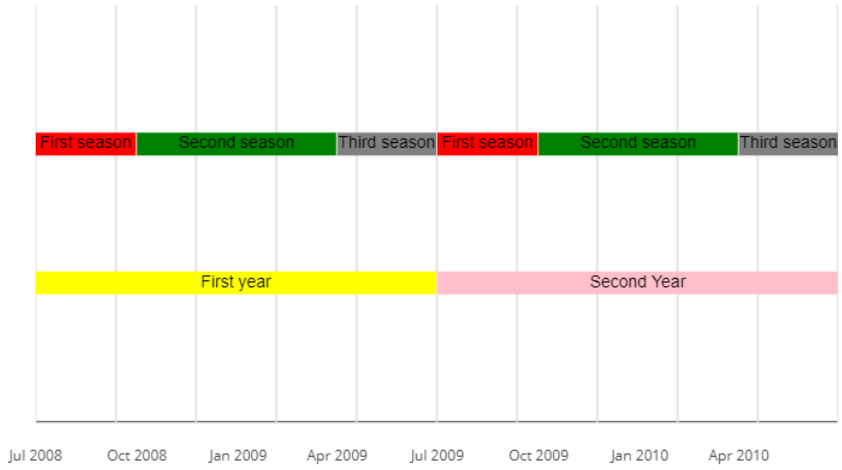

Figure 4.7: The three seasons and two years

We define two models namely $\boldsymbol{\mathcal{S}}_1$ and $\boldsymbol{\mathcal{S}}_2$ as follows :

$$h_{ij}(t|b_i) = \sum_{m=1}^{6} h_m \mathbb{1}_{[\tau_{m-1},\tau_m[}(t) \exp(Z_{ij}^t\beta + b_i) \text{ where } (b_i)_{1\leq i\leq N} \sim \mathcal{N}(0, \sigma^2 \exp(-\rho D)) \qquad (\boldsymbol{\mathcal{S}}_1)$$

$$h_{ij}(t|b_i) = \sum_{m=1}^{6} h_m \mathbb{1}_{[\tau_{m-1},\tau_m[}(t) \exp(Z_{ij}^t\beta + b_i) \text{ where } (b_i)_{1\leq i\leq N} \sim \mathcal{N}\Big(0, \sigma^2 \frac{1}{1+D^\rho}\Big) \qquad (\boldsymbol{\mathcal{S}}_2)$$

Modeling the baseline following the average daily rainfall

The seasonal effect of malaria incidence is largely due to rain that favors the breeding of mosquitoes. Therefore, it might be of interest to model the baseline hazard function h_0 following rainfall patterns instead of seasonal patterns as shown in Figure 4.8. The baseline hazard function is parameterized by a piecewise constant function to take into account the average daily rainfall during the two-year study. We define $h_0(t) = h_l$ for $t \in [\nu_{l-1}, \nu_l[$ for $l \in [1, 6]$ where $(\nu_l)_{l\in[1,6]}$ is a strictly increasing sequence and $\nu_0 = 0$. The cut-points $(\nu_l)_{l\in[1,6]}$ are chosen following the ones considered in Belay et al. (2017).

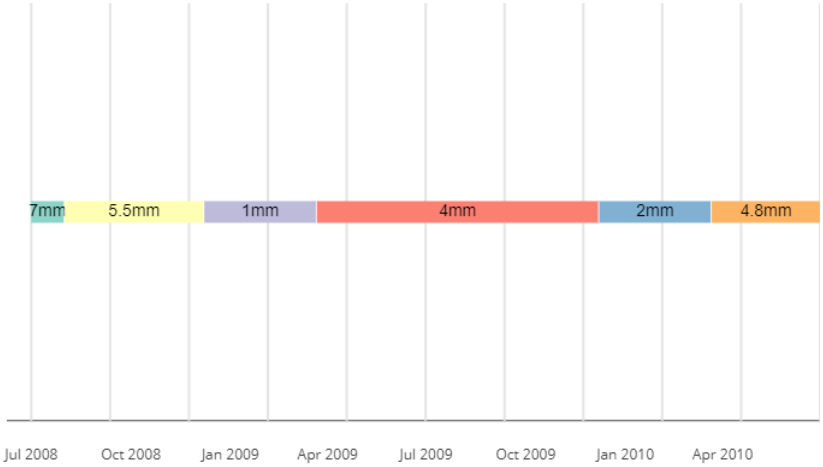

Figure 4.8: Time intervals based on average daily rainfall patterns

We define two models namely $\mathcal{S}_3$ and $\mathcal{S}_4$ as follows :

$$h_{ij}(t|b_i) = \sum_{l=1}^{6} h_l \mathbb{1}_{[\nu_{l-1},\nu_l[}(t) \exp(Z_{ij}^t\beta + b_i) \text{ where } (b_i)_{1\leq i\leq N} \sim \mathcal{N}(0, \sigma^2 \exp(-\rho D)) \qquad (\mathcal{S}_3)$$

$$h_{ij}(t|b_i) = \sum_{l=1}^{6} h_l \mathbb{1}_{[\nu_{l-1},\nu_l[}(t) \exp(Z_{ij}^t\beta + b_i) \text{ where } (b_i)_{1\leq i\leq N} \sim \mathcal{N}\Big(0, \sigma^2 \frac{1}{1+D^\rho}\Big) \qquad (\mathcal{S}_4)$$

4.7.3 Model comparison and parameter estimation

Table 4.4: Model comparison based on marginal log-likelihood values : malaria data analysis

	Seasonal baseline	Average rainfall baseline
$\sigma^2 \exp(-\rho D)$	-6094.2 ($\mathcal{S}_1$)	-6047.6 ($\mathcal{S}_3$)
$\sigma^2 \frac{1}{1+D^\rho}$	-6096.2 ($\mathcal{S}_2$)	-6021.4 ($\mathcal{S}_4$)

We estimate the parameters and compute the marginal log-likelihoods in models $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$ and $\mathcal{S}_4$. In the four models, the number of parameters is the same. We therefore directly compare the marginal log-likelihoods. We present the results in Table 4.4. Based on the marginal log-likelihoods, model $\mathcal{S}_4$ is the best fit for the data. We also present the estimates of model $\mathcal{S}_1$ to evaluate the impact of seasonal patterns on malaria incidence.

Model $\mathcal{S}_4$ estimates

The estimates obtained in model $\mathcal{S}_4$ are presented in Table 4.5. Likelihood-ratio tests are performed to assess the significance of the regression parameters. We consider the usual α level of 5%. The results obtained are displayed in Table 4.6. For each statistical test, we compare models that differ by one parameter. Therefore, under

Table 4.5: Mean and model-based standard errors in parentheses of parameters estimated in model $\mathcal{S}_4$

β_{sex}	β_{age}	β_{d}	β_{roof}	$(h_1, h_2, h_3, h_4, h_5, h_6) \times 10^{-4}$	σ^2	ρ
-0.0391	-0.0061	0.1057	0.0260	(5.40, 14.3, 3.98, 6.88, 2.42, 2.71)	0.364	0.794
(0.0659)	(0.0201)	(0.140)	(0.0342)	(0.48,0.97,0.22,0.49,0.11,0.18)	(0.088)	(0.11)

Table 4.6: Likelihood-ratio tests to test the significance of regression parameters β

Null hypothesis H_0	Likelihood-ratio test statistic	p-value
$\beta_{\text{sex}} = 0$	3.390	0.0656
$\beta_{\text{age}} = 0$	1.990	0.158
$\beta_{\text{d}} = 0$	0.800	0.371
$\beta_{\text{roof}} = 0$	2.156	0.142

the assumption of the null hypothesis, the test statistic will be asymptotically χ^2 distributed with one degree of freedom. Based on the p-values which are all greater than α, we fail to reject the null hypothesis in the four cases. This suggests that the regression parameters may not be significant. The other studies on this dataset conclude similarly except for Belay et al. (2017) where they find β_{d} to be significant.

We give a graphical representation of the estimations of the baseline components in Figure 4.9. The malaria risk seems to be highest during or just after periods of heavy rainfall. The limited number of observations in the last interval makes the interpretation of the last baseline component tricky. We expect the estimate $\hat{h}_6$ to be higher had data been available for a few more weeks.

In Figure 4.10, we give a graphical representation of how correlation between the children evolves with respect to the distance between them following estimates obtained ($\hat{\rho} = 0.794$).

Model $\mathcal{S}_1$ estimates

The estimates obtained in model $\mathcal{S}_1$ are presented in Table 4.7.

The piecewise baseline hazard estimated is presented in Figure 4.11. The estimated hazards for the different seasons are color coded accordingly. It seems from Figure 4.11 that malaria risk is highest just after or during the long rainy season. However, the estimates also point to a high malaria risk during the second dry season while the risk drops at the start of the moderate rainy season which is counter-intuitive. One possible explanation is that the seasons have slightly shifted along with the rainy periods.

In Figure 4.12, we give a graphical representation of how correlation between the children evolves with respect to the distance between them following estimates obtained ($\hat{\rho} = 1.50$). This behaviour is coherent with the maximum

Table 4.7: Mean and model-based standard errors in parentheses of parameters estimated in model $\mathcal{S}_1$

β_{sex}	β_{age}	β_{d}	β_{roof}	$(h_1, h_2, h_3, h_4, h_5, h_6) \times 10^{-4}$	σ^2	ρ
-0.0149	0.0079	0.0595	0.0076	(8.85,8.22,4.90,2.67,7.26,1.74)	0.449	1.50
(0.0751)	(0.0212)	(0.0622)	(0.0150)	(0.52,0.64,0.32,0.15,0.66,0.94)	(0.091)	(0.19)

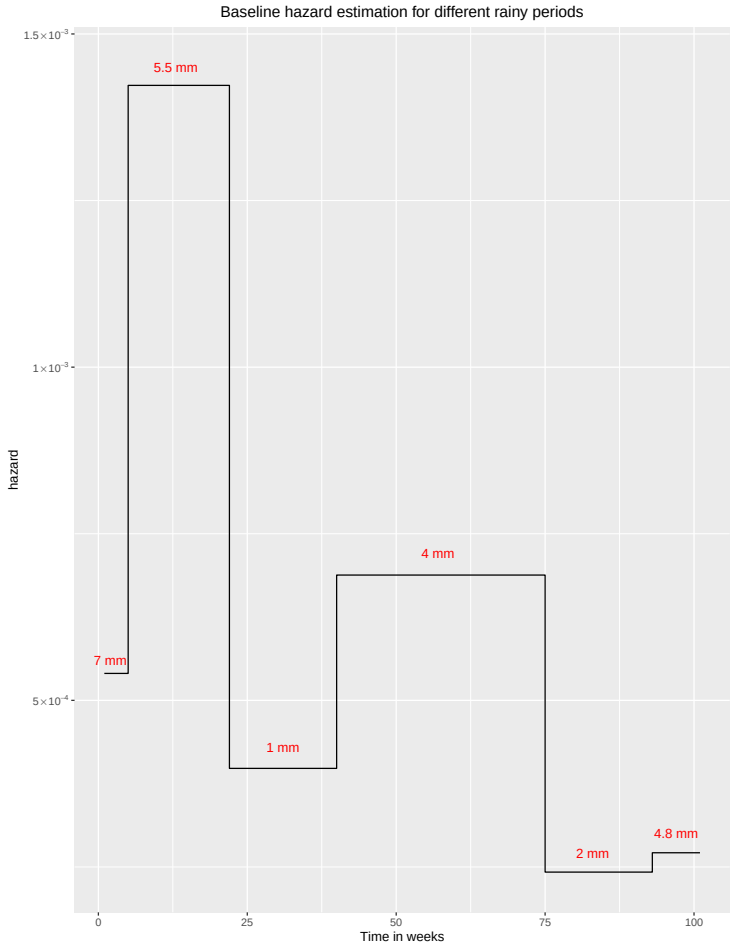

Figure 4.9: Hazard rates for different rain patterns. Avergae daily rainfall within different time periods annotated in red

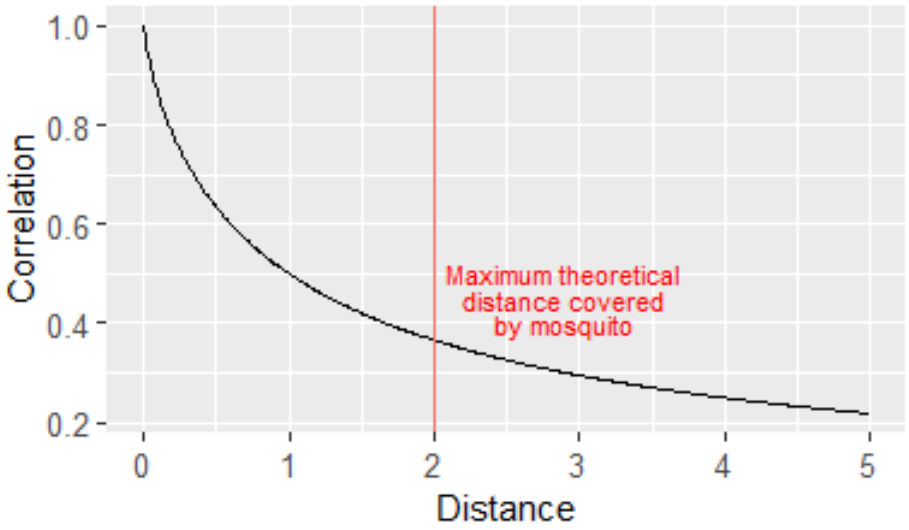

Figure 4.10: Graphical representation of correlation as a function of distance based on estimate $\hat{\rho} = 0.794$ in model $\mathcal{S}_4$

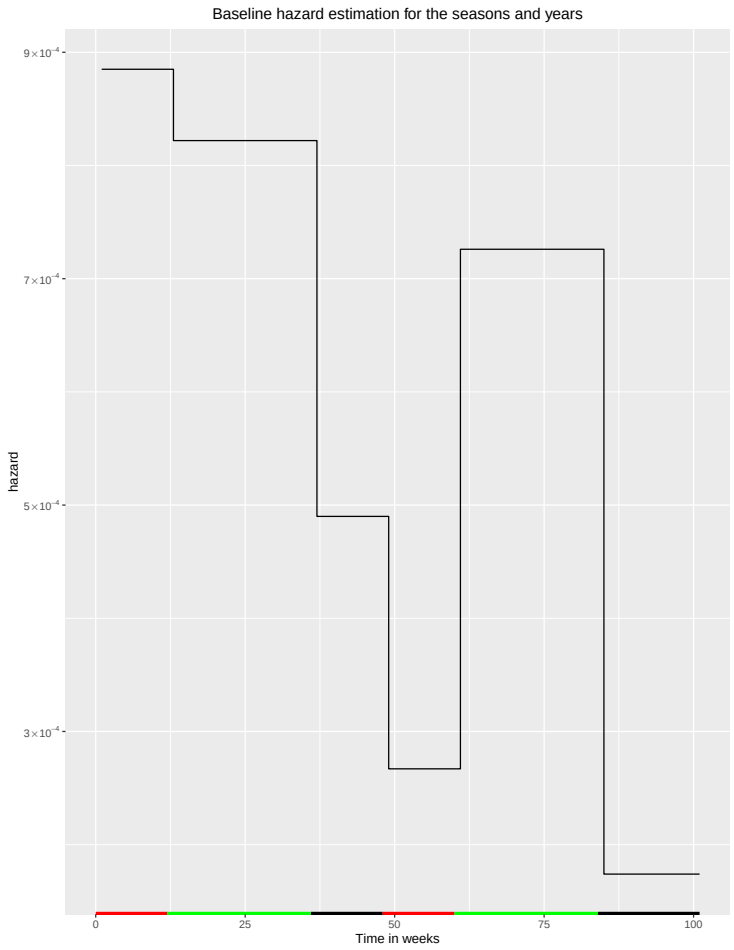

Figure 4.11: Hazard rate for each segment color coded according to the seasons: long rainy season (red), dry season (green) and moderate rainy season (black)

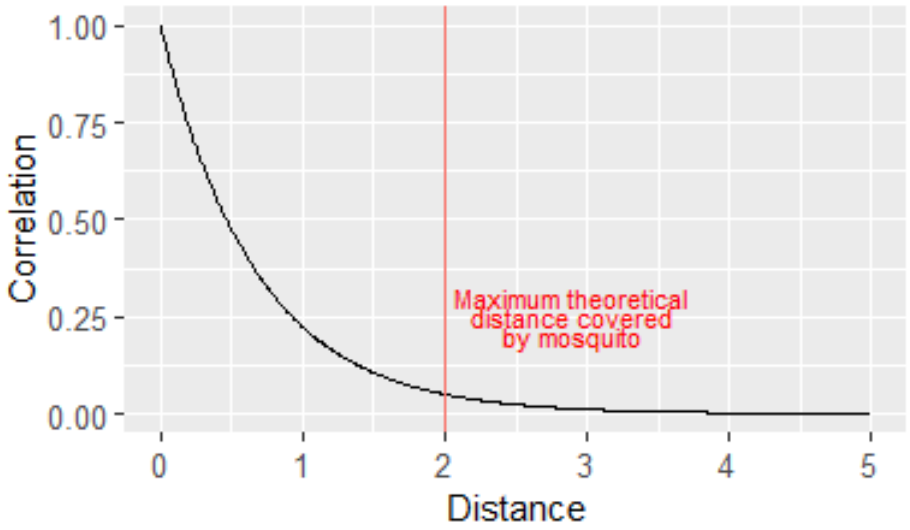

Figure 4.12: Graphical representation of correlation as a function of distance based on estimate $\hat{\rho} = 1.50$ in model $\mathcal{S}_1$

theoretical distance of 2 km covered by the mosquito. We can safely assume that children who live more than 2 km apart pose less risk to each other. Model $\mathcal{S}_4$ suggests stronger correlation between the children when compared to estimates from model $\mathcal{S}_1$.

4.8 Conclusion and perspectives

In this chapter, we first outlined the malaria problematic, its impact, transmission, diagnosis and treatment. The spread of malaria is still a major problem in the WHO Africa region. Although, Ethiopia has seen a drop in cases in the last few years, the disease remains active in the country. The various factors influencing the transmission from person to person contribute to the complexity of modeling and analyzing malaria data. We have at our disposal a detailed malaria dataset consisting of the times to malaria infection of children living in villages near a hydroelectric dam. After reviewing the different models applied to the dataset, we proposed a new spatial frailty model. Our model is inspired by the malaria data, more specifically by the spatial correlation present in the data. The spatial nature of the data stems from the increased risk for children living in close proximity. Contrary to existing spatial frailty models that take into account spatial correlation between regions, we considered spatial correlation at the child level. We implemented a stochastic EM algorithm to estimate parameters of the model. The almost sure convergence of the sequence of parameter estimates generated by the algorithm towards a critical point of the marginal likelihood is established. With the aim of analyzing the malaria dataset, we specified different correlation structures and baseline hazard functions to define four spatial frailty models. The model that best fits the data is then chosen using a likelihood based criterion. Following model comparison, the estimates are interpreted to get a better insight on the incidence of the disease in the region.

Let us now offer some perspectives on the analysis of the malaria data. As a result of the high computational cost of running the algorithm, we defined a limited number of spatial frailty models. It could be of interest to consider different combinations of cut-off points for the baseline function and compare a bigger set of models. Other more complex correlation structures could offer further information on the spatial correlation between the children. Also, the baseline hazard function is chosen as a piecewise constant function in our case but other distributions can be considered. Or, we could do without any parametric modeling of the baseline hazard function and estimate the regression and frailty parameters based on the integrated partial likelihood.

Spatial survival analysis is a relatively new research topic that will be more relevant with the increasing availability of geographical data. The recent COVID-19 outbreak and various tools based on GPS tracking come to mind. The spatial frailty model we propose is not limited to the malaria problematic. Even though the model was inspired by malaria data, it is defined through a very general framework and could easily be applied to other datasets.

Chapter 5

General conclusion of the thesis and perspectives

In this thesis, we proposed contributions to different aspects of frailty models.

First, we developed a new estimating method based on the integrated partial likelihood to estimate parameters in frailty models. The parameters of the model are estimated by maximizing the integrated partial likelihood. These estimates are computed through a stochastic Expectation Maximization algorithm. The almost sure convergence of the sequence generated by the algorithm to a critical point of the integrated partial likelihood is established. The estimation method allows for a wide choice of frailty distributions including also complex correlation structures. Also, the integrated partial likelihood does not require any modeling of the baseline hazard function. In particular, we showed via a simulation study that making a wrong choice of parametric distribution for the baseline has an adverse effect on parameter estimates. Furthermore, we tested the robustness of the estimation procedure under misspecification of the frailty distribution. We also applied the proposed estimation method to analyze real mastitis data and a bladder cancer dataset. We compared the estimates obtained to those of other studies on those datasets. Going forward, the new estimating method should be implemented in an R package.

Our second contribution concerned the study of convergence rates of maximum likelihood estimates (MLEs) in parametric shared frailty models. We proposed a conjecture based on the parameterization of the conditional likelihood as well as on the structure of covariates. First, we established theoretically the different convergence rates of MLEs in a simple linear mixed-effects model depending on the structure of covariates. However such kind of calculations could not be carried out in frailty models, whose likelihoods have complex analytical expressions involving in particular integrates. Therefore, we conducted an intensive simulation study on a Weibull shared frailty model. The simulation setting was set up so as to test different scenarios namely with respect to the parameterization of the conditional likelihood and the structure of covariates. Within the framework of parametric shared frailty models,

the following conjecture can be formulated : (1) MLEs of the frailty distribution parameters are $\sqrt{N}$-consistent. (2) MLEs of baseline parameters and regression parameters involved in the conditional likelihood with an additive frailty term are $\sqrt{N}$-consistent. (3) MLEs of regression parameters that are not subject to an additive frailty term in the conditional likelihood with associated covariates at the group level are $\sqrt{N}$-consistent. (4) All other parameters are $\sqrt{NJ}$-consistent.

There are numerous perspectives to our contribution. In particular, after the promising simulation study on the Weibull shared frailty model, it would be interesting to extend the simulation study to other frailty models. Then, the next logical step would consist in establishing a theoretical proof of the proposed conjecture. It would also be interesting to study the asymptotic normality of the MLE in parametric frailty models.

We proposed a third contribution in the form of a spatially correlated frailty model. The model we developed is inspired from malaria data collected in Ethiopia. The time to malaria infection of 2037 children living in villages located near a hydroelectric dam are analyzed. The spatial nature of the data stems from the fact that the children may live in close proximity. Consequently, we considered correlation at the child level rather than at the village level as done previously and estimated parameters in a spatially correlated frailty model. The parameters are estimated using a stochastic Expectation Maximization algorithm and the theoretical properties of the algorithm are established. We considered four models following different correlation structures and baseline hazard functions taking into account different seasonal effects. The models are compared by means of a likelihood-based criterion to select the one that best fits the data. We considered covariates namely sex, age, structure of the roof of households and distance to the dam and quantified their effects on the incidence of malaria. However, we concluded that the parameters may not be significant based on likelihood-ratio tests. Besides, the interpretation of parameter estimates associated to the correlation structure seemed to be in accordance with biological reality. In particular, the estimates that characterize the correlation structure point to weak between-child correlation beyond distances of $2km$. The distance of $2km$ is incidentally the maximum theoretical distance travelled by the mosquito vector.

The rich dataset put together coupled with the flexible spatially correlated frailty model we developed offers a lot of possibilities for future research work. In our study, we considered four possible models with two baseline hazard functions and two correlation structures. We could consider more models with the aim to better fit the data and come up with new insights on malaria incidence in the region. In particular, it might also be of interest to account for mosquito density in the model. Beyond the malaria problematic, the spatial frailty model we developed has the potential for wider applications. With the growing popularity of geographical data in the field of epidemiology in general, models that account for spatial correlations may offer deeper insights in the modeling and interpretation of the spatial nature of diseases. The development of mobile applications to limit the risk of spreading of the recent COVID-19 outbreak comes to mind.

Finally, it would be interesting to extend all the contributions of this thesis to the context of frailty models involving

covariates of high dimension. Indeed, nowadays more information on data is collected, in particular genetic information which usually involves high dimensional covariates. Including this information in the modeling task is critical to carry out relevant data analyses. However this requires the development of new adapted estimation tools, such as penalized estimation and variable selection, leading to new challenges from both numerical and theoretical points of view.

Bibliography

O. A. Adegboye, D. H. Leung, and Y.-G. Wang. Analysis of spatial data with a nested correlation structure. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(2):329–354, 2018.

H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

S. Allassonnière, E. Kuhn, and A. Trouvé. Construction of bayesian deformable models via a stochastic approximation algorithm: A convergence study. *Bernoulli*, 16:641–678, 2010.

P. Andersen and R. Gill. Cox's regression model for counting processes : a large sample study. *Annals of Statistics*, 10:1100–1120, 1982.

P. Andersen, J. Klein, K. Knudsen, and R. Tabanera y Palacios. Estimation of variance in cox's regression model with shared gamma frailties. *Biometrics*, 53:1475–1484, 1997.

D. A. Anderson and M. Aitkin. Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(2):203–210, 1985.

C. Andrieu, E. Moulines, and P. Priouret. Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim*, 44:283–312, 2005.

T. A. Balan and H. Putter. frailtyem: An r package for estimating semiparametric shared frailty models. *Journal of Statistical Software*, 90(1):1–29, 2019.

S. Banerjee. Spatial data analysis. *Annual review of public health*, 37:47–60, 2016.

S. Banerjee, M. M. Wall, and B. P. Carlin. Frailty modeling for spatially correlated survival data, with application to infant mortality in minnesota. *Biostatistics*, 4(1):123–142, 2003.

D. B. Belay, Y. G. Kifle, A. T. Goshu, J. M. Gran, D. Yewhalaw, L. Duchateau, and A. Frigessi. Joint bayesian modeling of time to malaria and mosquito abundance in ethiopia. *BMC infectious diseases*, 17(1):415, 2017.

M. N. K. Boulos and E. M. Geraghty. Geographical tracking and mapping of coronavirus disease covid-19/severe acute respiratory syndrome coronavirus 2 (sars-cov-2) epidemic and associated events around the world: how 21st century gis technologies are supporting the global fight against outbreaks and epidemics, 2020.

R. A. Bradley and J. J. Gart. The asymptotic properties of ml estimators when sampling from associated populations. *Biometrika*, 49(1/2):205–214, 1962.

N. Breslow and J. Crowley. A large sample study of the life table and product limit estimates under random censorship. *The Annals of statistics*, pages 437–453, 1974.

G. Broström and M. G. Broström. Package 'eha', 2019.

K. P. Burnham, D. R. Anderson, and K. P. Huyvaert. Aic model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral ecology and sociobiology*, 65(1):23–35, 2011.

G. Canfora, M. Ceccarelli, L. Cerulo, and M. Di Penta. How long does a bug survive? an empirical study. In *2011 18th Working Conference on Reverse Engineering*, pages 191–200. IEEE, 2011.

D. Clayton and J. Cuzick. Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society: Series A (General)*, 148(2):82–108, 1985.

D. G. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151, 1978.

E. Colosimo, F. v. Ferreira, M. Oliveira, and C. Sousa. Empirical comparisons between kaplan-meier and nelson-aalen survival function estimators. *Journal of Statistical Computation and Simulation*, 72(4):299–308, 2002.

D. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society*, 34:187—-220, 1972.

D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

M. Davidian and D. M. Giltinan. Some general estimation methods for nonlinear mixed-effects model. *Journal of Biopharmaceutical Statistics*, 3(1):23–55, 1993.

M. Davidian and D. M. Giltinan. *Nonlinear models for repeated measurement data*, volume 62. CRC press, 1995.

B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.*, 27(1):94–128, 1999. ISSN 0090-5364.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

I. Do Ha, M. Noh, and Y. Lee. frailtyhl: A package for fitting frailty models with h-likelihood. *The R Journal*, 4(2): 28–36, 2012.

L. Duchateau and P. Janssen. Penalized partial likelihood for frailties and smoothing splines in time to first insemination models for dairy cows. *Biometrics*, 60:608–614, 2004.

L. Duchateau and P. Janssen. *The Frailty Model*. Springer-Verlag, New York, 2008.

C. Elbers and G. Ridder. True and spurious duration dependence: The identifiability of the proportional hazard model. *The Review of Economic Studies*, 49(3):403–409, 1982.

A. Gamst, M. Donohue, and R. Xu. Asymptotic properties and empirical evaluation of the npmle in the proportional hazards mixed-effects model. *Statistica Sinica*, pages 997–1011, 2009.

C. Geerdens, G. Claeskens, and P. Janssen. Goodness-of-fit tests for the frailty distribution in proportional hazards models with shared frailty. *Biostatistics (Oxford, England)*, 14, 2012.

Y. Getachew, P. Janssen, D. Yewhalaw, N. Speybroeck, and L. Duchateau. Coping with time and space in modelling malaria incidence: a comparison of survival and count regression models. *Statistics in medicine*, 32(18):3224–3233, 2013.

T. Girum, T. Shumbej, and M. Shewangizaw. Burden of malaria in ethiopia, 2000-2016: findings from the global health estimates 2016. *Tropical Diseases, Travel Medicine and Vaccines*, 5(1):11, 2019.

M. Greenwood. The first life table. *Notes and Records of the Royal Society of London*, 1(2):70–72, 1938.

P. Greenwood and W. Wefelmeyer. Efficiency of estimators for partially specified filtered models. *Stochastic processes and their applications*, 36(2):353–370, 1990.

I. D. Ha, J.-H. Jeong, and Y. Lee. *Statistical Modelling of Survival Data with Random Effects*. Springer, Singapore, 2017.

H. Haario, E. Saksman, J. Tamminen, et al. An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

H. V. Henderson and S. R. Searle. On deriving the inverse of a sum of matrices. *Siam Review*, 23(1):53–60, 1981.

P. Hougaard. Shared frailty models. In *Analysis of multivariate survival data*, pages 215–262. Springer, 2000.

P. Hougaard. *Analysis of multivariate survival data*. Springer Science & Business Media, 2012.

K. J. Ishak, N. Kreif, A. Benedict, and N. Muszbek. Overview of parametric survival analysis for health-economic applications. *Pharmacoeconomics*, 31(8):663–675, 2013.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

J. Klein. Semiparametric of random effects using the cox model based on the em algorithm. *Biometrics*, 48:795–806, 1992.

J. P. Klein and M. L. Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2006.

X. Kong, K. J. Archer, L. H. Moulton, R. H. Gray, and M.-C. Wang. Parametric frailty models for clustered data with arbitrary censoring: application to effect of male circumcision on hpv clearance. *BMC medical research methodology*, 10(1):40, 2010.

E. Kuhn and C. El-Nouty. On a convergent stochastic estimation algorithm for frailty models. *Statistics and Computing*, 23:413–423, 2013.

E. Kuhn and M. Lavielle. Coupling a stochastic approximation version of em with an mcmc procedure. *ESAIM: Probability and Statistics*, 8:115–131, 2004.

E. Kuhn, K. Goethals, C. El-Nouty, and L. Duchateau. Assessing the correlation structure in cow udder quarter infection times through extensions of the correlated frailty model. *JABES*, 21:601—-618, 2016.

H. Laevens, H. Deluyker, Y. Schukken, L. De Meulemeester, R. Vandermeersch, E. De Muelenaere, and A. De Kruif. Influence of parity and stage of lactation on the somatic cell count in bacteriologically negative dairy cows. *Journal of dairy science*, 80(12):3219–3226, 1997.

Y. Li and L. Ryan. Modeling spatial survival data using semiparametric frailty models. *Biometrics*, 58(2):287–297, 2002.

P.-S. Lin. Analysis of spatial frailty models by a weighted estimating equation. *Journal of Statistical Planning and Inference*, 142(6):1436–1444, 2012.

T. Louis. Finding the observed information matrix when using the em algorithm. *J. Roy. Statist. Soc. Ser. B*, 44: 226–233, 1982.

W. M. Mendenhall and T. L. Sincich. *Statistics for Engineering and the Sciences*. CRC Press, 2016.

S. P. Meyn and R. Tweedie. Markov chains and stochastic stability. communications and control engineering series. *Springer-Verlag London Ltd*, 1993.

K. O. Mfuh, O. A. Achonduh-Atijegbe, O. N. Bekindaka, L. F. Esemu, C. D. Mbakop, K. Gandhi, R. G. Leke, D. W. Taylor, and V. R. Nerurkar. A comparison of thick-film microscopy, rapid diagnostic test, and polymerase chain reaction for accurate diagnosis of plasmodium falciparum malaria. *Malaria journal*, 18(1):73, 2019.

G. S. Mudholkar, D. K. Srivastava, and G. D. Kollia. A generalization of the weibull distribution with application to the analysis of survival data. *Journal of the American Statistical Association*, 91(436):1575–1583, 1996.

M. Munda, F. Rotolo, C. Legrand, et al. Parfm: parametric frailty models in r. *Journal of Statistical Software*, 51(11): 1–20, 2012.

S. A. Murphy. Consistency in a proportional hazards model incorporating a random effect. *Annals of Statistics*, 22: 712–731, 1994.

S. A. Murphy. Asymptotic theory for the frailty model. *Annals of Statistics*, 23:182–198, 1995.

J. M. Neuhaus and J. D. Kalbfleisch. Between-and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, pages 638–645, 1998.

L. Nie. Strong consistency of the maximum likelihood estimator in generalized linear and nonlinear mixed-effects models. *Metrika*, 63(2):123–143, 2006.

L. Nie. Convergence rate of mle in generalized linear and nonlinear mixed-effects models: Theory and applications. *Journal of Statistical Planning and Inference*, 137(6):1787–1804, 2007.

G. G. Nielsen, R. D. Gill, P. K. Andersen, and T. I. A. Sorensen. A counting process approach to maximum likelihood estimation in frailty models. *Scand. J. Statist.*, 19:25–44, 1992.

A. Oodally, L. Duchateau, and E. Kuhn. Convergent stochastic algorithm for parameter estimation in frailty models using integrated partial likelihood. *arXiv preprint arXiv:1909.07056*, 2019.

E. Parner. Asymptotic theory for the correlated gamma-frailty model. *Annals of Statistics*, 26:183–214, 1998.

S. Ripatti and J. Palmgren. Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56:1016—-1022, 2000.

S. Ripatti, K. Larsen, and J. Palmgren. Maximum likelihood inference for multivariate frailty models using an automated monte carlo em algorithm. *Lifetime Data Analysis*, 8(4):349–360, 2002.

C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

V. Rondeau, D. Commenges, and P. Joly. Maximum penalized likelihood estimation in a gamma-frailty model. *Lifetime data analysis*, 9(2):139–153, 2003.

V. Rondeau, Y. Mazroui, and J. Gonzalez. frailtypack: An r package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *Journal of Statistical Software*, 47: 1–28, 2012.

G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

N. Shiode, S. Shiode, E. Rod-Thatcher, S. Rana, and P. Vinten-Johansen. The mortality rates and the space-time patterns of john snow's cholera epidemic map. *International journal of health geographics*, 14(1):21, 2015.

A. L. Spivak and K. R. Damphousse. Who returns to prison? a survival analysis of recidivism among adult offenders released in oklahoma, 1985–2004. *Justice Research and Policy*, 8(2):57–88, 2006.

R. J. Sylvester, A. P. van der Meijden, W. Oosterlinck, J. A. Witjes, C. Bouffioux, L. Denis, D. W. Newling, and K. Kurth. Predicting recurrence and progression in individual patients with stage Ta T1 bladder cancer using EORTC risk tables: a combined analysis of 2596 patients from seven EORTC trials. *Eur. Urol.*, 49(3):466–465, Mar 2006.

H. S. Taffese, E. Hemming-Schroeder, C. Koepfli, G. Tesfaye, M.-c. Lee, J. Kazura, G.-Y. Yan, and G.-F. Zhou. Malaria epidemiology and interventions in ethiopia from 2001 to 2016. *Infectious diseases of poverty*, 7(1):1–9, 2018.

T. Therneau. Coxme and the laplace approximation. *Technical report*, 2018a.

T. Therneau. coxme: mixed effects cox models. r package version 2.2-10. 2018, 2018b.

A. A. Tsiatis et al. A large sample study of cox's regression model. *The Annals of Statistics*, 9(1):93–108, 1981.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

J. Vaupel, K. Manton, and E. Stallard. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16:439—-454, 1979.

A. Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20 (4):595–601, 1949.

G. C. Wei and M. A. Tanner. A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85(411):699–704, 1990.

WHO. World malaria report 2018, 2018.

A. Wienke. *Frailty Models in Survival Analysis*. Chapman & Hall/CRC Biostatistics Series. CRC Press, 2010. ISBN 9781420073911. URL https://books.google.fr/books?id=QBai7S-_2PoC.

S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, 9(1):60–62, 1938.

D. L. Wilson. The analysis of survival (mortality) data: fitting gompertz, weibull, and logistic functions. *Mechanisms of ageing and development*, 74(1-2):15–33, 1994.

A. I. Yashin and I. A. Iachine. Genetic analysis of durations: correlated frailty model applied to survival of danish twins. *Genetic epidemiology*, 12(5):529–538, 1995.

D. Yewhalaw, Y. Getachew, K. Tushune, W. Kassahun, L. Duchateau, N. Speybroeck, et al. The effect of dams and seasons on malaria incidence and anopheles abundance in ethiopia. *BMC infectious diseases*, 13(1):161, 2013.

D. Zeng and D. Lin. Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):507–564, 2007.

H. P. Zhu, X. Xia, H. Y. Chuan, A. Adnan, S. F. Liu, and Y. K. Du. Application of weibull model for survival of patients with gastric cancer. *BMC gastroenterology*, 11(1):1, 2011.

Valorisation des travaux de thèse

Article soumis

Convergent stochastic algorithm for estimation in general multivariate correlated frailty models using integrated partial likelihood, *version en ligne disponible sur arXiv*

Communications orales

Mai 2018 Journées de Statistique (JDS) à Paris-Saclay
Estimation dans des modèles de fragilité à partir d'une vraisemblance partielle

Avril 2019 Rencontres des Jeunes Statisticiens (RJS) à Porquerolles
Algorithme d'estimation convergent pour l'estimation dans des modèles de fragilité à partir d'une vraisemblance partielle intégrée

Avril 2019 Séminaire des doctorants de l'Ecole Doctorale de Mathématique Hadamard (EDMH) à Télécom Paris
Estimation dans des modèles de fragilité à partir d'une vraisemblance partielle intégrée

Juillet 2019 European Meeting of Statisticians (EMS) à Palerme
Convergent estimation algorithm for frailty models based on integrated partial likelihood

Juin 2020 Séminaire à l'Institut national de recherche en informatique et en automatique (INRIA) de Bordeaux
Analysis of time to malaria data via a spatially correlated frailty model

Appendix A

Description of the simulation procedure used to sample realizations for the unobserved frailties

We usually construct Π_η as a step of a Metropolis Hastings algorithm with proposal distribution q. Sample a candidate ξ^c:

$$\xi^c \sim q(.|\xi_{k-1}; \eta_{k-1})$$

We then calculate the acceptance ratio :

$$\alpha(\xi_{k-1}, \xi^c) = \min\left(1, \frac{\pi_{\eta_{k-1}}(\xi^c|\mathbf{X}, \mathbf{\Delta}) q(\xi_{k-1}|\xi^c; \eta_{k-1})}{\pi_{\eta_{k-1}}(\xi_{k-1}|\mathbf{X}, \mathbf{\Delta}) q(\xi^c|\xi_{k-1}; \eta_{k-1})}\right)$$

The simulated candidate is accepted with probability $\alpha(\xi_{k-1}, \xi^c)$.

$$\xi_k = \begin{cases} \xi^c & \text{with probability } \alpha(\xi_{k-1}, \xi^c) \\ \xi_{k-1} & \text{otherwise} \end{cases}$$

Maximum likelihood estimation with parametric Weibull baseline

We detail the equations used to define the maximum likelihood estimator following the choice of a Weibull baseline hazard function. The expression of the complete likelihood is given by:

$$\begin{aligned} L^{\text{weibull}}(\eta_{\text{weibull}}; \mathbf{X}, \mathbf{\Delta}, \xi) &= \prod_{i=1}^{N} g_\gamma(b_i) f_{\bar{\beta}}(\beta) \\ &\quad \times \prod_{i=1}^{N} \prod_{j=1}^{n_i} \left(\lambda \rho X_{ij}^{\rho-1} \exp(Z_{ij}^t \beta + b_i)\right)^{\Delta_{ij}} \exp(-\lambda X_{ij}^{\rho} \exp(Z_{ij}^t \beta + b_i)) \end{aligned}$$

The marginal likelihood is obtained by integrating over the extended frailty ξ:

$$L_{\text{marg}}^{\text{weibull}}(\eta; \mathbf{X}, \mathbf{\Delta}) = \int L^{\text{weibull}}(\eta_{\text{weibull}}; \mathbf{X}, \mathbf{\Delta}, \xi) d\xi$$

We denote by $\hat{\eta}_{\text{weibull}}$ the estimator of the maximum of the marginal likelihood:

$$\hat{\eta}_{\text{weibull}} = \underset{\eta_{\text{weibull}}}{\text{argmax}}\ L_{\text{marg}}^{\text{weibull}}(\eta_{\text{weibull}}; X, \Delta)$$

Appendix B

Effect of censoring on MLEs based on model $\mathcal{M}_2$ and $\mathcal{M}_3$

Effect of censoring on MLEs based on model $\mathcal{M}_2$

Table B.1: Reduction in variance based on empirical variances of the MLE of parameters from 200 repetitions of the datasets simulated following $h_{ij}(X_{ij}|b_i) = \lambda\rho X_{ij}^{\rho-1}$ exp$(Z_{i,1}(\beta_1 + b_i) + Z_{ij,2}\beta_2)$. $N = 200, \beta_0 = (2, 2), \sigma_0^2 = 0.7, \rho_0 = 1.5, \lambda_0 = 0.01$.

Parameter	**Change in J**				
	Censoring	J=4 $\rightarrow$ J=8	J=8 $\rightarrow$ J=16	J=16 $\rightarrow$ J=32	J=32 $\rightarrow$ J=64
$\hat{\beta}_1$		0.396	0.177	0.153	0.0183
$\hat{\beta}_2$		0.512	0.468	0.586	0.478
$\hat{\sigma}^2$	0 %	0.604	0.416	0.188	0.291
$\hat{\rho}$		0.516	0.485	0.578	0.475
$\hat{\lambda}$		0.502	0.442	0.429	0.351
$\hat{\beta}_1$		0.399	0.395	0.193	-0.0436
$\hat{\beta}_2$		0.624	0.533	0.407	0.654
$\hat{\sigma}^2$	40 %	0.604	0.585	0.408	0.229
$\hat{\rho}$		0.562	0.588	0.568	0.477
$\hat{\lambda}$		0.558	0.530	0.514	0.281
$\hat{\beta}_1$		0.471	0.250	0.342	0.354
$\hat{\beta}_2$		0.587	0.538	0.443	0.618
$\hat{\sigma}^2$	70 %	0.482	0.659	0.571	0.305
$\hat{\rho}$		0.524	0.535	0.429	0.546
$\hat{\lambda}$		0.518	0.535	0.360	0.545

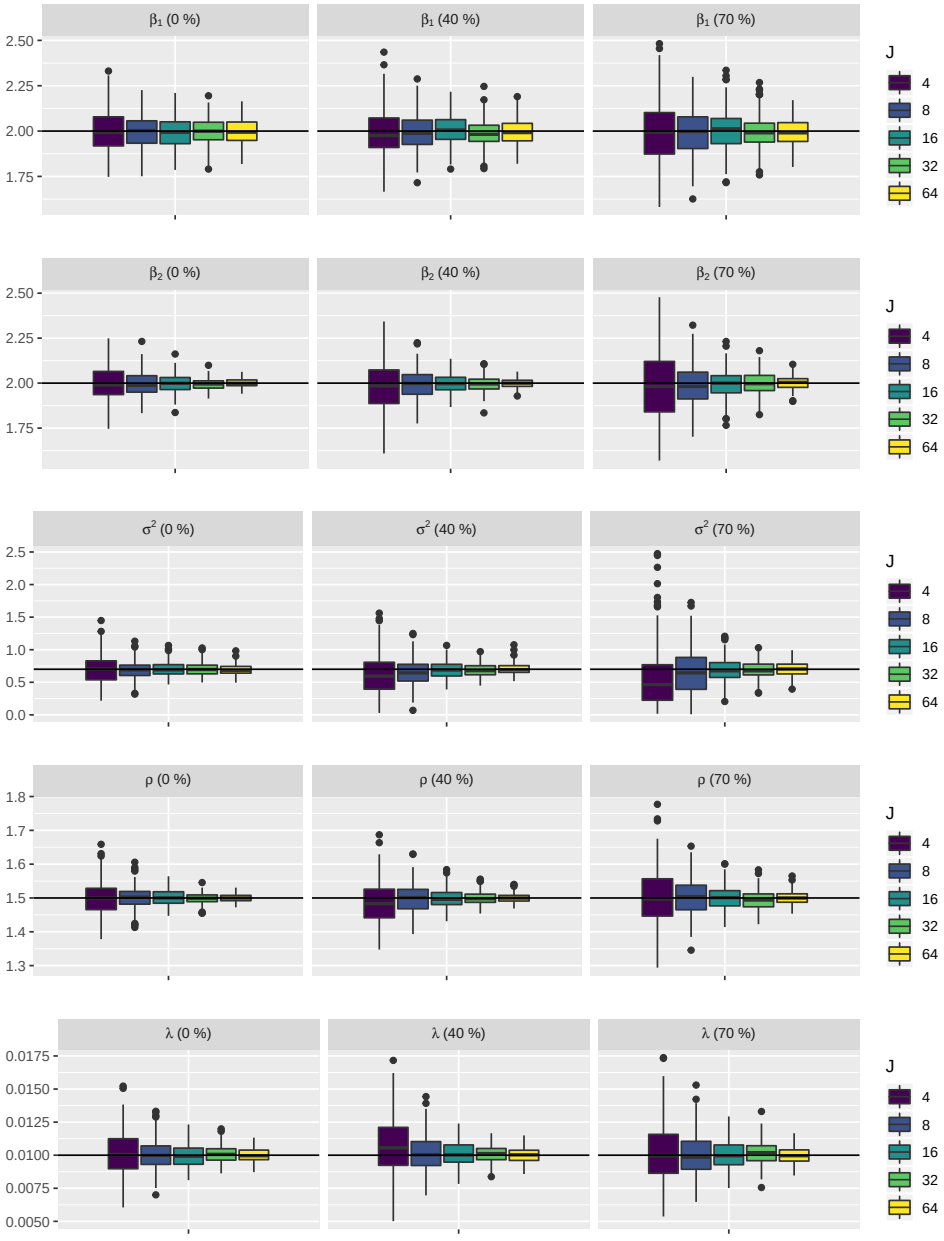

Figure B.1: MLE of parameters from 200 repetitions of the datasets simulated following $h_{ij}(X_{ij}|b_i) = \lambda\rho X_{ij}^{\rho-1}\exp(Z_{i,1}(\beta_1 + b_i) + Z_{ij,2}\beta_2)$. $N = 200, \beta_0 = (2, 2), \sigma_0^2 = 0.7, \rho_0 = 1.5, \lambda_0 = 0.01$.

Effect of censoring on MLEs based on model $\mathcal{M}_3$

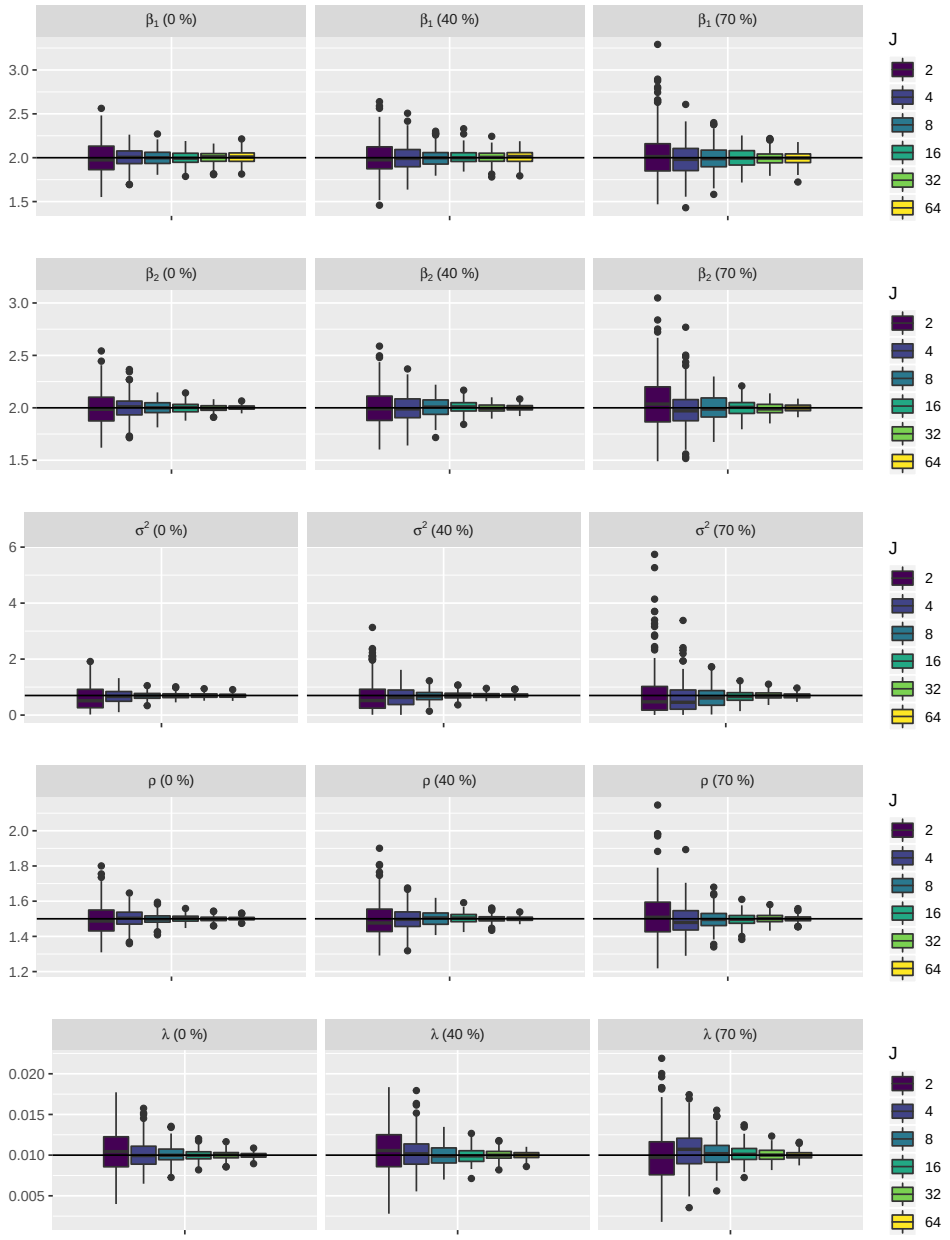

Figure B.2: MLE of parameters from 200 repetitions of the datasets simulated following $h_{ij}(X_{ij}|b_i) = \lambda\rho X_{ij}^{\rho-1}\exp(Z_{ij,1}(\beta_1+b_i)+Z_{ij,2}\beta_2)$. $N=200, \beta_0=(2,2), \sigma_0^2=0.7, \rho_0=1.5, \lambda_0=0.01$.

Table B.2: Reduction in variance based on empirical variances of the MLE of parameters from 200 repetitions of the datasets simulated following $h_{ij}(X_{ij}|b_i) = \lambda\rho X_{ij}^{\rho-1}$ exp$(Z_{ij,1}(\beta_1 + b_i) + Z_{ij,2}\beta_2)$. $N = 200, \beta_0 = (2, 2), \sigma_0^2 = 0.7, \rho_0 = 1.5, \lambda_0 = 0.01$.

Parameter	**Change in J**					
	Censoring	J=2 $\rightarrow$ J=4	J=4 $\rightarrow$ J=8	J=8 $\rightarrow$ J=16	J=16 $\rightarrow$ J=32	J=32 $\rightarrow$ J=64
$\hat{\beta}_1$		0.628	0.456	0.160	0.198	-0.109
$\hat{\beta}_2$		0.550	0.683	0.380	0.516	0.551
$\hat{\sigma}^2$	0	0.682	0.678	0.479	0.254	0.0727
$\hat{\rho}$		0.629	0.676	0.548	0.460	0.515
$\hat{\lambda}$		0.588	0.682	0.529	0.449	0.521
$\hat{\beta}_1$		0.403	0.517	0.394	0.136	0.140
$\hat{\beta}_2$		0.428	0.526	0.594	0.484	0.503
$\hat{\sigma}^2$	40	0.660	0.699	0.509	0.511	0.215
$\hat{\rho}$		0.599	0.491	0.582	0.512	0.584
$\hat{\lambda}$		0.525	0.546	0.532	0.472	0.599
$\hat{\beta}_1$		0.599	0.387	0.439	0.440	0.153
$\hat{\beta}_2$		0.545	0.524	0.593	0.455	0.614
$\hat{\sigma}^2$	70	0.667	0.515	0.724	0.546	0.460
$\hat{\rho}$		0.590	0.602	0.586	0.443	0.557
$\hat{\lambda}$		0.538	0.491	0.623	0.430	0.587

Appendix C

Assumptions for convergence of Expectation Maximization type algorithms

We state the classical assumptions **(M1)-(M5)** (cf. Delyon et al. (1999)) required to prove the almost sure convergence of EM like algorithms :

(M1) The complete data likelihood function is given by :

$$L(\theta; \mathbf{X}, \mathbf{b}) = \exp(-\Psi(\theta) + \langle \mathcal{S}(\mathbf{b}), \Phi(\theta) \rangle)$$

where $\mathcal{S}$, Ψ and Φ are Borel functions.

(M2) Define $L : \mathsf{S} \times \Theta \to \mathbb{R}$ as:

$$L(\theta; s) \triangleq -\Psi(\theta) + \langle s, \Phi(\theta) \rangle$$

The functions Ψ and Φ are twice continuously differentiable on Θ.

(M3) The function $\bar{s} : \Theta \to S$ defined as:

$$\bar{s}(\theta) = \int_{\mathbb{R}^N} \mathcal{S}(\mathbf{b}) \pi_\theta(\mathbf{b}|\mathbf{X}, \mathbf{\Delta}) d\mathbf{b}$$

is continuously differentiable on Θ.

(M4) The function $l : \Theta \to \mathbb{R}$ defined as the marginal extended log-likelihood

$$l(\theta) = \log \int_{\mathbb{R}^N} L(\theta; \mathbf{X}, \mathbf{\Delta}, \mathbf{b}) d\mathbf{b}$$

is continuously differentiable on Θ and

$$\partial_\theta \int_{\mathbb{R}^N} L(\theta; \mathbf{X}, \mathbf{\Delta}, \mathbf{b}) d\mathbf{b} = \int_{\mathbb{R}^N} \partial_\theta \, L(\theta; \mathbf{X}, \mathbf{\Delta}, \mathbf{b}) d\mathbf{b}$$

(M5) There exists a function $\hat{\theta} : \mathsf{S} \to \Theta$ such that:

$$\forall\ s \in \mathsf{S}, \forall\ \theta \in \Theta,\ L(\hat{\theta}(s), s) \geq L(\theta, s)$$

where $L : \mathsf{S} \times \Theta \to \mathbb{R}$ is defined as

$$L(\theta, s) = \exp(-\Psi(\theta) + \langle s, \Phi(\theta) \rangle) \tag{C.1}$$

Moreover, the function $\hat{\theta}$ is continuously differentiable on S.

We state assumptions **(SAEM1)-(SAEM3)**, **(C)** of Kuhn and Lavielle (2004)

(SAEM1') For all k in $\mathbb{N}$, $\mu_k \in \{0, 1\}$, $\sum_{k=1}^{\infty} \mu_k = \infty$ and there exists $\lambda \in [\frac{1}{2}, 1]$ such that $\sum_{k=1}^{\infty} \mu_k^{1+\lambda} < \infty$.

(SAEM2) $l : \Theta \to \mathbb{R}$ and $\hat{\theta} : \mathsf{S} \to \Theta$ are m times differentiable, where m is the integer such that S is an open subset of $\mathbb{R}^m$.

(SAEM3')

1. The chain $(\mathbf{b_k})_{\mathbf{k} \geq \mathbf{0}}$ takes its values in a compact subset $\mathcal{W}$ of $\mathbb{R}^N$.

2. For any compact subset V of Θ, there exists a real constant L such that for any (θ, θ') in V^2

$$\sup_{(x,y) \in \mathcal{W}^2} |\Pi_\theta(x, y) - \Pi_{\theta'}(x, y)| \leq L|\theta - \theta'|$$

3. The transition probability Π_θ generates a uniformly ergodic chain whose invariant probability is the conditional distribution $p(.|\theta)$:

$$\exists K_\theta \in \mathbb{R}^+, \exists \rho_\theta \in]0, 1[, \forall \mathbf{b} \in \mathcal{W}, \forall k \in \mathbb{N} \quad ||\Pi_\theta^k(b, .) - p(.|\theta)||_{\mathsf{TV}} \leq K_\theta \rho_\theta^k,$$

where $||.||_{\mathsf{TV}}$ denotes the total variation norm. We suppose also that:

$$K \triangleq \sup_\theta K_\theta < +\infty \quad \text{and} \quad \rho \triangleq \sup_\theta \rho_\theta < 1$$

4. The function $\mathcal{S}$ is bounded on $\mathcal{W}$.

(C) The sequence $(s_k)_{k \geq 0}$ takes its values in a compact subset of S.

université PARIS-SACLAY

ÉCOLE DOCTORALE
de mathématiques
Hadamard (EDMH)

Titre: Estimation dans des modèles de fragilité avec des structures de corrélation complexes via des algorithmes d'approximation stochastique

Mots clés: modèles de fragilité, structures de corrélation complexes, algorithmes d'approximation stochastique

Résumé: Cette thèse porte sur l'estimation dans les modèles de fragilité en analyse de survie. Notre première contribution concerne une nouvelle méthode d'estimation basée sur la vraisemblance partielle intégrée dans le modèle de fragilité. Cette méthode ne réalise aucune approximation de la vraisemblance partielle intégrée contrairement aux autres méthodes proposées dans la littérature. Nous mettons en œuvre une approximation stochastique de l'algorithme Expectation Maximization (EM) pour calculer les estimateurs du maximum de vraisemblance partielle intégrée des paramètres du modèle. De plus, nous établissons les propriétés théoriques de convergence de l'algorithme. Notre méthode permet de considérer des modèles de fragilité avec différentes structures de corrélations et une large gamme de lois de fragilité. Notre deuxième contribution porte sur l'étude des vitesses de convergence des estimateurs du maximum de vraisemblance dans les modèles à fragilités partagées paramétriques. En particulier, les vitesses diffèrent selon la factorisation de la vraisemblance conditionnelle. Nous étudions ce phénomène au travers d'une étude de simulation. Nous mettons aussi en évidence l'influence du niveau des covariables sur les vitesses de convergence à travers une étude de simulation intensive dans les modèles de fragilité et de façon analytique dans un modèle linéaire à effets mixtes. Notre troisième contribution présente un nouveau modèle de fragilité qui permet de prendre en compte les corrélations spatiales présentes dans les données. Cette nouvelle modélisation spatiale a été motivée par des données d'infection de malaria en Éthiopie. Les distances entre les enfants jouant un rôle important dans la transmission de la maladie, il s'avère judicieux de les prendre en compte dans le modèle. Une version stochastique de l'algorithme EM adaptée à ce contexte est mise en œuvre et étudiée. La méthode d'estimation est validée sur des données simulées puis mise en œuvre pour analyser les données de malaria.

Title: Estimation in frailty models with complex correlation structures through stochastic approximation algorithms

Keywords: frailty models, complex correlation structures, stochastic approximation algorithms

Abstract: This thesis deals with estimation in frailty models in survival analysis. Our first contribution concerns a new estimation method based on integrated partial likelihood in the frailty model. No approximation of the integrated partial likelihood is made as compared to other methods proposed in the literature. We implement a stochastic approximation of the Expectation Maximization (EM) algorithm to calculate the maximum of integrated partial likelihood estimators of the model parameters. We also establish the theoretical convergence properties of the algorithm. Our method allows for different correlation structures and for a wide range of frailty distributions. Our second contribution concerns the study of the convergence rates of maximum likelihood estimators in parametric shared frailty models. The convergence rates are notably different following the factorization of the conditional likelihood. We study this phenomenon via a simulation study. We also highlight the influence of the level of covariates on convergence rates analytically in a linear mixed effects model. We illustrate these differences via an intensive simulation study on a parametric frailty model. Our third contribution presents a new frailty model which takes into account spatial correlations which may be present in data. This new spatial modeling is motivated by malaria infection data collected in Ethiopia. Since the distances between children play an important role in the transmission of the disease, it may be relevant to take them into account in the model. A stochastic version of the EM algorithm adapted to this context is implemented and studied. The estimation method is validated on simulated data and then implemented to analyze the malaria data.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France