Contributions à la modélisation des données de durée en présence de censure : application à l'étude des résiliations de contrats d'assurance santé

Yohann Le Faou

École doctorale de sciences mathématiques de Paris centre

Laboratoire de Probabilités, Statistique et Modélisation, LPSM, Sorbonne Université

Thèse de doctorat

Discipline : Mathématiques

Spécialité : Statistique

présenté par

Yohann Le Faou

Contributions à la modélisation des données de durée en présence de censure : application à l'étude des résiliations de contrats d'assurance santé

---

dirigée par M. Olivier Lopez

Soutenue le 4 octobre 2019 devant le jury composé de :

Mme. Katrien Antonio	KU Leuven	Rapporteur
M. Gérard Biau	Sorbonne Université	Examinateur
M. Arnaud Cohen	Forsides	Examinateur
M. Olivier Lopez	Sorbonne Université	Directeur de thèse
M. Christian Robert	ENSAE	Rapporteur
M. Philippe Saint-Pierre	Université Toulouse III	Examinateur

Contributions à la modélisation des données de durée en présence de censure : application à l'étude des résiliations de contrats d'assurance santé

YOHANN LE FAOU

Contributions à la modélisation des données de durée en présence de censure

Application à l'étude des résiliations de contrats d'assurance santé

Laboratoire de Probabilités, Statistique et Modélisation (LPSM)
Sorbonne Université
Tours 15-25, 2ème étage
4 place Jussieu
75252 Paris Cedex 05

Forsides
52 rue de la victoire
75009 Paris

À Béatrice, Jean-luc, Morgan, et Nicolas

Remerciements

Mes premières pensées vont à mon directeur de thèse Olivier Lopez. Merci Olivier pour ton encadrement et ton soutien sans faille durant ces quatre années. Le chemin a été long, parfois tortueux, et tu m'as toujours orienté vers le bon cap, celui qui me permet aujourd'hui de conclure mon travail. Merci d'avoir partagé avec moi ta connaissance fine de la statistique, ton expérience et ta passion pour la discipline. Ce fut une aventure très enrichissante grâce à ta générosité.

Je tiens également à remercier chaleureusement Katrien Antonio et Christian Robert d'avoir accepté de rapporter ma thèse. Merci pour vos commentaires très encourageants. Christian, je garde un excellent souvenir de mon année d'études à l'ISFA qui constituait mes premiers pas dans le domaine de l'actuariat.

Je remercie également Arnaud Cohen, directeur de Forsides. Merci Arnaud de m'avoir orienté vers le domaine passionnant de la modélisation des durées, et de m'avoir ouvert des portes essentielles au lancement de mon projet de thèse. Aussi, merci d'avoir veillé à ce que j'accomplisse mon travail de recherche dans de bonnes conditions. Comme tu le sais, je garde d'excellents souvenirs de mes années passées à Forsides, et en particulier des personnes que j'y ai rencontrées.

Je remercie aussi Gérard Biau et Philippe Saint-Pierre d'avoir accepté d'être examinateurs de mon travail de thèse.

Je remercie Guillaume Gerber et Michael Trupin pour tout le temps qu'ils ont consacré à m'aider durant ma thèse. Merci Guillaume d'avoir partagé ton expertise actuarielle. Merci Michael pour ta disponibilité et ton dynamisme.

Je remercie ma tante Monique Brunet et mon oncle Heinz Weinmann qui m'ont aidé à relire ce manuscrit, le texte français pour l'une et l'anglais pour l'autre.

Je remercie tous les collègues de Forsides avec qui j'ai eu le plaisir de travailler. Il n'était pas toujours facile de comprendre quel était l'objet de mon travail, mais néanmoins vous avez toujours fait preuve d'une grande bienveillance à mon égard. Grâce à votre bonne humeur, les journées de travail sont passées bien vite. J'espère vous retrouver le

plus tôt possible.

Je remercie les doctorants du LPSM et de l'ex LSTA pour tous les bons moments que nous avons partagés. Les conférences, les séminaires, et nos discussions multiples, furent extrêmement enrichissantes. J'espère vous retrouver très vite également.

Je remercie toute l'équipe administrative du laboratoire et de l'école doctorale. Merci Louise pour ton aide durant la préparation de ma soutenance.

À l'heure de conclure mes études supérieures entamées il y a maintenant dix ans, je remercie le hasard de m'avoir mené jusqu'ici. Il a bien fait les choses une fois de plus et je suis heureux du chemin parcouru.

Je remercie Pierre de m'avoir aidé à me lancer dans le domaine de la data science il y a quelques années. Je remercie Côme pour son franc-parler sans égal, et je remercie Quentin de démontrer au quotidien que tout est possible.

Je salue tous les amis que j'ai rencontrés durant mes études, au lycée La Herdrie puis au lycée Clémenceau à Nantes, à l'ENS Rennes, à l'ISFA et enfin à Sorbonne Université. Je salue également tous les amis rencontrés dans mon parcours professionnel. En particulier, je remercie tous ceux qui ont tenu à assister à ma soutenance de thèse !

Enfin, je remercie mes parents, Béatrice et Jean-luc Le Faou, mes frères Morgan et Nicolas, pour leur soutien et leur accompagnement depuis de nombreuses années.

Sommaire

Chapitre 1

Introduction

1.1 Introduction : L'importance des modèles de durée en assurance

Dans sa structure même, l'assurance entretient un rapport particulier avec le temps, et notamment le temps long. Les acteurs du domaine ont donc dû s'adapter : aux comptables de repenser la manière de poser un bilan et de créer une comptabilité de l'assurance, aux gestionnaires de risque de définir des règles de solvabilité adaptées à travers des réformes prudentielles internationales, aux juristes de déterminer un code des assurances approprié. Enfin, aux mathématiciens et aux actuaires d'apprendre à composer avec les données de durée. L'actuaire collecte l'information en continu, il arrête le temps un instant donné pour étudier l'historique. Dans le temps long, les évènements étudiés par l'actuaire (e.g. décès, reprise du travail après un arrêt, résiliation d'un contrat d'assurance) sont rares et même souvent encore jamais observés, alors les données recueillies sont incomplètes, ou portent une information partielle. Dans l'exemple fondateur de l'étude de la durée de vie d'une population, aux applications démographiques (Graunt (1662)), viagères (Euler (n.d.)) et médicales (Bernoulli (1766)), la photographie instantanée informe sur l'état (vivant ou décédé) et l'âge (courant ou à la date du décès) de l'ensemble des individus étudiés, laissant pour ceux encore en vie l'incertitude sur la date du trépas. En ayant collecté ces informations durant les cinq dernières années, comment évaluer l'espérance de vie d'un nouveau né sans attendre, pendant plusieurs générations, que la totalité des évènements (décès) étudiés soient survenus ? De manière plus contemporaine, comment actualiser tous les ans les probabilités de décès à chaque âge à partir des informations enregistrées durant l'année passée, et ainsi incorporer aux modèles les évolutions de la

mortalité à court terme ? C'est à ce type de questions que répondent les modèles de durée en assurance.

L'étude des temps longs est ainsi une spécialité assurantielle, tant l'actuaire modélise les longues durées dans de nombreuses situations. Pour étudier la durée de vie donc (Boumezoued et al. (2017)), et construire des tables de mortalité (Guibert & Planchet (2017)) ou individualiser les prédictions (Hainaut (2018)). Pour estimer les lois d'incidence de la perte d'autonomie aussi (voir Biessy (2017) ou Guibert & Planchet (2018)), dans le cadre de l'assurance dépendance. Les durées de maintien en invalidité constituent également un enjeu de modélisation important (Lopez et al. (2016), Haberman & Pitacco (1998)), tout comme les durées de maintien au chômage (Verbelen et al. (2015)) en ce qui concerne l'assurance chômage. Enfin, on étudie également les durées de résiliation de contrats d'assurance (Milhaud (2013)), ainsi que les durées d'attente avant la survenance d'une défaillance de paiement (Hainaut & Robert (2014)) pour mesurer le risque de crédit. Tous ces travaux récents montrent l'intérêt de la communauté des actuaires pour la modélisation des durées.

Nous présentons dans cette partie une introduction aux notions clefs abordées dans notre travail de thèse. Dans la Section 1.2, nous présentons le problème de la censure des données de durée en assurance et nous introduisons la notion d'estimateur Kaplan-Meier ainsi que la notion de poids IPCW, en soulignant les avantages de l'utilisation de l'estimateur Kaplan-Meier conditionnel. Ce dernier point motive l'utilisation des poids IPCW conditionnels dans le Chapitre 4. Dans la Section 1.3, nous présentons la notion de copule comme outil de mesure de la dépendance entre les composantes d'un vecteur aléatoire. Puis nous abordons les sujets de l'estimation de copules et des copules conditionnelles, avant de faire le lien entre copules et durées. Cette introduction aux copules est préliminaire à notre travail du Chapitre 3 sur la mesure de la dépendance entre deux durées successives. Enfin, dans la Section 1.4, nous décrivons les applications des modèles de prédiction de durée en assurance, puis nous présentons les algorithmes d'arbre de régression et de forêt aléatoire, dans le cadre général dans un premier temps puis dans le cadre des données censurées à droite. Ces concepts sont l'objet de notre travail du Chapitre 4.

1.2 Les données de durée en assurance et leur exploitation

1.2.1 La censure et la troncature des données de durée

La collecte des données de durée est complexe à cause de son étalement dans le temps. Cela conduit au recueil de données tantôt incomplètes (dites censurées), tantôt sujettes à un biais d'observation (appelé troncature). Moralement, lorsque la durée d'intérêt est longue, on ne l'observe que sur un intervalle de temps $[D_{deb}, D_{fin}]$ qui n'englobe pas l'intégralité de la durée. Selon les situations pratiques, les conséquences sur l'étude de la durée sont différentes. Dans la suite, on appelle "individu observé" un individu dont on observe les caractéristiques à une date $d \in [D_{deb}, D_{fin}]$.

Dans l'exemple de l'étude de la durée de vie, la date de décès d'un individu observé est inconnue si elle intervient après la date D_{fin}. On dit alors que la durée de vie de l'individu est censurée à droite. En revanche, remarquons que bien que la période d'observation commence à une date D_{deb}, l'information sur la date de naissance d'un individu observé est toujours connue. Ainsi, l'information sur la durée de vie n'est pas censurée à gauche. De plus, un individu est observé si, et seulement si, il est toujours en vie à la date D_{deb}. Utiliser les observations disponibles revient donc à étudier la durée de vie conditionnellement à l'évènement "être encore en vie à la date D_{deb}", ce qui constitue un biais tendant à surestimer la durée de vie. On dit alors que les données recueillies sont tronquées à gauche. La situation de données tronquées à gauche et censurées à droite est représentée sur la Fig. 1.1. Planchet (2005) décrit les données, tronquées à gauche et censurées à droite, recueillies auprès des assureurs dans le but d'établir des tables de mortalité de référence pour des portefeuilles de rentiers.

Supposons maintenant que l'on ne s'intéresse plus à la durée de vie d'un humain mais à la durée de vie d'un appareil électronique fabriqué dans une usine, et supposons que depuis le début de la fabrication de l'appareil nous ayons enregistré dans une base de données l'ensemble des informations sur la durée de vie de chaque exemplaire sorti de l'usine. Alors en fixant D_{deb} à la veille de la mise en route de la chaîne de production, les données dont nous disposerions seraient toujours censurées à droite, mais ne seraient plus tronquées à gauche. Nous travaillons dans cette thèse à l'étude de données de cette nature. Précisément, nous étudions les durées de résiliation de contrats d'assurance santé souscrits par un courtier, et ce dernier ayant démarré son activité en 2006, la base de données comporte les informations sur l'ensemble des contrats souscrits depuis les débuts

de l'entreprise. L'information n'est donc pas tronquée.

Klein & Moeschberger (2006) décrivent en détails les différents types de censure et de troncature qu'il est possible de rencontrer en analyse de durée. Dans la plupart des situations actuarielles, comme celles citées dans la Section 1.1, les données de durée sont soit tronquées à gauche et censurées à droite, soit seulement censurées à droite.

Tout au long de ce manuscrit, nous nous intéressons au cas des données censurées à droite. Nous notons T la variable aléatoire réelle (v.a.r.) qui désigne la durée que nous étudions. La période d'observation débute avant que les premières réalisations de T ne se produisent, et donc T n'est pas tronquée à gauche. En revanche, la durée T n'est pas toujours observée entièrement ; on définit C la v.a.r. de censure qui rend compte du fait que la durée T n'est observée que jusqu'à une date D_{fin}. En pratique, la date de fin d'observation D_{fin} peut être causée par divers phénomènes : extraction de la base de données à une date fixée (et donc arrêt de l'observation à partir de cette date), date de fin de l'étude statistique (e.g. pour une étude médicale), arrêt du suivi d'un individu (résiliation de son contrat d'assurance prévoyance par exemple), impossibilité d'observer T à cause de la survenance d'un évènement adverse (e.g. dans l'étude du pronostic de survie à une maladie, décès dû à une autre cause que la maladie). On considère ainsi en toute généralité que la date D_{fin}, et donc C, est aléatoire. On suppose de plus que les v.a.r. T et C sont à valeurs positives et, sauf mention contraire, que T et C sont des v.a.r. continues. Le problème de censure est formalisé en définissant la durée observée $Y = \min(T, C)$ et l'indicateur binaire de la survenance de l'évènement $\delta = \mathbb{1}_{T \leq C}$. Les

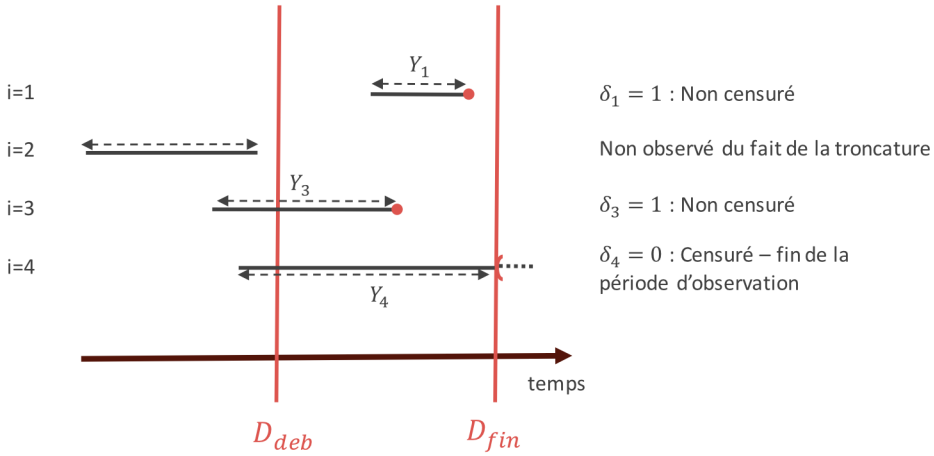

Fig. 1.1: Situation de données tronquées à gauche et censurées à droite. La donnée $i = 2$ n'est pas observée du fait de la troncature. Les données $i \in \{3, 4\}$ sont observées malgré qu'elles débutent avant la date de début d'observation.

quantités Y et δ sont biens les quantités observées dans le cadre de la censure à droite : si $\delta = 1$ (i.e. $T \leq C$) l'évènement d'intérêt est observé au temps $Y = T$, si $\delta = 0$ la donnée est censurée au temps $Y = C$.

Dans les travaux présentés dans la suite de ce manuscrit, les données d'étude ne sont donc pas constituées des informations $(T_i, C_i)_{i=1,\dots,n}$, mais d'un échantillon $(Y_i, \delta_i)_{i=1,\dots,n}$ de n réalisations indépendantes et identiquement distribuées (i.i.d.) du couple (Y, δ). Le schéma de la Fig. 1.2 résume de manière graphique les données de durée disponibles. De plus, on considère dans nos travaux que pour chaque observation i on dispose d'informations complémentaires (e.g. sur l'individu souscrivant le contrat d'assurance, ou sur le contrat lui-même) résumées dans un vecteur de variables explicatives $X_i \in \mathcal{X} \subset \mathbb{R}^d$.

1.2.2 Les outils mathématiques utilisés pour étudier les durées

Il convient dans un premier temps de définir certaines notions classiques à l'étude des durées, utilisées dans tout ce document. Nous renvoyons à l'ouvrage de Fleming & Harrington (2011) pour une introduction à tous ces concepts.

Pour U une v.a.r. positive continue représentant une durée, nous noterons $F_U(t) = P(U \leq t)$ la fonction de répartition de U, et $S_U(t) = 1 - F_U(t)$ la fonction de survie de U. De plus, le taux de risque instantané de U à l'instant t est défini par

$$\lambda_U(t) = \lim_{h \to 0} \frac{\mathrm{P}(t \leq U < t + h | U \geq t)}{h} = -\frac{S'_U(t)}{S_U(t)}$$

Il caractérise la probabilité que l'évènement U se produise dans un petit intervalle de

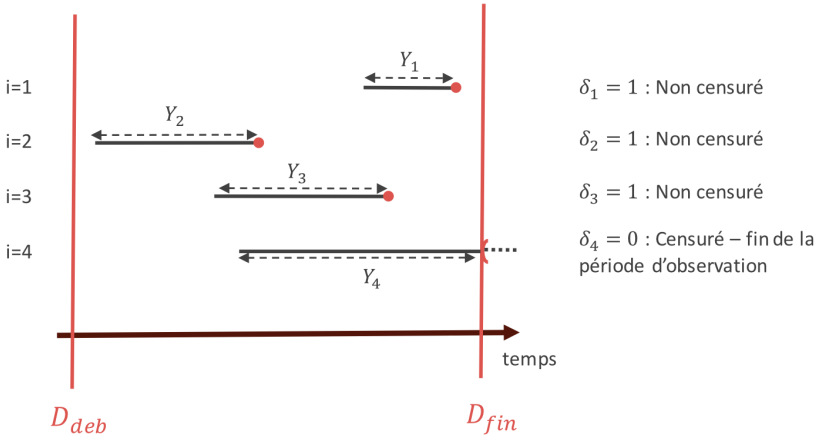

Fig. 1.2: Situation de données censurées à droite, non tronquées à gauche. La date de début d'observation précède le commencement de chaque durée.

temps après t, conditionnellement au fait que U ne se soit pas produit avant le temps t. La fonction de risque cumulé sera notée $\Lambda_U(t) = \int_{[0,t[} \lambda_U(s)ds$. On rappel que l'on a alors la relation $S_U(t) = \exp(-\Lambda_U(t))$.

Dans la situation de censure de la durée T par la variable de censure C, nous noterons également $\tau = \inf\{t \geq 0 : P(C > t) = 0\}$ la valeur maximale observable pour la durée T. De plus, $\mathcal{L}(T|X)$ désignera la loi conditionnelle de T sachant X.

1.2.3 L'estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier (Kaplan & Meier (1958a)) est l'outil de base de la statistique pour estimer de manière non paramétrique la distribution d'une v.a.r. T censurée à droite. C'est donc l'équivalent, pour une durée, de la fonction de répartition empirique. On définit

$$\hat{S}_T(t) = \prod_{Y_i \leq t} \left(1 - \frac{\delta_i}{\sum_{j=1}^n \mathbb{1}_{Y_j \geq Y_i}}\right), \tag{1.1}$$

l'estimateur de Kaplan-Meier (KM) de la fonction de survie de T, en utilisant les observations $(Y_i, \delta_i)_{i=1,\ldots,n}$ définies au 1.2.1. Dans le cas où les variables T et C ne sont pas supposées continues, plusieurs évènements ($\delta = 1$) ou censures ($\delta = 0$) peuvent se produire au même instant et la formule (1.1) n'est pas utilisée. Soient $t_1 < \ldots < t_H$ les instants distincts où un évènement est survenu. Pour $i = 1, \ldots, H$, on définit $d_i = \sum_{j=1}^n \delta_j \mathbb{1}_{Y_j = t_i}$ le nombre d'évènements survenus à l'instant t_i, et $n_i = \sum_{j=1}^n \mathbb{1}_{Y_j \geq t_i}$ le nombre d'observations à risque à l'instant t_i. Dans ce cas de figure, l'estimateur

$$\hat{S}_T(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{1.2}$$

généralise l'estimateur de Kaplan-Meier défini au (1.1).

Considérons l'hypothèse suivante, qui est essentielle pour l'étude de la convergence de l'estimateur KM.

Hypothèse

H0 *: T est indépendant de C.*

Dans certaines situations, il est possible de faire une hypothèse légèrement plus faible que l'hypothèse H0. Soit l'hypothèse ***H0'*** *: $P(C > T|T) = S_C(T)$. Remarquons que H0 implique H0'. Pour obtenir les égalités qui font intervenir les poids IPCW, que nous présentons dans la section suivante, nous verrons que l'hypothèse H0' est suffisante.*

L'hypothèse H0 constitue une hypothèse d'identifiabilité du modèle, qu'il est en général impossible de tester. En effet, Tsiatis (1975) a montré que pour toute distribution de données censurées (Y, δ), il existe un couple de variables aléatoires indépendantes (T, C) pour lequel la distribution censurée $(\min(T, C), \mathbb{1}_{T \leq C})$ possède la même loi que (Y, δ). Le problème de la dépendance entre la variable de censure et la durée d'intérêt a également été étudié par Lagakos (1979).

Stute & Wang (1993) ont montré le théorème suivant concernant l'estimateur KM.

Théorème 1 *Supposons que l'hypothèse H0 soit vérifiée. Alors en notant* $\tau = \inf\{t \geq 0 : \mathrm{P}(C > t) = 0\}$, *on a*

$$\sup_{t<\tau} \left| \hat{S}_T(t) - S_T(t) \right| \underset{n\to\infty}{\to} 0.$$

Notons que dans le cas où l'on suppose C continue, $\tau = +\infty$ et donc la convergence uniforme a lieu pour tout $t \geq 0$. Ce théorème est l'équivalent du Théorème de Glivenko-Cantelli qui existe pour les v.a.r. réelles.

Toujours sous l'hypothèse H0, Gill (1983) a établi la convergence en loi, ainsi que la normalité asymptotique, de l'estimateur de Kaplan-Meier. D'autres contributions importantes concernant la compréhension de l'estimateur KM ont été apportées par Stute (1995) ou encore Akritas et al. (2000).

L'estimateur KM est le modèle de durée le plus utilisé par les actuaires praticiens. Il intervient dans toutes les applications actuarielles qui requièrent la modélisation de durées, comme la construction et la certification des tables de mortalité (Planchet (2005), Institut des actuaires (2006)) et des lois de maintien en incapacité de travail et en invalidité (Aubin & Rolland (2010)). Dans les approches standards, l'estimateur KM est utilisé pour estimer les taux bruts de mortalité dans la population totale ou dans une sous-population (par tranches d'âge par exemple), qui sont retraités par la suite en utilisant des lissages ou des splines (Planchet & Winter (2010)).

Dans la partie suivante, nous présentons une interprétation de l'estimateur KM en termes de pondération des observations. Cela nous permet d'introduire la notion de poids IPCW (Inverse Probability of Censoring Weighting).

1.2.4 Les poids IPCW et l'estimateur de Kaplan-Meier

Les sauts de l'estimateur de Kaplan-Meier

La forme de l'estimateur KM $\hat{S}_T$ défini en (1.1) implique que $\hat{S}_T$ est une fonction décroissante et constante par morceaux dont les sauts se produisent aux temps correspondant à des

observations non censurées. On peut donc se demander quel est le poids attribué par l'estimateur KM à chaque observation non censurée. Une idée naturelle serait d'attribuer un poids nul aux observations censurées et un poids uniforme aux observations non censurées, ce qui correspondrait à ignorer dans l'estimation les observations censurées. Le problème est que cette idée conduirait à estimer la loi de T sachant $\delta = 1$, c'est à dire la loi $\mathcal{L}(T|T \leq C)$. La distribution de T serait donc sous-estimée par une telle méthode.

En remarquant que la situation de censure de T par C est symétrique à la censure de C par T, on peut définir

$$\hat{S}_C(t) = \prod_{Y_i \leq t} \left(1 - \frac{1-\delta_i}{\sum_{j=1}^n \mathbb{1}_{Y_j \geq Y_i}}\right),$$

l'estimateur KM de S_C. Alors en notant $\hat{S}_Y(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Y_i > t}$ la fonction de survie empirique de Y, on a la relation

$$\hat{S}_T \cdot \hat{S}_C = \hat{S}_Y,$$

ou en différentiant,

$$d\hat{S}_T \cdot \hat{S}_C + \hat{S}_T \cdot d\hat{S}_C = d\hat{S}_Y. \tag{1.3}$$

En un point $t = Y_i$ correspondant à une observation non censurée ($\delta_i = 1$), on a $d\hat{S}_Y(t) = \frac{1}{n}$ et $d\hat{S}_C(t) = 0$, ainsi on déduit de (1.3) que le saut de $\hat{S}_T$ en t vaut $d\hat{S}_T(t) = \frac{1}{n\hat{S}_C(t)}$. L'estimateur KM peut donc s'exprimer

$$\hat{S}_T(t) = \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\hat{S}_C(Y_i)} \mathbb{1}_{Y_i > t}. \tag{1.4}$$

La forme additive (1.4), analysée par Satten & Datta (2001), permet de faire l'analogie entre l'estimateur KM et la fonction de répartition empirique d'un v.a.r. non censurée. Là où on attribue un poids $1/n$ à chaque observation dans le cas non censuré, le poids a la forme $\hat{W}_i = n^{-1} \cdot \delta_i / \hat{S}_C(Y_i)$ en présence de censure. De façon heuristique, le poids $\hat{W}_i$ sert à compenser l'effet de la censure, qui tend à raréfier l'observation de grandes valeurs de T. On peut ainsi remarquer que l'expression $n^{-1} \cdot \delta_i / \hat{S}_C(Y_i)$ est croissante en Y_i, c'est à dire que le poids associé à une observation non censurée est plus important lorsque Y_i est grand. Remarquons également que bien que le poids des observations censurées soit nul, ces dernières interviennent dans l'estimation de $\hat{S}_C$, et impactent donc les poids $\hat{W}_i$.

Les poids IPCW

L'exemple de l'estimateur KM nous permet d'introduire la notion de poids IPCW (Inverse Probability of Censoring Weighting) qui est centrale dans notre travail de thèse. En effet, le poids $\hat{W}_i$ peut s'interpréter comme un estimateur du poids $W_i = n^{-1}\delta_i/S_C(Y_i)$. Or on a le résultat suivant.

Proposition 1 *Soit ψ une fonction réelle de support inclus dans $[0, \tau[$, avec $\tau = \inf\{t \geq 0 : P(C > t) = 0\}$. Alors sous l'hypothèse H0',*

$$E\left[\frac{\delta}{S_C(Y)} \cdot \psi(Y)\right] = E[\psi(T)]. \tag{1.5}$$

Preuve. Notons que

$$E\left[\frac{\delta}{S_C(Y)} \cdot \psi(Y)\right] = E\left[\frac{\mathbb{1}_{T\leq C}}{S_C(T)}\psi(T)\right].$$

Puis, en conditionnant par T dans la partie gauche de l'égalité, on obtient

$$E\left[\frac{\mathbb{1}_{T\leq C}}{S_C(T)}\psi(T)\right] = E\left[\frac{\psi(T)}{S_C(T)}E[\mathbb{1}_{T\leq C}|T]\right].$$

D'après notre hypothèse, $E[\mathbb{1}_{T\leq C}|T] = P(C \geq T|T) = S_C(T)$, ce qui conclut la preuve. ■

Cette proposition montre qu'il est possible d'estimer la distribution de T à partir des données censurées $(Y_i, \delta_i)_{i=1,\dots,n}$ moyennant une condition sur la dépendance entre T et C. Pour cela, il suffit d'allouer un poids $W_i = n^{-1}\delta_i/S_C(Y_i)$ à chaque observation Y_i. Pour une observation non censurée on a, en supposant l'hypothèse H0', $\delta_i/S_C(Y_i) = 1/P(\delta_i = 1|T_i)$. Le poids W_i correspond donc à l'inverse de la probabilité d'être observée sachant la valeur de T_i.

La notion de poids IPCW et son utilisation pour étudier les données censurées a été introduite par Van der Laan & Robins (2003). Notons que des approches similaires apparaissaient déjà auparavant dans la littérature, par exemple dans l'article de Koul et al. (1981), ou celui de Stute (1999). Le concept de poids IPCW est essentiel à notre travail de thèse puisqu'il est utilisé dans nos deux principales contributions, présentées aux Chapitres 3 et 4. Dans la section suivante nous abordons la question de la modélisation des données censurées en présence de variables explicatives X.

1.2.5 L'estimateur de Kaplan-Meier et les poids IPCW conditionnels

Les hypothèses faites dans le cas conditionnel

Nous avons vu dans les Sections 1.2.3 et 1.2.4 qu'il est possible d'estimer la fonction de survie de T sous réserve que l'hypothèse H0 soit satisfaite. Nous allons voir maintenant que si l'on observe, en plus de T et C, des variables explicatives X, des hypothèses similaires à H0 sont nécessaires pour pouvoir estimer la distribution conditionnelle de T sachant X. Définissons les hypothèses H1 et H2 suivantes.

Hypothèse

H1 *: (T,X) est indépendant de C,*

H2 *: T est indépendant de C conditionnellement à X.*

Comme pour l'hypothèse H0, il est parfois possible d'utiliser des hypothèses légèrement plus faibles que H1 et H2. On définit les hypothèses H1' et H2' par : ***H1'*** *:* $P(C > T|T, X) = S_C(T)$*, et* ***H2'*** *:* $P(C > T|T, X) = S_C(T|X)$*, où l'on a utilisé* $S_C(t|X) = P(C > t|X)$ *la fonction de survie de C conditionnelle à X. Remarquons que H1 (resp. H2) implique H1' (resp. H2') et donc que lorsque nous mentionnons un résultat obtenu sous l'hypothèse H1' (resp. H2'), il est également vrai sous l'hypothèse H1 (resp. H2).*

Il est nécessaire de faire l'hypothèse H1 ou l'hypothèse H2 si l'on souhaite étudier la loi de T sachant X. En effet, H1 et H2 sont des hypothèses d'identifiabilité du modèle, que l'on ne peut pas tester en général. Selon le contexte, nous nous placerons dans cette thèse, tantôt sous l'hypothèse H1, tantôt sous l'hypothèse H2. Notons que H1 implique H2, ainsi l'hypothèse H1 est plus forte que l'hypothèse H2.

Illustrons la différence entre ces deux hypothèses en prenant l'exemple de l'usine fabriquant des appareils électroniques (supposons que ce soient des écrans), que nous avions déjà évoqué dans la Section 1.2.1. Dans cet exemple, T désigne la durée de vie de l'écran, C est l'ancienneté de l'écran (différence entre la date d'aujourd'hui et la date de sortie de l'usine), et supposons que X désigne le modèle d'écran fabriqué. On suppose également que pour un modèle donné, la qualité des écrans produits n'évolue pas dans le temps (pas d'usure du matériel utilisé sur la chaîne de production). L'hypothèse H1 correspond au cas où l'usine fabrique les mêmes écrans depuis son lancement. Si en revanche, le modèle d'écran fabriqué a changé au cours du temps, alors C et X ne sont pas indépendants,

et l'hypothèse H1 n'est pas vérifiée. Néanmoins, dans ce dernier cas, l'hypothèse H2 est vérifiée et il est donc possible d'étudier la durée de vie des écrans conditionnellement à X en faisant l'hypothèse H2. Notons enfin que dans ce dernier cas de figure, la présence de covariables X permet aussi d'estimer la distribution (non conditionnelle) de T, ce qui ne serait pas possible autrement car l'hypothèse H0 du paragraphe 1.2.3 ne serait pas vérifiée.

L'estimateur de Kaplan-Meier conditionnel

L'estimateur de Kaplan-Meier conditionnel (KM conditionnel) proposé par Beran (1981) est défini par

$$\hat{S}_T(t|X=x) = \prod_{Y_i \leq t} \left(1 - \frac{\delta_i w_{i,n}(x)}{\sum_{j=1}^n w_{j,n}(x) \mathbb{1}_{Y_j \geq Y_i}}\right), \tag{1.6}$$

où $w_{i,n}(x)$ est un poids déterminé par l'équation

$$w_{i,n}(x) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^n K\left(\frac{X_j - x}{h}\right)}.$$

Ici, K désigne une fonction noyau (i.e. une fonction réelle positive, symétrique, et telle que $\int K(u)du = 1$). Sous l'hypothèse H1, la convergence de l'estimateur (1.6) vers la fonction de survie conditionnelle $S_T(\cdot|X=x)$ a été étudiée par Stute (1993), qui a également établi la normalité asymptotique (Stute (1996)). Un résultat de convergence uniforme sous H2 est démontré dans Dabrowska (1989). La convergence de l'estimateur KM conditionnel sous H2 est également étudiée dans Van Keilegom & Akritas (1999), Van Keilegom & Veraverbeke (2001) et Van Keilegom et al. (2001).

L'estimateur KM conditionnel vérifie lui aussi l'égalité $\hat{S}_T(\cdot|X) \cdot \hat{S}_C(\cdot|X) = \hat{S}_Y(\cdot|X)$, en définissant $\hat{S}_C(\cdot|X)$ de manière symétrique à (1.6) et $\hat{S}_Y(t|X=x) = \frac{1}{n}\sum_{i=1}^n w_{i,n}(x)\mathbb{1}_{Y_i>t}$. Il admet donc une représentation sous forme de somme similaire à celle de l'équation (1.4):

$$\hat{S}_T(t|X=x) = \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\hat{S}_C(Y_i|X_i)}\mathbb{1}_{Y_i>t}.$$

Les poids IPCW conditionnels

L'égalité (1.5) se généralise au cas conditionnel. Selon que l'on suppose l'hypothèse H1' ou H2', notons $W = \delta/S_C(Y)$ (si l'on suppose H1') ou $W = \delta/S_C(Y|X)$ (si l'on suppose

H2'). On montre, en utilisant la même technique que pour la Proposition 1, que

$$E\left[W \cdot \psi(Y, X)\right] = E\left[\psi(T, X)\right], \tag{1.7}$$

pour toute fonction ψ telle que $\forall x \in \mathcal{X}$, $\text{supp}(\psi(\cdot, x)) \subset [0, \tau(x)[$, avec $\tau(x) = \inf\{t \geq 0 : P(C > t|X = x) = 0\}$ et en notant $\text{supp}(f)$ le support d'une fonction f. Sous H1', le principe des poids IPCW est donc toujours valable en utilisant les poids $W_i = n^{-1}\delta_i/S_C(Y_i)$. C'est également le cas sous H2', mais à condition que les poids $W_i = n^{-1}\delta_i/S_C(Y_i|X_i)$ soient estimés conditionnellement à X.

De manière similaire à ce que nous avons observé dans le cas non conditionnel, l'égalité (1.7) montre que sous l'hypothèse H1' ou H2', il est possible d'estimer la loi jointe du couple (T, X). Pour cela, il suffit de considérer les observations (Y_i, X_i) pour lesquelles $\delta_i = 1$, et de pondérer ces observations par le poids W_i. On utilise les poids non conditionnels $W_i = \delta_i/S_C(Y_i)$ sous l'hypothèse H1', et les poids conditionnels $W_i = \delta_i/S_C(Y_i|X_i)$ sous H2'. En pratique, la fonction de survie de la variable de censure S_C (resp. fonction de survie conditionnelle $S_C(\cdot|X)$) est inconnue, et on utilise donc un estimateur $\hat{S}_C$ (resp. $\hat{S}_C(\cdot|X)$) pour estimer les poids par $\hat{W}_i = \delta_i/\hat{S}_C(Y_i)$ (resp. $\hat{W}_i = \delta_i/\hat{S}_C(Y_i|X_i)$). L'hypothèse H2' est moins contraignante que H1', néanmoins elle nécessite l'estimation de la fonction de survie conditionnelle $S_C(\cdot|X)$, qui est plus complexe.

L'approche IPCW permet de construire des modèles de prédiction de la durée T en fonction X. Cette méthode est utilisée par Koul et al. (1981) pour effectuer la régression linéaire d'une durée, par Molinaro et al. (2004) pour construire des arbres de régression, ou encore par Goldberg & Kosorok (2017) pour utiliser l'algorithme de Machine à vecteurs de support (SVM) sur des données censurées à droite. Dans le domaine de l'assurance, la technique IPCW est utilisée pour résoudre des problèmes de régression dans Lopez et al. (2016) et Lopez (2018).

Dans une situation où l'hypothèse H0 n'est pas satisfaite, la présence de covariables X peut permettre l'estimation de la distribution de T dans le cas où H2 est vérifiée (car H2 peut être vraie sans que H0 ne le soit). Ceci est illustré par Ferger et al. (2017) (voir le chapitre 2) qui utilisent l'estimateur KM conditionnel pour estimer la distribution de T sous l'hypothèse H2.

En pratique, nous avons vu qu'il est en général difficile de tester les hypothèses H1 et H2, et comme l'hypothèse H2 est la moins forte, utiliser H2 apparaît souvent comme la solution la plus prudente. Dans le cas d'application que nous étudions dans cette thèse (voir le Chapitre 2), la censure C correspond à l'ancienneté d'un contrat d'assurance, alors

que T correspond à la durée de résiliation du contrat. La situation est donc similaire à l'exemple de l'usine évoqué ci-dessus, dans le sens où dans les deux cas de figure l'aléa de la variable C porte sur la date de début du risque (sortie de l'usine ou prise d'effet du contrat), la date de fin étant la même pour toutes les observations (date d'extraction de la base). Supposer H1 revient donc à supposer que la date de début du risque est indépendante de T et X. Si les caractéristiques X des contrats signés ont évolué dans le temps on ne pourra pas supposer H1, en revanche il sera possible de supposer H2 en général. L'influence du choix de l'hypothèse H1 ou H2 sur le résultat de la régression de T à partir de variables X est un sujet traité au Chapitre 4.

1.3 L'étude de la dépendance en présence de censure : application à l'assurance

1.3.1 Une introduction aux copules

En statistique, les copules sont centrales dans l'étude de la dépendance entre deux (copules bivariées) ou plusieurs (copules multivariées) variables aléatoires. Dans le cas bivarié, la dépendance entre deux variables V_1 et V_2 désigne l'ensemble des liens qui existent entre les deux variables. Étant donné une information sur la variable V_1, quelle est la conséquence pour la variable V_2 ? Si aucune information concernant V_1 n'a de conséquence sur la distribution de V_2, alors les variables V_1 et V_2 sont dites indépendantes. À contrario, s'il existe une information qui concerne V_1 et qui modifie la valeurs attendue pour V_2, alors les variables V_1 et V_2 sont dites dépendantes. Le Théorème de Bayes (1763) permet de montrer que la relation de dépendance statistique, définie comme ci-dessus, est symétrique, c'est à dire que si une telle information existe pour V_1 alors elle existe également pour V_2. On peut donc bien parler de dépendance, et d'indépendance, entre deux variables V_1 et V_2.

La notion de copule, introduite par Sklar (1959), permet de caractériser la dépendance entre les variables d'un vecteur aléatoire $V = (V_1, \ldots, V_p)$. L'ensemble des copules multivariées $\mathfrak{C} : [0,1]^p \to [0,1]$ correspond à l'ensemble des fonctions de répartition associées à un vecteur aléatoire $(U_1, \ldots, U_p)$ dont les lois marginales sont uniformes sur $[0,1]$. Le Théorème de Sklar (1959) établit que la fonction de répartition jointe $F(v_1, \ldots, v_p) = P(V_1 \leq v_1, \ldots, V_p \leq v_p)$ peut s'écrire

$$F(v_1, \ldots, v_p) = \mathfrak{C}(F_{V_1}(v_1), \ldots, F_{V_p}(v_p)), \tag{1.8}$$

avec F_{V_k} $(k = 1, \dots, p)$ la fonction de répartition de V_k, et $\mathfrak{C}$ une copule. De plus, la copule $\mathfrak{C}$ est unique si les lois marginales de V sont continues. En effet, dans ce cas de figure les fonctions F_{V_k} $(k = 1, \dots, p)$ sont inversibles et pour $(u_1, \dots, u_p) \in [0, 1]^p$ la copule $\mathfrak{C}$ est donnée par

$$\mathfrak{C}(u_1, \dots, u_p) = F(F_{V_1}^{-1}(u_1), \dots, F_{V_p}^{-1}(u_p)). \tag{1.9}$$

Dans la factorisation (1.8), toute l'information sur la dépendance entre les variables du vecteur $(V_1, \dots, V_p)$ est contenue dans la copule $\mathfrak{C}$. A ce titre, la copule $\mathfrak{C}$ caractérise la dépendance dans le vecteur V et il arrive qu'on l'appelle *fonction de dépendance*. Les ouvrages Joe (1997), Nelsen (2007), ou encore Charpentier (2013) sont des introductions au concept de copule.

Les copules ont connu un essor important à partir des années 2000, et on les trouve aujourd'hui dans de nombreux domaines d'application. En assurance, les copules sont utilisées pour agréger les risques, étudier la dépendance entre des risques extrêmes ou des risques de catastrophe naturelle, ou encore prévoir l'évolution de la mortalité. Dans la Section 1.3.5, nous revenons en détails sur les utilisations des modèles multivariés dans le domaine de l'assurance. Les copules sont également très utilisées en finance comme indiqué par Genest et al. (2013), par exemple pour des applications au trading haute fréquence dans Dias et al. (2004) ou à la gestion de portefeuille dans Patton (2004). En théorie de la fiabilité, Benoumechiara et al. (2018) proposent une estimation conservative des risques, qui requiert la prise en compte des dépendances entre les évènements adverses. L'étude des corrélations entre les événements extrêmes est aussi nécessaire en hydrologie, où Favre et al. (2004) s'intéressent aux fortes crues.

1.3.2 Familles de copules usuelles et mesures de concordance

Nous abordons dans cette section différentes notions utilisées au Chapitre 3 de notre travail, en présentant les familles de copules les plus classiques, ainsi que les mesures de concordance non paramétriques les plus courantes. Un traitement complet des éléments présentés ici est proposé dans le livre de Nelsen (2007). Pour une copule $\mathfrak{C}$, nous notons

$$c(u_1, \dots, u_p) = \frac{\partial^p}{\partial u_1 \dots \partial u_p} \mathfrak{C}(u_1, \dots, u_p)$$

la densité de copule associée.

Les Copules archimédiennes

Dans cette section, nous nous plaçons dans le cas bivarié pour simplifier la présentation. Étant donné une fonction décroissante convexe $\phi : (0,1] \to [0,+\infty[$ telle que $\phi(1) = 0$ et $\lim_{t\to 0} \phi(t) = +\infty$, la copule archimédienne de générateur ϕ, telle que définie par Genest & MacKay (1986a,b), est la copule donnée par la formule

$$\mathfrak{C}(u_1, u_2) = \phi^{-1}(\phi(u_1) + \phi(u_2)), \tag{1.10}$$

avec $(u_1, u_2) \in [0,1]^2$. Les copules archimédiennes sont symétriques, dans le sens où elles vérifient $\mathfrak{C}(u_1, u_2) = \mathfrak{C}(u_2, u_1)$.

La copule de Clayton de paramètre $\theta \in]-1, +\infty[\backslash\{0\}$, introduite dans Clayton (1978), est obtenue en considérant la fonction $\phi(t) = t^{-\theta} - 1$ dans la construction ci-dessus. On obtient alors

$$\mathfrak{C}_\theta(u_1, u_2) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}.$$

Les cas limites de la copule de Clayton, pour $\theta \to 0$, $\theta \to -1$, et $\theta \to +\infty$ sont intéressants à noter. On prolonge souvent la copule de Clayton en $\theta = 0$ en prenant $\phi(t) = -\log(t)$ dans (1.10). On trouve alors $\mathfrak{C}^{\perp\!\!\!\perp}(u_1, u_2) = u_1 u_2$ qui correspond à la copule du couple (U_1, U_2) lorsque U_1 et U_2 sont indépendants. Pour $\theta \to +\infty$, on trouve la copule comonotone (ou copule de dépendance positive maximale), donnée par $\mathfrak{C}^+(u_1, u_2) = \min(u_1, u_2)$. Elle caractérise la dépendance du couple (U_1, U_2) lorsque $U_2 = U_1$. A l'opposé, pour $\theta \to -1$, on trouve la copule anticomonotone (ou copule de dépendance négative maximale) $\mathfrak{C}^-(u_1, u_2) = \max(0, u_1 + u_2 - 1)$, qui représente la dépendance du couple (U_1, U_2) lorsque $U_2 = 1 - U_1$. La copule comonotone (resp. la copule anticomonotone) correspond à la borne supérieure (resp. inférieure) de Fréchet-Hoeffding de l'ensemble des copules, c'est à dire que pour toute copule $\mathfrak{C}$ on a

$$\mathfrak{C}^-(u_1, u_2) \leq \mathfrak{C}(u_1, u_2) \leq \mathfrak{C}^+(u_1, u_2).$$

La copule de Gumbel de paramètre $\theta \geq 1$, introduite par Gumbel (1960), s'obtient en prenant $\phi(t) = (-\log(t))^\theta$. Elle s'exprime donc

$$\mathfrak{C}_\theta(u_1, u_2) = \exp\left[-\left((-\log(u_1))^\theta + (-\log(u_2))^\theta\right)^{1/\theta}\right].$$

Une particularité de la copule de Gumbel est d'être max-stable, c'est à dire que pour tout $t \geq 0$, $(\mathfrak{C}_\theta(u_1, u_2))^t = \mathfrak{C}_\theta(u_1^t, u_2^t)$. Cela lui confère un rôle particulier dans l'étude de la

dépendance entre évènements extrêmes.

La copule de Frank de paramètre $\theta \neq 0$, introduite par Frank (1979), est la copule archimédienne de générateur $\phi(t) = -\log\big((e^{-\theta t}-1)/(e^{-\theta}-1)\big)$. Elle est donnée par la formule

$$\mathfrak{C}_\theta(u_1,u_2) = -\frac{1}{\theta}\log\left[1+\frac{(\exp(-\theta u_1)-1)(\exp(-\theta u_2)-1)}{\exp(-\theta)-1}\right].$$

Ici aussi, on peut prolonger la famille de copule en $\theta = 0$ par la copule d'indépendance $\mathfrak{C}^{\perp\!\!\!\perp}$. La copule de Frank est la seule copule archimédienne qui vérifie la propriété de symétrie radiale, c'est à dire que sa densité de copule vérifie $c(u_1,u_2) = c(1-u_1,1-u_2)$ (symétrie centrale par rapport au point $(1/2,1/2)$).

De nombreuses autres familles de copules archimédiennes sont présentées dans Nelsen (2007). Voir notamment le tableau récapitulatif p. 116.

Les copules elliptiques

Pour $\Sigma \in M_p(\mathbb{R})$ une matrice carrée symétrique définie positive, $\mu = (\mu_1,\dots,\mu_p) \in \mathbb{R}^p$, et $g : [0,+\infty[\to [0,+\infty[$ une fonction telle que $\int_{\mathbb{R}^p} g(\|v\|^2)dv = 1$, la loi elliptique $\xi(\mu,\Sigma,g)$ est la loi sur $\mathbb{R}^p$ donnée par la densité

$$f_{\mu,\Sigma,g}(x) = (\det\Sigma)^{-1/2} g\left({}^t(x-\mu)\Sigma^{-1}(x-\mu)\right). \tag{1.11}$$

Cette loi est dite elliptique car les courbes de niveau de la densité $f_{\mu,\Sigma,g}$ sont des ellipsoïdes (à la condition que g ait une forme non dégénérée, i.e. ne soit pas constante sur aucun intervalle). On définit les copules elliptiques comme les copules associées aux vecteurs aléatoires $(V_1,\dots,V_p)$ qui suivent une loi elliptique. Grâce à la formule (1.9), ces copules s'expriment à partir de la fonction de répartition jointe F et des fonctions de répartition marginales $(F_{V_k})_{k=1,\dots,p}$.

On obtient la famille des copules gaussiennes lorsque $g(t) = (2\pi)^{-p/2}\exp(-t/2)$, $\mu = 0$ et Σ décrit l'ensemble des matrices de corrélation (i.e. Σ est une matrice de covariance qui n'a que des 1 sur sa diagonale). Alors, les densités $f_{\mu,\Sigma,g}$ correspondent à l'ensemble des densités gaussiennes centrées multivariées de lois marginales $\mathcal{N}(0,1)$, où l'on note $\mathcal{N}(0,1)$ la loi gaussienne centrée réduite. Soit Φ la fonction de répartition de la loi gaussienne centrée réduite et Φ_Σ la fonction de répartition de la loi gaussienne multivariée de corrélation Σ, la copule gaussienne de corrélation Σ est donnée par

$$\mathfrak{C}_\Sigma(u_1,\dots,u_p) = \Phi_\Sigma(\Phi^{-1}(u_1),\dots,\Phi^{-1}(u_p)).$$

En dérivant cette formule, on obtient que la densité de la copule gaussienne s'exprime

$$c_\Sigma(u_1,\dots,u_p) = \det(\Sigma)^{-1/2}\exp\left(-1/2\ {}^t\beta(\Sigma^{-1}-I_p)\beta\right),$$

avec $\beta = {}^t(\Phi^{-1}(u_1),\dots,\Phi^{-1}(u_p))$.

En dimension 2, la copule gaussienne possède un unique paramètre $\theta \in]-1,1[$ car la matrice de corrélation Σ vaut

$$\Sigma_\theta = \begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix}.$$

La densité de copule est alors égale à

$$c_\theta(u_1,u_2) = \frac{1}{1-\theta^2}\exp\left(-\frac{\theta^2\Phi^{-1}(u_1)^2+\theta^2\Phi^{-1}(u_1)^2-2\theta\Phi^{-1}(u_1)\Phi^{-1}(u_2)}{2(1-\theta^2)}\right).$$

Une autre famille importante de copules elliptiques est constituée des copules de Student (ou *t-copula*). On les obtient en considérant les fonctions

$$g_\nu(t) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)(\pi\nu)^{1/2}}\,(1+t/\nu)^{-(\nu+1)/2}$$

dans la formule (1.11), où $\nu > 0$ correspond aux nombres de degrés de liberté de la loi de Student et Γ désigne la fonction gamma.

Sur la Fig. 1.3, nous avons représenté des nuages de points simulés avec les cinq familles de copule présentées ci-dessus.

Les principales mesures de concordance entre variables aléatoires réelles

De nouveau, nous nous plaçons dans le cas bivarié dans cette section. Le coefficient de corrélation linéaire (Bravais (1844)) entre deux v.a.r. de carré intégrable V_1 et V_2, aussi appelé coefficient de corrélation de Pearson, est défini par

$$\begin{aligned}\mathrm{Corr}(V_1,V_2) &= \frac{\mathrm{Cov}(V_1,V_2)}{\sqrt{\mathrm{Var}(V_1)\,\mathrm{Var}(V_2)}} \\ &= \frac{E([V_1-E(V_1)][V_2-E(V_2)])}{\sqrt{E([V_1-E(V_1)]^2)E([V_2-E(V_2)]^2)}}.\end{aligned}$$

La covariance $\mathrm{Cov}(V_1,V_2)$ entre deux v.a.r. V_1 et V_2 a une interprétation géométrique puisqu'elle correspond à un produit scalaire. On déduit de l'inégalité de Cauchy-Schwarz que $\mathrm{Corr}(V_1,V_2) \in [-1,1]$. De plus, la corrélation linéaire entre V_1 et V_2 vaut 1 (resp.

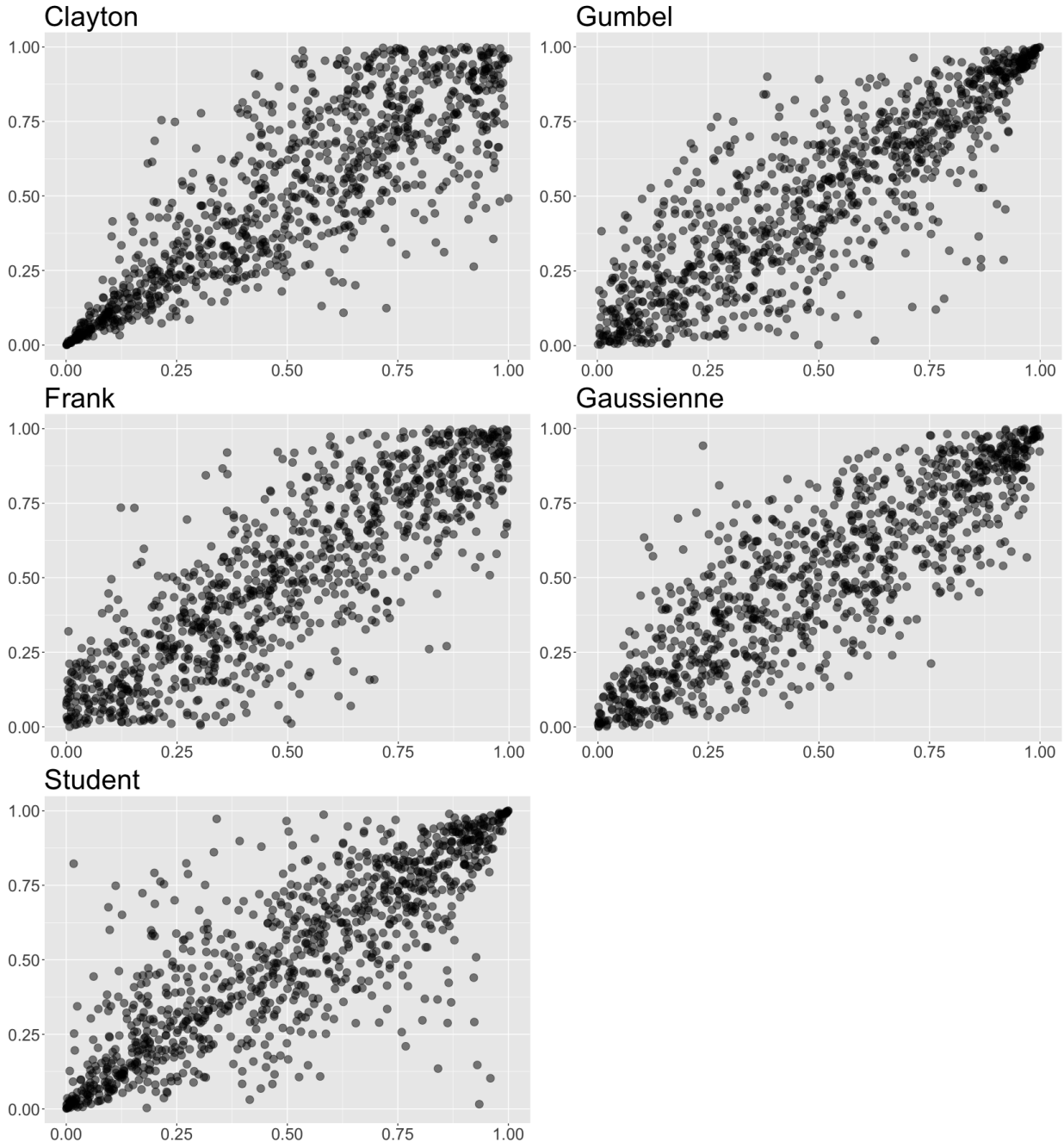

Fig. 1.3: Nuages de points correspondant aux cinq familles de copules présentées à la Section 1.3.2. Le tau de Kendall est fixé à 0.65. Pour la copule de Student, le paramètre donnant le nombre de degrés de liberté est fixé à 3.

-1) si et seulement si il existe $a > 0$ (resp. $a < 0$) et $b \in \mathbb{R}$ tels que $V_2 = aV_1 + b$. La corrélation linéaire intervient aussi dans le problème de régression linéaire. Si l'on considère la régression de V_2 à partir de V_1, les coefficients de régression $\hat{a}$ et $\hat{b}$ qui minimisent $\mathrm{E}[V_2 - (aV_1 + b)]^2$ sont

$$\hat{a} = \mathrm{Corr}(V_1, V_2)\sqrt{\frac{\mathrm{Var}(V_2)}{\mathrm{Var}(V_1)}}, \text{ et } \hat{b} = \mathrm{E}(V_2) - \hat{a}\,\mathrm{E}(V_1).$$

Enfin, si (V_1, V_2) possède des marginales gaussiennes et a pour copule la copule gaussienne de paramètre θ, alors $\mathrm{Corr}(V_1, V_2) = \theta$. En pratique, soit $(V_{1i}, V_{2i})_{i=1,..,n}$ un échantillon de taille n de réalisations du couple (V_1, V_2). Alors en notant $\overline{V_k} = n^{-1}\sum_{i=1}^n V_{ki}$ la moyenne empirique de V_k ($k = 1, 2$), on dispose des estimateurs classiques de la covariance

$$\hat{\sigma}_{V_1,V_2} = \frac{1}{n-1}\sum_{i=1}^{n}(V_{1i} - \overline{V}_1)(V_{2i} - \overline{V}_2),$$

et de la variance

$$\hat{\sigma}^2_{V_k} = \frac{1}{n-1}\sum_{i=1}^{n}(V_{ki} - \overline{V}_k)^2, \ k = 1, 2.$$

On peut donc estimer $\mathrm{Corr}(V_1, V_2)$ en utilisant ces estimateurs.

Le rho de Spearman (1904) entre les v.a.r. continues V_1 et V_2 est défini comme le coefficient de corrélation linéaire entre $U_1 = F_{V_1}(V_1)$ et $U_2 = F_{V_2}(V_2)$. Il est égal à

$$\rho(V_1, V_2) = \mathrm{Corr}(U_1, U_2) = \frac{E(U_1U_2) - 1/4}{1/12} = 12E(U_1U_2) - 3.$$

En effet, les variables U_1 et U_2 étant uniformément réparties sur $[0, 1]$, $\mathrm{E}(U_1) = \mathrm{E}(U_2) = 1/2$ et $\mathrm{Var}(U_1) = \mathrm{Var}(U_2) = 1/12$. On peut donc exprimer $\rho(V_1, V_2)$ en fonction de $\mathfrak{C}$ la copule du couple (V_1, V_2) grâce à la formule

$$\rho(V_1, V_2) = 12\int_0^1\int_0^1 \mathfrak{C}(u_1, u_2)du_1du_2 - 3.$$

Là où la corrélation linéaire mesure la relation linéaire entre deux variables, le rho de Spearman mesure la relation de monotonie (linéaire ou non linéaire) entre les variables, en se basant sur les rangs de chaque variable. Le rho de Spearman vaut 1 (resp. -1) si et seulement si V_1 et V_2 sont comonotones (resp. anticomonotones).

Le tau de Kendall entre les v.a.r. continues V_1 et V_2 est défini par

$$\begin{aligned}\tau(V_1,V_2) &= \mathrm{E}[\mathrm{sign}((\tilde{V}_1-\breve{V}_1)(\tilde{V}_2-\breve{V}_2))] \qquad (1.12)\\ &= 2\,\mathrm{P}((\tilde{V}_1-\breve{V}_1)(\tilde{V}_2-\breve{V}_2))>0)-1\\ &= 4\,\mathrm{P}(\tilde{V}_1<\breve{V}_1,\tilde{V}_2<\breve{V}_2)-1\\ &= 4\,\mathrm{P}(F_{V_1}(\tilde{V}_1)<F_{V_1}(\breve{V}_1),F_{V_2}(\tilde{V}_2)<F_{V_2}(\breve{V}_2))-1 \qquad (1.13)\end{aligned}$$

avec $(\tilde{V}_1,\tilde{V}_2)$ et $(\breve{V}_1,\breve{V}_2)$ deux couples de même loi que (V_1,V_2) et tels que $(\tilde{V}_1,\tilde{V}_2)$ est indépendant de $(\breve{V}_1,\breve{V}_2)$. De manière empirique, on définit

$$\begin{aligned}n_c &= \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{1}_{V_{1i}<V_{1j},V_{2i}<V_{2j}},\\ n_d &= \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{1}_{V_{1i}<V_{1j},V_{2i}>V_{2j}},\end{aligned}$$

le nombre de paires concordantes (n_c) et discordantes (n_d) dans l'échantillon. Le tau de Kendall, tel que défini dans Kendall (1938) est alors estimé par

$$\hat{\tau}=\frac{n_c-n_d}{n(n-1)/2}=\frac{n_c-n_d}{n_c+n_d},$$

conformément à (1.12). D'après (1.13), on peut voir que le tau de Kendall s'exprime lui aussi en fonction de la copule $\mathfrak{C}$:

$$\tau(V_1,V_2)=4\int_0^1\int_0^1\mathfrak{C}(u_1,u_2)d\mathfrak{C}(u_1,u_2)-1.$$

Si le couple (V_1,V_2) est de copule archimédienne donnée par le générateur ϕ, alors Genest & MacKay (1986a,b) ont établi que

$$\tau(V_1,V_2)=1+4\int_0^1\frac{\phi(t)}{\phi'(t)}dt. \qquad (1.14)$$

Cette propriété permet de montrer que pour les familles de copules archimédiennes présentées plus haut dans cette section, il y a une correspondance bijective entre le paramètre de copule θ et tau de Kendall. Par exemple si (V_1,V_2) est de copule de Clayton $\mathfrak{C}_\theta$, on a $\tau(V_1,V_2)=\theta/(\theta+2)$. Pour la copule gaussienne, il n'y a pas de formule telle que (1.14), néanmoins il y a également une bijection entre θ et τ. Cette bijection entre tau de Kendall

et paramètre de copule est utilisée au Chapitre 3.

Les trois mesures de concordance présentées ci-dessus sont non paramétriques, et donc très utilisées en pratique car elles ne nécessitent aucune hypothèse sur le couple (V_1, V_2).

1.3.3 L'estimation de copule

Dans cette section, on considère $(V_{1i}, \ldots, V_{pi})_{i=1,..,n}$ un échantillon de réalisations i.i.d. du vecteur $V = (V_1, \ldots, V_p)$.

L'estimation non paramétrique

Étant donné un vecteur aléatoire $(V_1, \ldots, V_p)$, nous avons vu avec l'équation (1.9) que la copule du vecteur s'exprime

$$\mathfrak{C}(u_1, \ldots, u_p) = F(F_{V_1}^{-1}(u_1), \ldots, F_{V_p}^{-1}(u_p)).$$

Une méthode non paramétrique d'estimation de copules, étudiée par Deheuvels (1979), consiste à remplacer dans cette formule les fonctions de répartition F et F_{V_k} $(k = 1, \ldots, p)$ par leurs estimateurs non paramétriques

$$\hat{F}^{(1)}(v_1, \ldots, v_p) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{V_{1i}\leq v_1,\ldots,V_{pi}\leq v_p}, \tag{1.15}$$

$$\hat{F}_{V_k}^{(1)}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{V_{ki}\leq t},\ k = 1, \ldots, p. \tag{1.16}$$

On définit alors la fonction de dépendance empirique

$$\begin{aligned}\hat{\mathfrak{C}}^{(1)}(u_1, \ldots, u_p) &= \hat{F}^{(1)}([\hat{F}_{V_1}^{(1)}]^{-1}(u_1), \ldots, [\hat{F}_{V_p}^{(1)}]^{-1}(u_p)) \qquad (1.17)\\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{V_{1i}\leq[\hat{F}_{V_1}^{(1)}]^{-1}(u_1),\ldots,V_{pi}\leq[\hat{F}_{V_p}^{(1)}]^{-1}(u_p)}\\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\hat{F}_{V_1}^{(1)}(V_{1i})\leq u_1,\ldots,\hat{F}_{V_p}^{(1)}(V_{pi})\leq u_p}. \qquad (1.18)\end{aligned}$$

L'opérateur inverse $^{-1}$ désigne ici l'inverse généralisé, c'est à dire que pour une fonction réelle croissante G,

$$G^{-1}(t) = \inf\{s \in \mathbb{R} : G(s) \geq t\}.$$

La quantité $[\hat{F}_{V_k}^{(1)}]^{-1}(u_k)$ désigne alors le quantile d'ordre u_k de de l'échantillon $(V_{ki})_{i=1,\dots,n}$. L'équation (1.18) montre que la fonction de dépendance empirique correspond à l'estimateur non paramétrique de la fonction de répartition associée à l'échantillon $(\hat{F}_{V_1}^{(1)}(V_{1i}),\dots,\hat{F}_{V_p}^{(1)}(V_{pi}))_{i=1,\dots,n}$. Les quantités $(\hat{F}_{V_1}^{(1)}(V_{1i}),\dots,\hat{F}_{V_p}^{(1)}(V_{pi}))_{i=1,\dots,n}$ sont appelées pseudo-observations. Plus généralement, si $(\hat{F}_{V_k})_{k=1,\dots,p}$ désignent des estimateurs de $(F_{V_k})_{k=1,\dots,p}$, alors les quantités $\hat{U}_{ki}=\hat{F}_{V_k}(V_{ki})$ $(k=1,\dots,p,\ i=1,\dots,n)$ sont appelées pseudo-observations associées à l'échantillon $(V_{ki})_{k,i}$.

L'estimateur de la copule donné par la fonction de dépendance de Deheuvels (1979) a été étudiée plus récemment par Fermanian et al. (2004) et Doukhan et al. (2005). Il est robuste, et universel, car il ne suppose aucune hypothèse sur le vecteur $(V_1,\dots,V_p)$. Cependant il n'est pas lisse et ne permet donc pas d'estimer la densité de copule.

Une méthode non paramétrique qui permet d'aboutir à un estimateur de copule lisse consiste à utiliser des estimateurs à noyau pour estimer F et F_{V_k}. Soit $K:\mathbb{R}\to\mathbb{R}$ (resp. $K_p:\mathbb{R}^p\to\mathbb{R}$) une fonction lisse, positive, et symétrique, telle que $K(t)=0$ si $|t|\geq 1$ et $\int K(t)dt=1$ (resp. $K_p(v)=0$ si $\|v\|\geq 1$ et $\int K_p(v)dv=1$, avec $v=(v_1,\dots,v_p)$). On définit les primitives des noyaux

$$\mathcal{K}(t)=\int_{-\infty}^{t}K(s)ds,$$

et

$$\mathcal{K}_p(v)=\int_{-\infty}^{v_1}\dots\int_{-\infty}^{v_p}K_p(r)dr.$$

Alors en posant

$$\begin{aligned}\hat{F}^{(2)}(v_1,\dots,v_p) &= \frac{1}{n}\sum_{i=1}^{n}\mathcal{K}_p\left(\frac{v_1-V_{1i}}{h},\dots,\frac{v_p-V_{pi}}{h}\right),\\ \hat{F}_{V_k}^{(2)}(t) &= \frac{1}{n}\sum_{i=1}^{n}\mathcal{K}\left(\frac{t-V_{ki}}{h_k}\right),\ k=1,\dots,p,\end{aligned}$$

pour des paramètres de "fenêtre" h et $(h_k)_{k=1,\dots,p}$ fixés, on définit

$$\hat{\mathfrak{C}}^{(2)}(u_1,\dots,u_p)=\hat{F}^{(2)}([\hat{F}_{V_1}^{(2)}]^{-1}(u_1),\dots,[\hat{F}_{V_p}^{(2)}]^{-1}(u_p)),$$

qui est un estimateur lisse de $\mathfrak{C}$. On peut déduire une densité de copule en dérivant $\hat{\mathfrak{C}}^{(2)}$. La méthode des estimateurs à noyau est donc utile, toutefois elle se heurte au fléau de la dimension. On sait en effet que la vitesse de convergence des estimateurs à noyau diminue

très vite avec la dimension (e.g. voir Stone (1980)). En pratique, cette méthode n'est donc pas utilisée si la dimension est élevée. Les propriétés théoriques de l'estimateur $\hat{\mathfrak{C}}^{(2)}$ sont étudiées dans Fermanian & Scaillet (2003).

L'estimation paramétrique

Considérons une famille paramétrique de copules $\mathcal{C} = \{\mathfrak{C}_\theta : \theta \in \Theta\}$, avec Θ un sous-ensemble compact de $\mathbb{R}^k$, et supposons que la copule $\mathfrak{C}$ du vecteur V appartient à $\mathcal{C}$. De même, supposons que chaque loi marginale F_{V_k} appartient à une famille paramétrique $\mathcal{F}_k = \{F_{k,\alpha_k}, \alpha_k \in \Delta_k\}$.

Les méthodes d'estimation paramétriques consistent à proposer des estimateurs pour les paramètres $(\theta, \alpha_1, \ldots, \alpha_p)$. Parmi elles, la méthode du maximum de vraisemblance est la plus utilisée. En dérivant l'équation (1.8), on voit que la densité du vecteur V s'exprime

$$f(v_1, \ldots, v_p) = c(F_{V_1}(v_1), \ldots, F_{V_p}(v_p)) \prod_{k=1}^{p} f_{V_k}(v_k),$$

où pour tout k, f_{V_k} est la densité de la V_k. Ainsi, la log-vraisemblance des données $(V_{1i}, \ldots, V_{pi})_{i=1,\ldots,n}$ s'écrit

$$\log \mathcal{L}_n(\theta, \alpha_1, \ldots, \alpha_p) = \sum_{i=1}^{n} \left(\log c_\theta(F_{1,\alpha_1}(V_{1i}), \ldots, F_{p,\alpha_p}(V_{pi})) + \sum_{k=1}^{p} \log f_{k,\alpha_k}(V_{ki}) \right).$$

La méthode du maximum de vraisemblance consiste à poser

$$(\hat{\theta}, \hat{\alpha_1}, \ldots, \hat{\alpha_p}) = \underset{\theta,\alpha_1,\ldots,\alpha_p}{\arg\max} \ \log \mathcal{L}_n(\theta, \alpha_1, \ldots, \alpha_p).$$

Les propriétés de convergence des estimateurs de maximum de vraisemblance sont étudiées dans A. Van der Vaart (1998), et les principaux résultats s'appliquent en général au cas de figure ci-dessus. En particulier, sous réserve d'hypothèses de régularité par rapport aux paramètres du modèle, les estimateurs du paramètre d'association ainsi que des paramètres décrivant le comportement des marges convergent à la vitesse $n^{1/2}$.

Du point de vue numérique, l'optimisation simultanée des paramètres θ et $(\alpha_k)_{k=1,\ldots,p}$ peut s'avérer difficile, ainsi il est courant de procéder à une optimisation en deux étapes comme cela est étudié par Shih & Louis (1995). Cette méthode consiste à estimer dans un premier temps chaque paramètre $\hat{\alpha}_k$ $(k = 1, \ldots, p)$ séparément par maximum de

vraisemblance (i.e. $\hat{\alpha}_k = \arg\max_{\alpha_k} \sum_{i=1}^{n} \log f_{k,\alpha_k}(V_{ki})$), puis à estimer θ par

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log c_\theta(F_{1,\hat{\alpha}_1}(V_{1i}), \ldots, F_{p,\hat{\alpha}_p}(V_{pi})). \tag{1.19}$$

Quelque soit la méthode d'optimisation utilisée, l'utilisation de modèles paramétriques comporte un risque de biais de modèle, c'est à dire le risque que l'une des marginales F_k n'appartienne pas à la famille $\mathcal{F}_k$ ou que $\mathfrak{C}$ n'appartienne pas à $\mathcal{C}$. Dans un tel cas de figure l'estimation de la copule est fortement détériorée. Par exemple, Fermanian & Scaillet (2005) ont montré grâce à des simulations numériques que l'hypothèse de loi marginale gaussienne pour un vecteur bivarié possédant des lois de Student comme marginales impacte fortement l'estimation du paramètre de la copule de Frank.

A cause des nombreux risques de biais de modèle causés par l'estimation des lois marginales, l'approche paramétrique que nous avons décrite dans cette section est assez peu utilisée. Les praticiens préfèrent souvent utiliser une méthode semi-paramétrique, que nous allons décrire dans la section suivante.

L'estimation semi-paramétrique

La méthode semi-paramétrique la plus courante consiste à conserver seulement l'hypothèse $\mathfrak{C} \in \mathcal{C} = \{\mathfrak{C}_\theta : \theta \in \Theta\}$, et à estimer θ avec l'estimateur

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log c_\theta(\hat{F}^{(1)}_{V_1}(V_{1i}), \ldots, \hat{F}^{(1)}_{V_p}(V_{pi})), \tag{1.20}$$

c'est à dire à remplacer dans l'équation (1.19) les estimateurs paramétriques $F_{k,\hat{\alpha}_k}$ ($k = 1, \ldots, p$) par les estimateurs non paramétrique $\hat{F}^{(1)}_{V_k}$. Rappelons que les quantités $\hat{U}^{(1)}_{ki} = \hat{F}^{(1)}_{V_k}(V_{ki})$ ($k = 1, \ldots, p,\ i = 1, \ldots, n$), sont alors appelées pseudo-observations. En cohérence avec cette appellation, la fonction à optimiser

$$\sum_{i=1}^{n} \log c_\theta(\hat{F}^{(1)}_{V_1}(V_{1i}), \ldots, \hat{F}^{(1)}_{V_p}(V_{pi})), \tag{1.21}$$

est souvent appelée pseudo-vraisemblance du modèle. Les propriétés asymptotiques de l'estimateur (1.20) ont été étudiées dans Genest et al. (1995a) et Shih & Louis (1995). Comme illustré dans Fermanian & Scaillet (2005), ainsi que dans Silvapulle et al. (2004), l'estimateur semi-paramétrique de copule (1.20) aboutit en pratique à de faibles erreurs

d'estimation, tout en réduisant drastiquement les risques de biais de modèle. Des remarques intéressantes sur les effets inattendus de réduction de variance liés à l'utilisation des pseudo-observations pour l'inférence de copule sont faites dans Charpentier et al. (2007).

Signalons que dans l'estimateur (1.20) et la pseudo-vraisemblance (1.21), il est tout à fait possible de remplacer les estimateurs empiriques $\hat{F}_{V_k}^{(1)}$ par les estimateurs à noyau $\hat{F}_{V_k}^{(2)}$, ou même d'autres estimateurs non paramétriques des fonctions de répartition marginales. On aboutit alors à un nouvel estimateur semi-paramétrique de θ. Les estimateurs de θ de ce type sont appelés *omnibus estimator* dans la littérature. Plus généralement, les sections précédentes montrent que le problème d'estimation de copules se découple en un problème d'estimation des lois marginales, et un problème d'estimation de la loi jointe. Il est alors possible d'utiliser, pour chacune des deux étapes, les différentes approches paramétriques ou non paramétriques vu précédemment.

Dans le cas bivarié, une autre méthode semi-paramétrique couramment employée repose sur l'estimation du tau de Kendall. Comme on l'a mentionné à la Section 1.3.2, le tau de Kendall et le paramètre des copules archimédiennes, mais aussi de la copule gaussienne, sont en bijection. Ainsi à partir d'un estimateur du tau de Kendall $\hat{\tau}$, on peut déduire un estimateur du paramètre de copule $\hat{\theta}$. La même technique peut être employée avec le rho de Spearman lorsqu'il est en bijection avec le paramètre de copule. Par exemple si (V_1, V_2) a pour copule la copule de Frank $\mathfrak{C}_\theta$ alors $\rho(V_1, V_2) = 1 - 12\theta^{-1}(D_1(\theta) - D_2(\theta))$, avec

$$D_k(x) = \frac{k}{x^k} \int_0^x \frac{t^k}{e^t - 1} dt.$$

Cette méthode d'estimation peut être envisagée comme l'analogue de la méthode des moments dans le contexte de l'estimation de copules. En effet, comme dans le cas de la méthode des moments, il s'agit d'ajuster la distribution sur un certain nombre de grandeurs (ici des mesures de dépendance) qui ont été préalablement estimées non paramétriquement. Cette technique est étudiée dans un cadre multidimensionnel dans Genest et al. (2011). Notons que dans le cas où l'on estime le paramètre θ de la copule gaussienne et où les lois marginales sont également gaussiennes, il est préférable d'utiliser la corrélation linéaire comme estimateur de θ plutôt que la méthode des moments ci-dessus.

Au Chapitre 3, nous utilisons un estimateur de copule semi-paramétrique tel que 1.20) pour estimer la dépendance entre deux durées successives. Cette estimation est faite conditionnellement à des variables explicatives. Nous présentons donc le concept de copule conditionnelle dans la section suivante.

1.3.4 Les copules conditionnelles

Motivations

Avec l'essor du Big Data et des pratiques visant à mettre l'analyse de données toujours plus au cœur des processus de décision, les besoins en méthodes statistiques spécifiques à chaque problème progressent dans tous les domaines. Dans notre cas, les modèles statistiques multivariés qui étudient la dépendance nécessitent, de plus en plus, de prendre en compte des variables explicatives, notamment dans le domaine de l'économétrie. Un exemple donné dans Gijbels et al. (2011) illustre très bien l'intérêt de faire intervenir des facteurs explicatifs dans l'étude de la dépendance entre deux variables V_1 et V_2. Notons $(V_{1i}, V_{2i})_{i=1,\dots,n}$ un échantillon où V_{1i} (resp. V_{2i}) correspond à l'espérance de vie des hommes (resp. femmes) dans le pays i (on supposera que l'on a numéroté un ensemble de pays entre 1 et n), alors on peut s'intéresser à la dépendance entre l'espérance de vie des hommes et des femmes. Une question naturelle est de se demander si cette dépendance est différente dans les pays riches et dans les pays pauvres. Si l'on dispose d'observations X_i qui correspondent aux PIB par habitant dans les pays i, alors l'utilisation de copules conditionnelles permet d'obtenir des éléments de réponse (voir Gijbels et al. (2011)).

Le concept de copule conditionnelle

L'étude de la dépendance conditionnelle, qui est un sujet auquel on s'intéresse au Chapitre 3, consiste à modéliser la loi du vecteur $V = (V_1, \dots, V_p)$ en prenant en compte l'influence d'un vecteur de variables explicatives $X \in \mathcal{X} \subset \mathbb{R}^d$. C'est à dire que la dépendance entre les composantes du vecteur V est susceptible de varier en fonction de la valeur de X. Dans ce cas de figure, le Théorème de Sklar (1.8) s'applique pour tout $X = x$, et

$$F(v_1, \dots, v_p|x) = \mathfrak{C}^{(x)}(F_{V_1}(v_1|x), \dots, F_{V_p}(v_p|x)), \tag{1.22}$$

avec $F(v_1, \dots, v_p|x) = P(V_1 \leq v_1, \dots, V_p \leq v_p|X = x)$ la fonction de répartition jointe conditionnelle, $F_{V_k}(v_k|x) = P(V_k \leq v_k|X = x)$ $(k = 1, \dots, p)$ les fonctions de répartition marginales conditionnelles, et $\mathfrak{C}^{(x)}$ la copule de $(V_1, \dots, V_p)$ sachant $X = x$. L'étude des copules conditionnelles est un sujet récent en statistique. Les premiers travaux sont dus à Patton (2006a,b), dans le domaine de l'économétrie.

L'estimation de copules dans le cas conditionnel

Dans cette section, nous noterons $X \in \mathcal{X} \subset \mathbb{R}^d$ un vecteur de variables explicatives et $(V_{1i}, \ldots, V_{pi}, X_i)_{i=1,\ldots,n}$ un échantillon d'observations. Les différentes méthodes vues dans la Section 1.3.3 se généralisent en général au cas conditionnel.

Par analogie avec les équations (1.15) et (1.16), on définit les estimateurs

$$\begin{aligned} \hat{F}^{(1)}(v_1, \ldots, v_p|x) &= \frac{1}{n}\sum_{i=1}^{n} w_{i,n}(x)\mathbb{1}_{V_{1i}\leq v_1,\ldots,V_{pi}\leq v_p}, \\ \hat{F}^{(1)}_{V_k}(t|x) &= \frac{1}{n}\sum_{i=1}^{n} w_{i,n}(x)\mathbb{1}_{V_{ki}\leq t},\ k = 1, \ldots, p, \end{aligned}$$

de la fonction de répartition jointe conditionnelle et des fonctions de répartition marginales conditionnelles. Dans ces expressions, les termes $w_{i,n}(x)$ désignent des poids attribués à chaque observation qui permettent de localiser les estimateurs en chaque point de l'espace $\mathcal{X}$. Par exemple, il est possible d'utiliser les poids Nadaraya-Watson (voir Nadaraya (1964) ou Watson (1964)) définis par

$$w_{i,n}(x) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)}, \tag{1.23}$$

avec $K : \mathbb{R}^d \to \mathbb{R}$ une fonction noyau (i.e. une fonction positive, symétrique, telle que $K(x) = 0$ si $\|x\| \geq 1$, et $\int K(x)dx = 1$). D'autres types de poids sont suggérés dans Gijbels et al. (2011). En utilisant le Théorème de Sklar conditionnel (1.22), on peut estimer, comme dans la formule (1.17), la copule conditionnelle par

$$\begin{aligned} \hat{F}^{(1)}([\hat{F}^{(1)}_{V_1}]^{-1}(u_1|x), \ldots, [\hat{F}^{(1)}_{V_p}]^{-1}(u_p|x)|x) &= \frac{1}{n}\sum_{i=1}^{n} w_{i,n}(x)\mathbb{1}_{V_{1i}\leq[\hat{F}^{(1)}_{V_1}]^{-1}(u_1|x),\ldots,V_{pi}\leq[\hat{F}^{(1)}_{V_p}]^{-1}(u_p|x)} \\ &= \frac{1}{n}\sum_{i=1}^{n} w_{i,n}(x)\mathbb{1}_{\hat{F}^{(1)}_{V_1}(V_{1i}|x)\leq u_1,\ldots,\hat{F}^{(1)}_{V_p}(V_{pi}|x)\leq u_p} \end{aligned} \tag{1.24}$$

En pratique, il est plus simple de remplacer les termes $\hat{F}^{(1)}_{V_k}(V_{ki}|x)$ par $\hat{F}^{(1)}_{V_k}(V_{ki}|X_i)$. En effet, cela permet de définir des pseudo-observations $\hat{U}_{kix} = \hat{F}^{(1)}_{V_k}(V_{ki}|X_i)$ $(k = 1, \ldots, p,$ $i = 1, \ldots, n)$ qui ne dépendent pas de x, et qu'il n'est donc pas nécessaire de recalculer pour chaque valeur de x. De plus, le poids $w_{i,n}(x)$ dans la somme (1.24) est non nul lorsque X_i et x sont proches. Cela implique, en supposant que les marginales conditionnelles soient suffisamment régulières par rapport à x (hypothèse qui est vérifiée en générale), que les

quantités $\hat{F}^{(1)}_{V_k}(V_{ki}|x)$ et $\hat{F}^{(1)}_{V_k}(V_{ki}|X_i)$ sont voisines. Ainsi, un estimateur non paramétrique de la copule conditionnelle est donné par

$$\hat{\mathfrak{C}}^{(x,1)}(u_1,\dots,u_p) = \frac{1}{n}\sum_{i=1}^{n} w_{i,n}(x)\mathbb{1}_{\hat{F}^{(1)}_{V_1}(V_{1i}|X_i)\leq u_1,\dots,\hat{F}^{(1)}_{V_p}(V_{pi}|X_i)\leq u_p}.$$

Les propriétés de cet estimateur non paramétrique sont étudiées dans Veraverbeke et al. (2011).

Comme dans le cas non conditionnel, on peut considérer une famille de copules $\mathcal{C} = \{\mathfrak{C}_\theta, \theta \in \Theta\}$ et faire l'hypothèse que pour tout x dans $\mathcal{X}$, il existe $\theta(x) \in \Theta$ tel que $\mathfrak{C}^{(x)} = \mathfrak{C}_{\theta(x)}$. Alors, la log-vraisemblance des observations s'écrit

$$\log \mathcal{L}_n = \sum_{i=1}^{n}\left(\log c_{\theta(X_i)}(F_{V_1}(V_{1i}|X_i),\dots,F_{V_p}(V_{pi}|X_i)) + \sum_{k=1}^{p}\log f_{V_k}(V_{ki}|X_i) + \log f_X(X_i)\right) \tag{1.25}$$

en notant f_X la densité de X.

Supposons désormais que l'on dispose d'estimateurs $\hat{F}_{V_1}(\cdot|x),\dots,\hat{F}_{V_p}(\cdot|x)$ des lois marginales conditionnelles. L'approche paramétrique consiste à supposer que la fonction $\theta(x)$ dépend uniquement d'un paramètre fini-dimensionnel $\beta \in \Pi \subset \mathbb{R}^q$, i.e. $\theta(x) = \psi(x,\beta)$ avec ψ une fonction connue. Le paramètre β peut alors être estimé avec une méthode de maximum de vraisemblance, par

$$\hat{\beta} = \arg\max_{\beta\in\Pi}\sum_{i=1}^{n}\log c_{\psi(X_i,\beta)}(\hat{F}_{V_1}(V_{1i}|X_i),\dots,\hat{F}_{V_p}(V_{pi}|X_i)).$$

Dans le cas où les estimateurs $\hat{F}_{V_1}(\cdot|x),\dots,\hat{F}_{V_p}(\cdot|x)$ sont paramétriques, éventuellement conditionnels, c'est à dire

$$\hat{F}_{V_k}(\cdot|x) = F_{k,\psi_k(x,\hat{\alpha}_k)} \in \mathcal{F}_k \ (k=1,\dots,p),$$

avec ψ_k une fonction connue et $\mathcal{F}_k = \{F_{k,\alpha_k}, \alpha_k \in \Delta_k\}$ une famille paramétrique, alors il est bien sûr possible d'utiliser une optimisation en une étape de la vraisemblance complète (1.25), même si cela peut s'avérer complexe sur le plan numérique. On trouve des applications de cette technique paramétrique dans le domaine de l'économétrie, par exemple dans Jondeau & Rockinger (2006) ou Bartram et al. (2007).

Une méthode non paramétrique consiste à ne faire aucune hypothèse sur la forme de

la fonction $\theta(x)$. Il est alors possible d'estimer la log-vraisemblance conditionnelle

$$\mathrm{E}\left[\log c_{\theta(x)}(F_{V_1}(V_1|X),\ldots,F_{V_p}(V_p|X))|X=x\right],$$

par un estimateur à noyau faisant intervenir les poids $w_{i,n}(x)$ de l'équation (1.23). Ainsi, on estime $x \to \theta(x)$ par

$$\hat{\theta}(x) = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} w_{i,n}(x) \log c_\theta(\hat{F}_{V_1}(V_{1i}|X_i),\ldots,\hat{F}_{V_p}(V_{pi}|X_i)). \tag{1.26}$$

C'est cette approche non paramétrique que nous utilisons au Chapitre 3, pour étudier un cas de dépendance conditionnelle. De façon plus générale, cet estimateur à noyau peut être remplacé par un estimateur par polynômes locaux, comme cela est étudié par Abegaz et al. (2012).

Nous avons déjà vu dans la Section 1.3.3, que les estimateurs à noyau ne sont pas performants si la dimension de X est grande. Ainsi, Fermanian & Lopez (2018) ont proposé une méthode intermédiaire, entre le cas paramétrique et le cas non paramétrique, qui permet d'utiliser des estimateurs à noyau en grande dimension. Pour cette méthode, dite single-index, l'hypothèse faite est que $\theta(x) = \psi(\beta,\ {}^t\beta \cdot x)$, avec une fonction ψ inconnue, à estimer de manière non paramétrique, β un paramètre fini-dimensionnel.

Une question, qui émerge naturellement en pratique, est de déterminer s'il est nécessaire, lorsque l'on dispose de variables explicatives X, d'utiliser un modèle de copule conditionnel plutôt qu'un modèle non conditionnel. Des tests ont été proposés par Derumigny & Fermanian (2017) et Gijbels et al. (2017) comme outils d'aide à la décision dans ce contexte.

1.3.5 L'utilisation de modèles multivariés en assurance

Les modèles multivariés sont importants en assurance car ils permettent de modéliser plus fidèlement les risques qui, souvent, présentent des aspects multidimensionnels. Dans cette section, chaque composante du vecteur $V = (V_1,\ldots,V_p)$ peut donc être interprétée comme un risque, c'est à dire une variable aléatoire réelle à valeurs positives. Étant donné un risque R, une mesure de risque $\mathfrak{M}$ est définie comme une fonction réelle du risque $\mathfrak{M}(R)$ telle que $\mathfrak{M}(R) \in [0,+\infty]$. Par exemple, l'espérance d'un risque $\mathrm{E}(R)$ est une mesure de risque. Deux autres mesures de risque couramment utilisées en assurance sont la value-at-risk (VaR) et la tail value-at-risk (TVaR). La VaR de niveau (ou de probabilité) $\alpha \in]0,1[$

est définie comme le quantile de niveau α de la variable aléatoire R :

$$\mathrm{VaR}_\alpha(R) = F_R^{-1}(\alpha),$$

avec F_R la fonction de répartition du risque R. La TVaR de niveau $\alpha \in]0,1[$ est, quant à elle, définie par

$$\begin{aligned} \mathrm{TVaR}_\alpha(R) &= \mathrm{E}[R|R > \mathrm{VaR}_\alpha(R)] \\ &= \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_\xi(R)d\xi, \end{aligned}$$

c'est à dire comme la valeur moyenne prise par R sachant $R > \mathrm{VaR}_\alpha(R)$. Pour illustrer l'importance des modèles multivariés en assurance, nous traitons dans cette section les sujets de l'agrégation de risques et de l'étude des risques extrêmes, avant de donner d'autres exemples d'utilisations de modèles multivariés tirés de l'assurance. Le livre Denuit et al. (2006) est une introduction détaillée aux notions abordées ici.

L'agrégation de risques

Un agrégat de p risques $V_1, \ldots, V_p$ est une fonction $g(V_1, \ldots, V_p) \in \mathbb{R}^+$. L'agrégat le plus commun en assurance est la somme des risques, que nous noterons $S = V_1 + \ldots + V_p$. On sait que l'espérance est additive, i.e.

$$\mathrm{E}(S) = \mathrm{E}(V_1) + \ldots + \mathrm{E}(V_p). \tag{1.27}$$

Par ailleurs, la fonction de répartition de S est donnée par

$$\begin{aligned} F_S(t) &= \mathrm{P}(V_1 + \ldots + V_p \leq t) \\ &= \int\limits_{v_1+\ldots+v_p\leq t} \ldots \int f(v_1, \ldots, v_p)dv_1 \ldots dv_p \\ &= \int\limits_{v_1+\ldots+v_p\leq t} \ldots \int c(F_{V_1}(v_1), \ldots, F_{V_p}(v_p)f_{V_1}(v_1) \ldots f_{V_p}(v_p)dv_1 \ldots dv_p, \end{aligned}$$

en reprenant les notations des sections précédentes. On voit donc que la fonction de répartition de S dépend de la densité de copule c du vecteur V. Ainsi, si l'égalité (1.27) est toujours vérifiée, il n'est pas vrai en général que $\mathrm{VaR}_\alpha(S)$ est égal à $\mathrm{VaR}_\alpha(V_1) + \ldots + \mathrm{VaR}_\alpha(V_p)$. Cela dépend de la copule du vecteur V. En fait, cela est vrai si les risques

$V_1, \ldots, V_p$ sont comonotones car dans ce cas on peut montrer que l'on a

$$(V_1, \ldots, V_p) = (F_{V_1}^{-1}(U), \ldots, F_{V_p}^{-1}(U)),$$

avec U une variable de loi uniforme sur $[0, 1]$. Alors, pour $t = F_{V_1}^{-1}(\alpha) + \ldots + F_{V_p}^{-1}(\alpha)$,

$$\begin{aligned} F_S(t) &= \mathrm{P}(F_{V_1}^{-1}(U) + \ldots + F_{V_p}^{-1}(U) \leq t) \\ &= \mathrm{P}((F_{V_1}^{-1} + \ldots + F_{V_p}^{-1})(U) \leq (F_{V_1}^{-1} + \ldots + F_{V_p}^{-1})(\alpha)) \\ &= \mathrm{P}(U \leq \alpha) = \alpha, \end{aligned}$$

d'où

$$\mathrm{VaR}_\alpha(S) = \mathrm{VaR}_\alpha(V_1) + \ldots + \mathrm{VaR}_\alpha(V_p). \tag{1.28}$$

Mais l'égalité (1.28) n'est pas vraie en général. Cet exemple montre que pour estimer une mesure de risque d'un agrégat $\mathfrak{M}(g(V_1, \ldots, V_p))$, il est nécessaire de prendre en compte la dépendance entre les différents risques.

De nombreux articles s'intéressent à la question de l'agrégation de risques dans la littérature. Un sujet majeur est de calculer une mesure de risque, par exemple la VaR ou la TVaR, pour un agrégat donné. Ainsi, Gijbels & Herrmann (2014) étudient mathématiquement la distribution de la somme de variables aléatoires dont la dépendance est spécifiée par une copule, alors que Cossette et al. (2014) proposent une méthode numérique pour calculer des bornes pour la somme, le produit, et le ratio, de différents risques. La value-at-risk d'une somme est étudiée dans Cuberos et al. (2019) et Cossette, Marceau, Nguyen, & Robert (2018). Une méthode analytique, ainsi qu'une méthode de Monte-Carlo sont proposées dans Cossette, Marceau, Nguyen, & Robert (2018) pour étudier la probabilité de dépassement de seuil d'une somme de v.a.r. de copule archimédienne. Enfin, Bargès et al. (2009) estiment le niveau de risque de stratégies d'investissement en évaluant la TVaR d'un portefeuille composé de différents actifs.

En pratique, l'agrégation de risques est en effet un sujet très important en assurance. Par exemple, en assurance non-vie, chaque police d'assurance constitue un risque et on suppose pour calculer les tarifs ou les provisions d'assurance que les risques associés à chaque contrat sont indépendants. C'est le cas par exemple en assurance auto, ou en assurance habitation. Cette hypothèse est naturelle et on peut supposer qu'elle est proche de la réalité. De plus, elle permet d'utiliser pour l'évaluation du risque du portefeuille d'assurances les modèles statistiques de régression qui reposent sur l'hypothèse

d'indépendance entre les observations: le modèle linéaire généralisé (GLM), l'algorithme CART, etc. Voir par exemple Henckaerts et al. (2018) pour un problème de tarification, et Wüthrich (2018) pour du provisionnement non-vie. Dans ce cas de figure, la quantité d'intérêt est l'espérance $\mathrm{E}(S)$ de la sommes des risques assurés, qui définit la prime pure du portefeuille d'assurances. Si l'on ne suppose plus que les observations sont i.i.d., Cossette, Marceau, & Mtalai (2018) donnent des résultats théoriques sur la somme des risques assurés. L'assurance sur les revenus agricoles, qui comporte une composante de prix et une composante d'incertitude sur la récolte, est étudiée dans Goodwin & Hungerford (2014). C'est un exemple de risque non-vie où les observations ne sont pas indépendantes, car un même phénomène peut affecter les revenus de nombreux agriculteurs. Par exemple, des intempéries ou une chute des prix du marché. Dans Solvabilité II, le règlement européen qui encadre le calcul des fonds propres minimums devant être détenus par les assurances pour faire face à leurs risques, l'agrégation de risques a une place centrale car l'approche utilisée est une approche *bottom-up*, où les risques sont d'abord évalués par sous-catégories (par exemple en vie : risque de longévité, risque de mortalité, risque de pandémie, risque de rachat etc.), puis agrégés en catégories (dans le cas précédent, le risque vie ; le risque de marché est un autre exemple de catégorie), elles-mêmes agrégées pour évaluer le risque total d'une compagnie d'assurance. La mesure de risque retenue dans Solvabilité II est la VaR à $0,5\%$ de la valeur des fonds propres économiques projetés à un horizon d'un an, ce qui garantie avec une probabilité de 99,5% qu'une compagnie ne soit pas en faillite dans un an. Comme expliqué dans Devineau & Loisel (2009), la méthode d'agrégation utilisée dans Solvabilité II repose sur l'hypothèse de distribution multivariée elliptique, qui n'est pas toujours vérifiée (voir Maume-Deschamps et al. (2017) pour des résultats sur l'évaluation de quantiles sous l'hypothèse elliptique). Devineau & Loisel (2009) ont donc proposé une nouvelle méthode d'agrégation qui permet de reproduire les résultats obtenus en utilisant un modèle interne. Dans ce contexte, Chauvigny et al. (2011) abordent la question de l'évaluation du quantile à $0,5\%$ d'une distribution, qui est un quantile *éloigné* mais pas extrême, alors que Bargès et al. (2009) proposent une méthode d'agrégation des lignes d'affaire d'un assureur basée sur la TVaR.

L'évaluation des risques extrêmes

Une situation typique où les modèles multivariés interviennent en assurance pour modéliser les risques extrêmes est la suivante. Considérons deux risques V_1 et V_2 qui correspondent chacun au niveau d'une rivière à deux endroits différents. Un évènement extrême survient

lorsque le niveau de la rivière dépasse un seuil donné dans l'une ou l'autre des localisations, i.e. $V_1 \geq v_{1,cat}$ ou $V_2 \geq v_{2,cat}$. Deux problèmes étudiés en théorie des extrêmes multivariés sont d'estimer la probabilité $\mathrm{P}(\{V_1 \geq v_{1,cat}\} \cup \{V_2 \geq v_{2,cat}\})$ de survenance d'un évènement extrême, et la probabilité $\mathrm{P}(\{V_1 \geq v_{1,cat}\} \cap \{V_2 \geq v_{2,cat}\})$ de survenance simultanée des deux extrêmes. Ces quantités sont bien sûr liées à la copule $\mathfrak{C}$ du couple (V_1, V_2) car on a, par exemple pour la première,

$$\begin{aligned} \mathrm{P}(\{V_1 \geq v_{1,cat}\} \cup \{V_2 \geq v_{2,cat}\}) &= 1 - \mathrm{P}(V_1 < v_{1,cat}, V_2 < v_{2,cat}) \\ &= 1 - \mathfrak{C}(F_{V_1}(v_{1,cat}), F_{V_2}(v_{2,cat})). \end{aligned}$$

Une introduction aux risques extrêmes multivariés est présente dans Joe et al. (1992), ainsi que dans Embrechts et al. (2000).

En assurance, les risques extrêmes multivariés peuvent comme dans l'exemple précédent concerner des risques de catastrophe naturelle (Favre et al. (2004)), mais aussi le cyber-risque, le risque nucléaire, le risque opérationnel, le risque d'image ou encore le risque terroriste. Leur appréhension est particulièrement importante pour les réassureurs. Dans le cas du risque tempête/ouragan, Lescourret & Robert (2006) ont modélisé la dépendance entre la survenance de dégats extrêmes (en termes de montant de sinistre) sur les automobiles (assurance auto) et les maisons (assurance habitation). Dans le domaine de la santé, Cebrian et al. (2003) ont analysé parmi les sinistres extrêmes, la dépendance entre les montants de sinistres (Losses) et les frais liés au traitement des sinistres (ALAE : Allocated Loss Adjustment Expenses) à partir de la base de données Loss-ALAE fournie par la Society of Actuaries. D'autres analyses, ainsi que des résultats théoriques dans le cas bivarié, sont donnés dans Di Bernardino et al. (2013).

Les mesures de risque multivariées sont également un sujet important. Différentes extensions de la VaR au cas multivarié ont été étudiées dans Cousin & Di Bernardino (2013). Les mêmes auteurs ont aussi proposé des extensions multivariées pour la TVaR (Cousin & Di Bernardino (2014)).

Autres exemples d'utilisations de modèles multivariés en assurance

Boudreault et al. (2014) proposent une modélisation à l'aide de chaînes de Markov pour le risque Ouragan, et l'appliquent aux données enregistrées en Floride. Boudreault et al. (2017) ont étudié la dépendance entre la fréquence et la magnitude des tremblements de terre à partir des données disponibles pour la ville de Montreal. Enfin, Guibert et al.

(2017) ont proposé un modèle multivarié pour prédire l'évolution de la mortalité simultanément dans différentes populations.

Les modèles multivariés sont également utilisés en assurance pour faire la modélisation jointe de plusieurs durées. Nous approfondissons ce sujet dans la section suivante.

1.3.6 Les modèles de durée multivariés

Dans cette section, nous revenons à l'étude des modèles de durée et nous reprenons donc les notations utilisées dans la Section 1.2. Aussi, nous nous restreignons au cas bivarié. On note (T, U) le couple de durées duquel on souhaite étudier la distribution. Les durées T et U sont sujettes à la censure et l'on note C et D les variables de censure respectives de T et U. On définit alors $Y = \min(T, C)$, $\eta = \mathbb{1}_{T \leq C}$, $Z = \min(U, D)$ et $\gamma = \mathbb{1}_{U \leq D}$, qui correspondent aux variables observées du fait de la censure. Les échantillons de données étudiés sont donc de la forme $(Y_i, Z_i, \eta_i, \gamma_i)_{i=1,\dots,n}$. On suppose par les suite l'hypothèse d'identifiabilité

$$\boldsymbol{H3} : (T, U) \perp\!\!\!\perp (C, D),$$

qui est l'équivalent bivarié de l'hypothèse H0 donnée à la Section 1.2.3.

Pour l'étude de la dépendance entre plusieurs durées, la notion de copule de survie est souvent utilisée. Étant donné une copule bivariée $\mathfrak{C}$, la copule de survie $\mathfrak{C}^*$ associée à $\mathfrak{C}$ est définie par $\mathfrak{C}^*(a, b) = a + b - 1 + \mathfrak{C}(1 - a, 1 - b)$. Soient $S(t, u) = \mathrm{P}(T > t, U > u)$ la fonction de survie jointe, et $S_T = \mathrm{P}(T > t)$, $S_U(u) = \mathrm{P}(U > u)$ les fonctions de survie marginales, alors le Théorème de Sklar (1.8) s'écrit avec la copule de survie

$$S(t, u) = \mathfrak{C}^*(S_T(t), S_U(u)).$$

Le livre de Hougaard (2012) est une introduction aux modèles de durée multivariés.

Les différentes approches d'estimation

Comme dans les Sections 1.3.3 et 1.3.4, nous allons passer en revue les techniques d'estimation non paramétriques, paramétriques et semi-paramétriques.

Dans le cas non paramétrique l'estimation repose, comme à l'équation (1.17), sur la construction d'estimateurs non paramétriques de la fonction de répartition jointe du couple (T, U) : $F(t, u) = \mathrm{P}(T \leq t, U \leq u)$, et des fonctions de répartition marginales $F_T(t) = \mathrm{P}(T \leq t$ et $F_U(u) = \mathrm{P}(U \leq u)$. Étant donné l'hypothèse *H3*, les candidats

naturels pour estimer S_T et S_U sont les estimateurs de Kaplan-Meier (voir Section 1.2.3) de chacune de ces variables, i.e.

$$\begin{aligned}\hat{S}_T(t) &= \prod_{Y_i \leq t} \left(1 - \frac{\eta_i}{\sum_{j=1}^n \mathbb{1}_{Y_j \geq Y_i}}\right), \\ \hat{S}_U(t) &= \prod_{Z_i \leq t} \left(1 - \frac{\gamma_i}{\sum_{j=1}^n \mathbb{1}_{Z_j \geq Z_i}}\right).\end{aligned}$$

On déduit de $\hat{S}_T$ et $\hat{S}_U$ les estimateurs des fonctions de répartition marginales $\hat{F}_T = 1 - \hat{S}_T$ et $\hat{F}_U = 1 - \hat{S}_U$. Concernant l'estimation de la fonction de répartition jointe sous l'hypothèse *H3*, différents estimateurs non paramétriques ont été proposés dans Dabrowska et al. (1988), Van Der Laan et al. (1996), Akritas & Keilegom (2003), ou encore Lopez (2012). Notons donc $\hat{F}^{(d)}$ un tel estimateur. Alors, la copule du couple (T, U) peut être estimée par

$$\hat{\mathfrak{C}}^{(d)}(a, b) = \hat{F}^{(d)}(\hat{F}_T^{-1}(a), \hat{F}_U^{-1}(b)). \tag{1.29}$$

Toujours pour estimer F, Gribkova & Lopez (2015) proposent de considérer des estimateurs de la forme

$$\hat{F}^{(d)}(y, z) = \frac{1}{n} \sum_{i=1}^n W_{i,n} \mathbb{1}_{Y_i \leq y, Z_i \leq z}, \tag{1.30}$$

avec $W_{i,n} = \eta_i \gamma_i \hat{g}(Y_i, Z_i)$ où $\hat{g}$ désigne un estimateur de la fonction $g(y, z) = \mathrm{P}(C \geq y, D \geq z)^{-1}$. Les facteurs $W_{i,n}$ utilisés ici correspondent à des poids IPCW tels que présentés à la Section 1.2.4. En effet, pour une fonction ψ donnée, on peut voir en utilisant l'hypothèse *H3* que

$$\mathrm{E}\left[\frac{\eta\gamma}{g(Y, Z)}\psi(Y, Z)\right] = \mathrm{E}\left[\psi(T, U)\right].$$

En général, l'estimation de la fonction g est complexe et il est possible de l'estimer avec l'un des estimateurs de Dabrowska et al. (1988), Van Der Laan et al. (1996), Akritas & Keilegom (2003) ou Lopez (2012), déjà cités plus haut. Ces cas de figure présentent toutefois un intérêt limité car on pourrait utiliser les estimateurs ci-dessus directement dans la formule (1.29). Gribkova & Lopez (2015) ont étudié différents cas particuliers où la forme d'estimateur (1.30) est très utile. Par exemple, dans le cas où la copule du couple (C, D) est connue, l'estimation de la fonction g est possible à partir des estimateurs KM des marginales $\hat{S}_C$ et $\hat{S}_D$ (Lopez & Saint-Pierre (2012)).

Un problème classique en assurance vie est de modéliser la mortalité dans un couple. Il existe en effet des contrats d'assurance vie, dit sur deux têtes, où en cas de décès de l'un des membres du couple, le conjoint perçoit une indemnité, par exemple sous la forme d'une rente jusqu'à son décès. Pour étudier les risques liés à ce type de contrats, il est donc nécessaire de modéliser conjointement les durées de vie T et U des deux membres du couple. Dans ce contexte, les censures C et D correspondent aux âges des membres du couple à la sortie de l'étude. Les variables C et D sont en fait liées car pour un même contrat, la date de sortie de l'étude, si elle intervient, est la même pour les deux membres du couple. Ainsi, les durées C et D ne diffèrent que par la différence d'âge, connue, entre les deux assurés. En notant $\varepsilon = D - C$ cette différence d'âge, le couple de variables de censure s'écrit donc $(C, C + \varepsilon)$, où ε est une variable observée. La fonction g devient dans ce cas $g(y, z) = \mathrm{P}(C \geq \max(y, z - \varepsilon))^{-1}$, et on peut considérer les poids $W_{i,n} = \eta_i \gamma_i / \hat{S}_C(\max(Y_i, Z_i - \varepsilon_i))$ dans la formule (1.30), avec $\hat{S}_C$ un estimateur KM de la censure. On voit donc que dans cette situation les poids $W_{i,n}$ sont facile à estimer. L'exemple décrit ici est étudié dans Gribkova et al. (2013). Le sujet de l'assurance vie sur deux têtes est également traité dans Carriere (2000) et Youn & Shemyakin (1999, 2001).

Les méthodes d'estimation de copules paramétriques en présence de censure ont été étudiées par Shih & Louis (1995) ainsi que Shih (1998). La méthode du maximum de vraisemblance se généralise au cas censuré même si l'expression de la log-vraisemblance est plus complexe dans ce cas. En effet, pour calculer la vraisemblance des données $(Y_i, Z_i, \eta_i, \gamma_i, X_i)_{i=1,\ldots,n}$, il est nécessaire de prendre en compte quatre cas de figure selon que chacune des durées T et U est censurée ou pas. La vraisemblance des données s'exprime alors

$$\begin{aligned} \mathcal{L}_n = \prod_{i=1}^{n} \Big(& \mathrm{P}(T = Y_i, \eta = 1, U = Z_i, \gamma = 1)^{\eta_i \gamma_i} \\ & \times \mathrm{P}(C = Y_i, \eta = 0, D = Z_i, \gamma = 0)^{(1-\eta_i)(1-\gamma_i)} \\ & \times \mathrm{P}(T = Y_i, \eta = 1, D = Z_i, \gamma = 0)^{\eta_i (1-\gamma_i)} \\ & \times \mathrm{P}(C = Y_i, \eta = 0, U = Z_i, \gamma = 1)^{(1-\eta_i)\gamma_i} \Big), \end{aligned} \tag{1.31}$$

où, pour simplifier l'écriture, on a noté $\mathrm{P}(T = Y_i) = \mathrm{P}(T \in [Y_i, Y_i + dy])$. On trouve donc

$$\begin{aligned}\mathcal{L}_n = \prod_{i=1}^n \Big(& \Big[c(F_T(Y_i), F_U(Z_i)) f_T(Y_i) f_U(Z_i)\Big]^{\eta_i\gamma_i} \\ & \times \Big[1 - F_T(Y_i) - F_U(Z_i) - \mathfrak{C}(F_T(Y_i), F_U(Z_i))\Big]^{(1-\eta_i)(1-\gamma_i)} \\ & \times \Big[f_T(Y_i)\Big(1 - \frac{\partial \mathfrak{C}}{\partial a}(F_T(Y_i), F_U(Z_i))\Big)\Big]^{\eta_i(1-\gamma_i)} \\ & \times \Big[f_U(Z_i)\Big(1 - \frac{\partial \mathfrak{C}}{\partial b}(F_T(Y_i), F_U(Z_i))\Big)\Big]^{(1-\eta_i)\gamma_i} \Big) \times \Big[\dots\Big],\end{aligned} \tag{1.32}$$

où $[\dots]$ désigne une suite de termes qui ne dépendent que de la distribution de la censure (C, D). À partir de cette expression de la vraisemblance, il est possible de calibrer les modèles paramétriques qui portent sur les marginales où sur la copule. Comme nous l'avons vu à la Section 1.3.3, les approches d'optimisation en une étape ou deux étapes sont possibles. Shih (1998) précise que l'impact négatif sur l'estimation des lois marginales est faible si le modèle de copule est mal spécifié. En revanche, une mauvaise spécification des lois marginales peut avoir un impact significatif sur l'estimation de la copule.

De nombreux modèles multivariés paramétriques, non basés sur l'estimation de copules, sont présentés dans Hougaard (2012). Notamment, l'auteur approfondit le sujet des modèles de fragilité multivariés, qui sont également étudiés dans Xue & Brookmeyer (1996). Un modèle mélange multivarié de loi d'Erlang est étudié dans Verbelen et al. (2016).

Le modèle à chocs communs, de Marshall & Olkin (1967), donne une approche paramétrique pour traiter le problème de la modélisation jointe des durées de vie dans un couple. Ce modèle permet de calculer les annuités d'un contrat d'assurance vie sur deux têtes, comme cela est expliqué dans Frees & Valdez (1998). Le modèle de Marshall et Olkin est donc populaire en assurance.

À partir de l'expression de la vraisemblance (1.32), il est possible d'utiliser des méthodes semi-paramétriques pour estimer le paramètre θ d'une copule $\mathfrak{C}$ appartenant à une famille $\mathcal{C} = \{\mathfrak{C}_\theta, \theta \in \Theta\}$. La log-vraisemblance des données s'écrit

$$\begin{aligned}\log \mathcal{L}_n = \sum_{i=1}^n \Big(& \eta_i\gamma_i \log\Big[c_\theta(\hat{F}_T(Y_i), \hat{F}_U(Z_i))\Big] \\ & + (1-\eta_i)(1-\gamma_i)\log\Big[1 - \hat{F}_T(Y_i) - \hat{F}_U(Z_i) - \mathfrak{C}(\hat{F}_T(Y_i), \hat{F}_U(Z_i))\Big] \\ & + \eta_i(1-\gamma_i)\log\Big[1 - \frac{\partial \mathfrak{C}}{\partial a}(\hat{F}_T(Y_i), \hat{F}_U(Z_i))\Big] \\ & + (1-\eta_i)\gamma_i \log\Big[1 - \frac{\partial \mathfrak{C}}{\partial b}(\hat{F}_T(Y_i), \hat{F}_U(Z_i))\Big] \Big) + \Big[\dots\Big],\end{aligned}$$

L'article de Shih & Louis (1995) utilise cette approche d'estimation semi-paramétrique. Dans l'article de Geerdens et al. (2018), cette méthode est également utilisée mais pour estimer une copule conditionnelle, c'est à dire qu'en plus des données $(Y_i, Z_i, \eta_i, \gamma_i)_{i=1,\dots,n}$, des variables explicatives $(X_i)_{i=1,\dots,n}$ sont également observées. Geerdens et al. (2018) utilisent alors, pour l'estimation conditionnelle, le type d'estimateur que nous avons présenté à l'équation (1.26), en pondérant chaque terme de la log-vraisemblance avec un poids $w_{i,n}(x)$ (tel que les poids Nadaraya-Watson (1.23)) qui permet de localiser l'estimation de θ dans un voisinage de $X = x$. De plus, l'estimateur Kaplan-Meier conditionnel présenté dans la Section 1.2.5 est utilisé pour estimer les lois marginales conditionnelles $\hat{F}_T(\cdot|x)$ et $\hat{F}_U(\cdot|x)$, de sorte que la (pseudo) log-vraisemblance optimisée en chaque point x est

$$\begin{aligned}\log \mathcal{L}_n = \sum\nolimits_{i=1}^{n} w_{i,n}(x) \quad \Big(\quad & \eta_i\gamma_i \log\Big[c_\theta(\hat{F}_T(Y_i|X_i), \hat{F}_U(Z_i|X_i))\Big] \\ & + (1-\eta_i)(1-\gamma_i)\log\Big[1 - \hat{F}_T(Y_i|X_i) - \hat{F}_U(Z_i|X_i) \\ & \qquad - \mathfrak{C}(\hat{F}_T(Y_i|X_i), \hat{F}_U(Z_i|X_i))\Big] \\ & + \eta_i(1-\gamma_i)\log\Big[1 - \frac{\partial \mathfrak{C}}{\partial a}(\hat{F}_T(Y_i|X_i), \hat{F}_U(Z_i|X_i))\Big] \\ & + (1-\eta_i)\gamma_i \log\Big[1 - \frac{\partial \mathfrak{C}}{\partial b}(\hat{F}_T(Y_i|X_i), \hat{F}_U(Z_i|X_i))\Big] \quad \Big).\end{aligned} \tag{1.33}$$

Lopez (2018) utilise également une approche semi-paramétrique et conditionnelle pour mesurer la dépendance entre la durée d'un sinistre et le montant de sinistre ultime. Dans cette situation la variable de censure des deux phénomènes étudiés est identique, i.e. $C = D$, de plus T (la durée de sinistre) et U (le montant de sinistre) sont observés, ou non observés, simultanément (i.e. $\eta = \gamma$). En faisant l'hypothèse que le paramètre de copule est constant (i.e. $\theta(x) = \theta$: parfois appelée *simplifying assumption*), Lopez (2018) tire profit de la méthode des poids IPCW pour proposer d'utiliser la log-vraisemblance conditionnelle

$$\log \mathcal{L}_n = \sum\nolimits_{i=1}^{n} W_{i,n} \log\big[c_\theta(\hat{F}_T(Y_i|X_i), \hat{F}_U(Z_i|X_i))\big], \tag{1.34}$$

avec $W_{i,n} = \eta_i\gamma_i/\hat{S}_C(Y_i)$, et $\hat{S}_C$ l'estimateur de Kaplan-Meier de la durée de censure C. Dans ce cas de figure, l'utilisation de poids IPCW permet donc de proposer une forme simplifiée de log-vraisemblance, qui diffère uniquement de la log-vraisemblance des données non censurées par l'introduction des poids $W_{i,n}$ et l'absence des poids $w_{i,n}(x)$ du fait de la *simplifying assumption*. D'un point de vue pratique, cette approche présente un avantage car il est possible d'utiliser pour le cas censuré des outils développés pour le cas non censuré, simplement en pondérant les observations par les poids $W_{i,n}$. Par ailleurs,

l'estimateur obtenu en utilisant la log-vraisemblance (1.34) est compétitif à distance finie, comparé à celui obtenu avec la vraisemblance (1.33), grâce à la simplicité d'optimisation qui évite de tomber sur des maxima locaux.

Les mesures de concordance en présence de censure

Dans la Section 1.3.2, nous avons présenté différentes mesures non paramétriques de concordance utilisées fréquemment dans le cas non censuré pour évaluer le degré d'association entre deux variables. En présence de censure bivariée, il n'est pas possible de calculer ces mesures. Un problème naturel est donc de chercher des quantités calculables dans le cas censuré, que l'on pourrait utiliser en remplacement des mesures vues à la Section 1.3.2.

Lorsque seule une des deux variables est censurée, par exemple T est censuré et U ne l'est pas (i.e. $D = +\infty$ donc $Z = U$), le C-index a été proposé par Harrell et al. (1982) comme une généralisation du tau de Kendall. Rappelons que le tau de Kendall (1.12) est égal à la proportion de paires bien ordonnées (ou concordantes i.e. $(T_i - U_i)(T_j - U_j) > 0$), moins la proportion de paires mal ordonnées. Le principe est le même pour le C-index, seulement du fait de la censure, toutes les paires d'observations ne sont pas ordonnées. En fait, deux observations (Y_i, η_i, Z_i) et (Y_j, η_j, Z_j) (rappelons que $\gamma_i = \gamma_j = 1$) sont ordonnées si et seulement si

$$\begin{cases} \eta_i = \eta_j = 1, \\ \quad \text{ou} \\ \eta_i = 1, \eta_j = 0 \text{ et } Y_i < Y_j, \\ \quad \text{ou} \\ \eta_i = 0, \eta_j = 1 \text{ et } Y_j < Y_i. \end{cases}$$

Cette condition est en fait équivalente à

$$\begin{cases} \eta_i = 1 \text{ et } Y_i < Y_j, \\ \quad \text{ou} \\ \eta_j = 1 \text{ et } Y_j < Y_j. \end{cases}$$

Ainsi le nombres de paires ordonnées est égal à

$$\sum_{i=1}^{n} \sum_{j>i} \eta_i \mathbb{1}_{Y_i<Y_j} + \eta_j \mathbb{1}_{Y_j<Y_i} = \sum_{i=1}^{n} \sum_{j=1}^{n} \eta_i \mathbb{1}_{Y_i<Y_j}.$$

Le C-index est égal à la proportion de paires bien ordonnées dans l'ensemble des paires ordonnées, i.e.

$$C_{index} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\eta_i \mathbb{1}_{Y_i<Y_j, Z_i<Z_j}}{\sum_{i=1}^{n}\sum_{j=1}^{n}\eta_i \mathbb{1}_{Y_i<Y_j}}. \tag{1.35}$$

En notant

$$n_c^{(c1)} = \sum_{i=1}^{n}\sum_{j=1}^{n}\eta_i \mathbb{1}_{Y_i<Y_j, Z_i<Z_j}, \quad n_d^{(c1)} = \sum_{i=1}^{n}\sum_{j=1}^{n}\eta_i \mathbb{1}_{Y_i<Y_j, Z_i>Z_j},$$

le nombre de paires (ordonnées) concordantes et discordantes, on a $C_{index} = n_c^{(c1)}/(n_c^{(c1)} + n_d^{(c1)})$. Remarquons qu'une autre différence entre le tau de Kendall et le C-index est la transformation affine $Aff(x) = (x+1)/2$, qui fait que $C_{index} \in [0,1]$, alors le tau de Kendall $\tau \in [-1,1]$. Enfin, notons que le C-index est en général un estimateur biaisé de la proportion de paires bien ordonnées. A l'aide de poids IPCW une forme du C-index non biaisée asymptotiquement a été proposée par Uno et al. (2011). Toutefois, du fait de sa simplicité, la forme du C-index la plus utilisée en pratique est celle de l'équation (1.35).

Lorsque les deux durées sont censurées, un estimateur du tau de Kendall construit suivant le même principe que le C-index ci-dessus a été proposé par Oakes (2008). Ce dernier correspond à la proportion de paires $((Y_i, \eta_i, Z_i, \gamma_i), (Y_j, \eta_j, Z_j, \gamma_j))$ bien ordonnées (dans l'ensemble des paires ordonnées), moins la proportion de paires mal ordonnées. En faisant un raisonnement similaire à celui fait pour le C-index, on peut voir que l'estimateur de Oakes (2008) s'exprime

$$\hat{\tau}^{(c2)} = \frac{n_c^{(c2)} - n_d^{(c2)}}{n_c^{(c2)} + n_d^{(c2)}},$$

avec

$$n_c^{(c2)} = \sum_{i=1}^{n}\sum_{j=1}^{n}\eta_i\gamma_i \mathbb{1}_{Y_i<Y_j, Z_i<Z_j}, \quad n_d^{(c2)} = \sum_{i=1}^{n}\sum_{j=1}^{n}\eta_i\gamma_j \mathbb{1}_{Y_i<Y_j, Z_i>Z_j}.$$

Comme pour le C-index, une version de cette estimateur sans biais asymptotique, qui utilise les poids IPCW, a été proposée par Lakhal et al. (2009). Dans l'article de Lakhal et al. (2009), différentes situations de censure sont étudiées (variables de censure C et D indépendantes, censure univariée $C = D$), à l'image des cas particuliers que nous avons mentionnés au début de cette section pour estimer la fonction g. L'estimation du tau de Kendall en présence de censure est également l'objet des articles de Wang & Wells (2000) et Oakes (2008). Dans la section suivante, nous définissons un estimateur du tau de

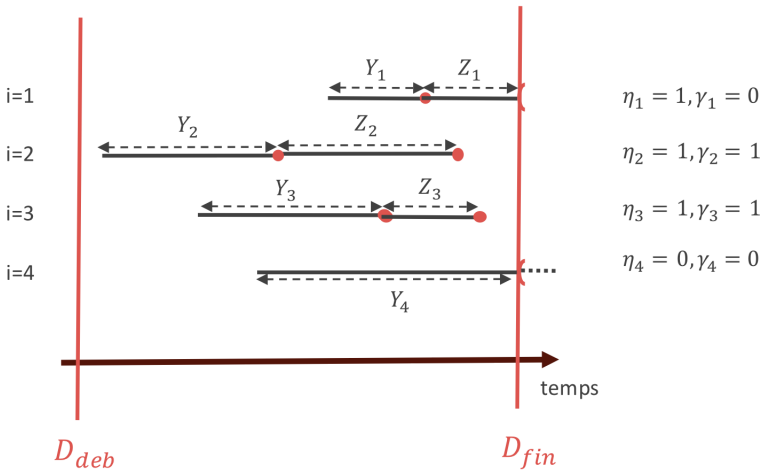

Fig. 1.4: Situation de deux durées successives. Seuls 3 cas de censure sont possibles car le cas $\eta = 0$ et $\gamma = 1$ ne peux pas se produire.

Kendall conditionnel, où seules les paires d'observations non censurées (i.e. $\eta_i\gamma_i\eta_j\gamma_j = 1$) sont utilisées.

Le cas de la dépendance entre deux durées successives

Nous terminons cette section sur les modèles de durée bivariés en abordant le sujet de la dépendance entre deux durées successives, qui est l'objet de notre travail du Chapitre 3. Dans l'exemple que nous étudions au Chapitre 3, la première durée T correspond à la durée entre la souscription d'un contrat d'assurance et la prise d'effet de ce contrat, alors que la seconde durée U correspond à la durée entre la prise d'effet du contrat et sa résiliation. Les données étudiées, qui sont décrites plus en détails au Chapitre 2, concernent des contrats d'assurance santé.

La situation décrite ci-dessus correspond au cas où $D = C - T$, c'est à dire qu'il n'y a qu'une seule variable de censure C pour les deux durées successives. Le schéma de la Fig. 1.4 illustre la situation. On peut alors remarquer que l'hypothèse *H3* n'est pas vérifiée. Toutefois, on suppose que le couple (T, U) est indépendant de C. Dans ce contexte, Wang & Wells (1998) ont étudié un estimateur non paramétrique de la fonction de répartition bivariée, alors que Meira-Machado et al. (2016) ont proposé un estimateur non paramétrique de la fonction de répartition de U sachant T.

Nous abordons maintenant les approches d'estimation semi-paramétriques. En partant de la vraisemblance (1.31), on peut voir que sur les 4 cas possibles rencontrés dans les situations de censure précédentes, on passe à seulement 3 cas possibles. En effet, le cas où la première durée T est censurée ($\eta = 0$) et la seconde durée U n'est pas censurée

($\gamma = 1$) ne peut pas se produire. En considérant une famille paramétrique de copules $\mathcal{C} = \{\mathfrak{C}_\theta, \theta \in \Theta\}$, ainsi que des estimateurs non paramétriques des marginales $\hat{F}_T$ et $\hat{F}_U$, on peut développer la vraisemblance (1.31) pour obtenir que la (pseudo) log-vraisemblance vaut

$$\log \mathcal{L}_n = \sum_{i=1}^{n} \eta_i \gamma_i \log \Big(c_\theta(\hat{F}_T(Y_i), \hat{F}_U(Z_i)) \Big) + \eta_i (1-\gamma_i) \log \Big(1 - \frac{\partial \mathfrak{C}_\theta}{\partial a}(\hat{F}_T(Y_i), \hat{F}_U(Z_i)) \Big).$$

Comme nous l'avons fait pour l'équation (1.34), il est également possible dans la situation de censure étudiée ici, d'utiliser un estimateur de la log-vraisemblance qui fasse intervenir les poids IPCW. Nous montrons au Chapitre 3 que les poids IPCW à considérer dans cet exemple sont de la forme $W_{i,n} = \eta_i \gamma_i / \hat{S}_C(Y_i + Z_i)$, avec $\hat{S}_C$ un estimateur de la fonction de survie de la censure C. Dans ce cas, on trouve la log-vraisemblance

$$\log \mathcal{L}_n = \sum_{i=1}^{n} W_{i,n} \log \Big(c_\theta(\hat{F}_T(Y_i), \hat{F}_U(Z_i)) \Big).$$

Au Chapitre 3, nous nous plaçons dans le cas conditionnel et disposons donc de réalisations $(X_i)_{i=1,\dots,n}$ d'un vecteur X de variables explicatives dans le but d'étudier la dépendance conditionnelle entre T et U. Nous supposons de plus que (T, U, X) est indépendant de C. Nous considérons alors, comme nous l'avons fait à l'équation (1.34), des poids Nadaraya-Watson $w_{i,n}(x)$ ainsi que des estimateurs des lois marginales conditionnelles $\hat{F}_T(t|x)$ et $\hat{F}_U(u|x)$. Un choix possible pour ces derniers est

$$\begin{aligned} \hat{F}_T(t|x) &= \sum_{i=1}^{n} W_{i,n} w_{i,n}(x) \mathbb{1}_{Y_i \le t}, \\ \hat{F}_U(u|x) &= \sum_{i=1}^{n} W_{i,n} w_{i,n}(x) \mathbb{1}_{Z_i \le u}, \end{aligned}$$

avec toujours $W_{i,n} = \eta_i \gamma_i / \hat{S}_C(Y_i + Z_i)$. L'estimateur de $x \to \theta(x)$ étudié au Chapitre 3 est donné par

$$\hat{\theta}(x) = \arg\max_{\theta \in \Theta} \ \log \mathcal{L}_n, \tag{1.36}$$

avec

$$\log \mathcal{L}_n = \sum_{i=1}^{n} w_{i,n}(x) W_{i,n} \log \Big(c_\theta(\hat{F}_T(Y_i|X_i), \hat{F}_U(Z_i|X_i)) \Big).$$

Pour évaluer l'adéquation des différentes copules avec les données, nous utilisons au

Chapitre 3 un estimateur du tau de Kendall conditionnel. En reprenant les idées développées dans le paragraphe précédent sur les mesures de concordance en présence de censure, nous utilisons l'estimateur suivant. Soit $W_{i,n}(x) = w_{i,n}(x)W_{i,n}$, nous définissons $n_c^{(w,x)}$ et $n_d^{(w,x)}$ les estimateurs conditionnels du nombre de paires concordantes et discordantes par

$$n_c^{(w,x)} = \sum_{i=1}^{n}\sum_{j=1}^{n} W_{i,n}(x)W_{j,n}(x)\mathbb{1}_{Y_i<Y_j, Z_i<Z_j}, \quad n_d^{(w,x)} = \sum_{i=1}^{n}\sum_{j=1}^{n} W_{i,n}(x)W_{j,n}(x)\mathbb{1}_{Y_i<Y_j, Z_i>Z_j}.$$

Nous pouvons alors estimer le tau de Kendall conditionnel avec l'estimateur

$$\tau^{(w,x)} = \frac{n_c^{(w,x)} - n_d^{(w,x)}}{n_c^{(w,x)} + n_d^{(w,x)}}.$$

1.4 Méthodes d'arbre de régression pour le problème de la prédiction de durée : applications à l'assurance

Les méthodes d'arbre de régression qui s'appliquent aux données censurées à droites sont l'objet du Chapitre 4 de notre thèse. Dans cette section, nous commençons par motiver l'utilisation de telles méthodes dans le cadre assurantiel, puis nous présentons les principaux algorithmes présents dans l'état de l'art, utilisés au Chapitre 4.

1.4.1 Les enjeux liés à la prédiction de durée en assurance

Nous avons vu dans la Section 1.2.5 que l'utilisation d'informations supplémentaires portant sur les observations (vecteur X_i) peut, si l'hypothèse H2 est vérifiée, permettre l'estimation de la distribution de T dans un cas où l'hypothèse H0 ne serait pas satisfaite. Les applications du recueil des données clients ne s'arrêtent pas, toutefois, à l'évaluation juste du risque moyen. En effet, un enjeu important est d'individualiser les prédictions faites pour chaque observation. Une telle politique est bénéfique car elle permet de mieux mesurer les risques et de les anticiper, mais aussi de personnaliser les offres. Étant donné un ensemble de caractéristiques connues X, une question récurrente en assurance est de prédire à quelle date, ou horizon de temps, va intervenir un évènement tel un décès ou une résiliation de contrat. Cela revient alors à modéliser la loi $\mathcal{L}(T|X)$ d'une durée T sachant les informations X ou bien l'espérance de cette loi $E[T|X]$.

L'influence de divers facteurs tels que l'âge, la catégorie socioprofessionnelle, ou le sexe, sur la durée de maintien en invalidité suite à une maladie ou à un accident a ainsi été étudiée par Lopez et al. (2016). Sur un portefeuille d'assurance vie, Denuit & Legrand (2018) ont mesuré, à l'aide de modèles additifs généralisés (GAM), l'impact de l'âge et du capital assuré sur le risque de mortalité d'un individu. Milhaud (2013) a caractérisé la durée de rachat de contrats d'assurance vie à partir de variables économiques (niveau de taux d'intérêt, taux de chômage...) et de variables individuelles (sexe, âge, montant assuré, option de partage des profits...). Pour des crédits à la consommation, le risque de crédit d'un individu (ou la durée avant la survenance d'un défaut de paiement) a été évalué par Stepanova & Thomas (2002) en prenant en compte des variables telles que le montant du crédit ou le nombre d'enfants à charge. Henebry (1997) a utilisé les historiques des opérations de prêt des sociétés pour prédire leurs risques de défaut de paiement. Enfin, en modélisant la durée avant la survenance d'un accident de voiture avec un modèle de Cox (1972), Caragata Nasvadi & Wister (2009) ont réalisé une étude d'impact pour une mesure visant à allouer des permis de conduire restreints aux personnes âgées présentant un risque d'accident élevé.

Le trait commun à toutes ces applications est qu'elles consistent en la prédiction d'une durée T à partir de caractéristiques X. Dans la suite de cette partie, nous présentons les méthodes d'arbre de régression qui répondent à ce problème en s'appliquant spécifiquement aux données censurées à droite. Dans la section suivante, nous commençons par présenter les arbres CART (Classification And Regression Tree) et la forêt aléatoire dans un contexte non spécifique aux données censurées.

1.4.2 Les arbres CART et la forêt aléatoire

Dans cette section exclusivement, nous nous plaçons dans le cas où la variable à expliquer, notée V, n'est pas censurée. L'algorithme CART (Breiman et al. (1984)) et l'algorithme de forêt aléatoire (Breiman (2001)) sont des méthodes d'apprentissage statistique très utilisées, appliquées sur des données non censurées pour résoudre des problèmes de régression et de classification. Nous allons présenter ces algorithmes dans le contexte de la régression, où le problème consiste à estimer la fonction $f(x) = E[V|X = x]$, dite fonction de régression.

Les arbres CART

Soit $(V_i, X_i)_{i=1,\dots,n}$ un échantillon de n observations et $\mathcal{D} = \{1, \dots, n\}$ l'ensemble des indices de l'échantillon. Pour une partie $B \subset \mathcal{X}$, nous notons $n^{\mathcal{D}}(B) = \sum_{i \in \mathcal{D}} \mathbb{1}_{X_i \in B}$ le nombre d'observations telles que $X_i \in B$ et $\bar{V}_B^{\mathcal{D}} = 1/n^{\mathcal{D}}(B) \sum_{i \in \mathcal{D}} V_i \mathbb{1}_{X_i \in B}$ la moyenne empirique de V pour les observations appartenant à B. Pour un ensemble E, nous notons également $Card(E)$ le cardinal de cet ensemble (i.e. son nombre d'éléments).

L'algorithme CART consiste à construire un arbre binaire, tel que celui représenté à la Fig. 1.5, à partir du noeud initial constitué de l'échantillon complet $\mathcal{D}$. Chaque séparation en deux branches correspond à une segmentation d'un sous-ensemble d'observations en deux parties. Par exemple, la séparation issue du noeud d'origine segmente $\mathcal{D}$ en deux groupes d'observations. Les coupures effectuées en chaque noeud correspondent à des règles binaires de type $(X^j \leq u)$, où X^j désigne la j-ième dimension de X ($j \in \{1, \dots, p\}$) et u est un réel. Une fois construit, un arbre binaire détermine une partition de l'espace $\mathcal{X}$ en sous-ensembles $(\mathcal{X}_k)_{k=1,\dots,K}$ (aussi appelés feuilles), ainsi qu'une partition de l'échantillon $\mathcal{D}$ en sous-échantillons. On associe alors à cet arbre l'estimateur, constant par morceaux, de la fonction de régression $f(x)$ donné par $\hat{m}(x) = \sum_{k=1}^{K} \bar{V}_{\mathcal{X}_k}^{\mathcal{D}} \mathbb{1}_{X_i \in \mathcal{X}_k}$.

À taille de partition K fixée, les coupures $(X^j \leq u)$ sont optimisées pour que l'estimateur $\hat{m}$ soit à distance minimale du nuage de points formé par les observations $(V_i, X_i)_{i \in \mathcal{D}}$. En effet, on définit pour un noeud de l'arbre $A \subset \mathcal{X}$,

$$L(u, j, A, \mathcal{D}) = \frac{1}{n^{\mathcal{D}}(A)} \cdot \sum_{i \in \mathcal{D}} \left(V_i - \bar{V}_{A_l}^{\mathcal{D}} \mathbb{1}_{X_i^j \leq u} - \bar{V}_{A_r}^{\mathcal{D}} \mathbb{1}_{X_i^j > u} \right)^2 \cdot \mathbb{1}_{X_i \in A}, \tag{1.37}$$

avec $A_l = A \cap \{X^j \leq u\}$ et $A_r = A \cap \{X^j > u\}$. Étant donné un couple (u, j), L mesure une distance (erreur quadratique) entre les observations contenues dans A : $\{i \in \mathcal{D} / X_i \in A\}$, et le modèle simple qui attribue les prédictions $\bar{V}_{A_l}^{\mathcal{D}}$ et $\bar{V}_{A_r}^{\mathcal{D}}$ aux ensembles A_l et A_r, respectivement. Les paramètres j et u sont alors choisis en chaque noeud de sorte qu'ils minimisent le critère L.

Un autre élément important de l'algorithme CART est le critère qui décide de l'arrêt de la subdivision d'un noeud en sous-ensembles. Dans la version originale de Breiman et al. (1984), un noeud A n'est pas divisé si son effectif $n(A)$ est inférieur à une constante spécifiée en paramètre à la procédure (nous appellerons ce paramètre *minsplit*). Toutefois, d'autres critères sont couramment employés, notamment pour l'usage de CART dans le cadre de l'algorithme de forêt aléatoire (voir Algorithme 2). La profondeur d'un noeud A dans un arbre $\mathcal{T}$ (évaluée dans l'Algorithme 1 en appelant la fonction $Profondeur$)

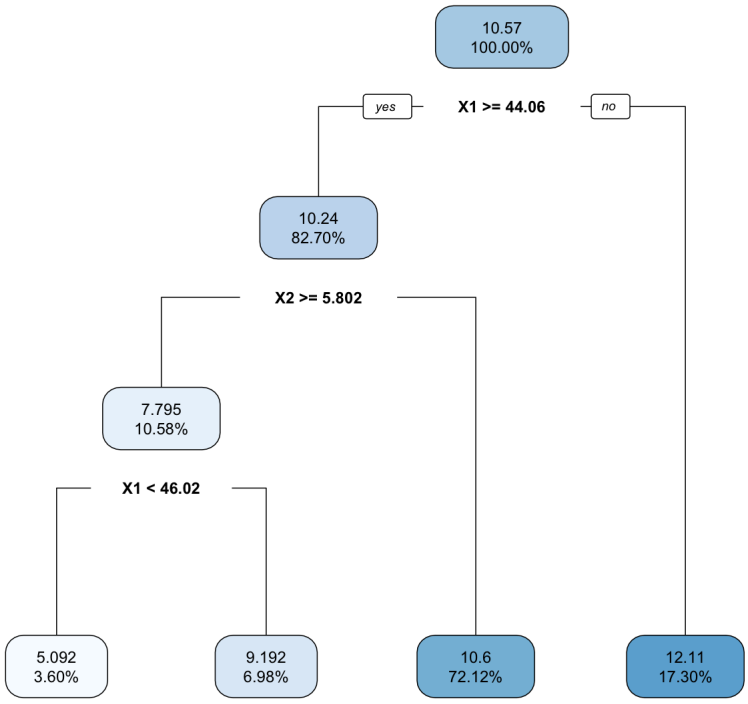

Fig. 1.5: Exemple d'arbre CART avec deux variables explicatives X1 et X2. La profondeur de l'arbre est de 4.

correspond à la longueur du chemin qui descend depuis le noeud d'origine de $\mathcal{T}$ dans le noeud A, avec la convention que la profondeur du noeud d'origine est 1. Par extension, la profondeur de l'arbre $\mathcal{T}$ correspond à la profondeur maximale d'une feuille de l'arbre. Une autre méthode pour stopper la division d'un noeud consiste alors à fixer une profondeur maximale (paramètre que nous appellerons *maxdepth*) au-delà de laquelle un noeud n'est pas divisé.

La procédure CART est détaillée dans l'Algorithme 1. Notons que nous avons introduit 3 modifications majeures dans cet algorithme par rapport à l'algorithme original de Breiman et al. (1984). En effet, pour simplifier la présentation de l'algorithme de forêt aléatoire dans la partie suivante, nous avons ajouté les paramètres *maxdepth* (cf. paragraphe précédent) et *mtry* à la procédure. Dans la version originale, on ne tirerait pas de sous-ensemble de variables $\mathcal{M}_{try}$ et la ligne 10 serait "**pour** $j \in \{1, \ldots, p\}$ **faire**". De plus, l'Algorithme 1 ne fait pas mention de la partie *élagage* présente dans l'algorithme original.

Algorithme 1 : CART

Entrée : Données : $(V_i, X_i)_{i\in\mathcal{D}}$, $mtry \in \{1,\dots,p\}$, $maxdepth \in \mathbb{N}^*$, $minsplit \in \mathbb{N}^*$
Sortie : Arbre binaire $\mathcal{T}$, estimateur $\hat{m}$.

`// Construction de l'arbre` $\mathcal{T}$

Initialiser $\mathcal{T} = (\mathcal{X})$ l'arbre formé du noeud initial.
Initialiser $\mathcal{A} = \{\mathcal{X}\}$ l'ensemble des noeuds terminaux de $\mathcal{T}$.
tant que $Card(\mathcal{A}) > 0$ **faire**
  Définir A le premier élément de $\mathcal{A}$.
  si $Profondeur(A) = maxdepth$, $n(A) \leq minsplit$ *ou si tous les* $X_i \in A$ *sont égaux* **alors**
    Retirer l'élément A de $\mathcal{A}$.
  fin
  sinon
    Tirer uniformément, sans remise, un sous-ensemble $\mathcal{M}_{try} \subset \{1,\dots,p\}$ de cardinal $mtry$.
    pour $j \in \mathcal{M}_{try}$ **faire**
      Trouver u_j tel que

$$u_j = \underset{u\in\left\{X_i^j /\ i\in\mathcal{D}\ t.q.\ X_i\in A\right\}}{\arg\min} L(u, j, A, \mathcal{D}).$$

    fin
    Trouver $j^* = \underset{j\in\mathcal{M}_{try}}{\arg\min}\ L(u_j, j, A, \mathcal{D})$.
    Définir $A_l^* = A \cap \{X^{j^*} \leq u_{j^*}\}$ et $A_r^* = A \cap \{X^{j^*} > u_{j^*}\}$.
    Ajouter A_l^* et A_r^* à l'ensemble $\mathcal{A}$.
    Augmenter l'arbre $\mathcal{T}$ des noeuds A_l^* et A_r^*.
    Retirer l'élément A de $\mathcal{A}$.
  fin
fin

`// Construction de l'estimateur` $\hat{m}$

Définir $\mathcal{X}_1,\dots,\mathcal{X}_K$ les K noeuds terminaux de l'arbre $\mathcal{T}$.
Définir $\hat{m}(x) = \sum_{k=1}^{K} \bar{V}_{\mathcal{X}_k}^{\mathcal{D}} \mathbb{1}_{X_i\in\mathcal{X}_k}$.
renvoyer : $\mathcal{T}$, $\hat{m}$

La forêt aléatoire

L'algorithme de forêt aléatoire (Breiman (2001)) est construit à partir de l'algorithme CART. Nous allons voir qu'il s'agit en fait d'un assemblage de différents arbres CART construits en parallèles et en respectant une même procédure.

L'idée du bagging (contraction de bootstrap-aggregating) a été introduite par Breiman (1996). Le principe se résume en deux phases : dans un premier temps, tirer uniformément et avec remise un nombre M d'échantillons bootstrap (que nous noterons $\mathcal{D}_n$) à partir de l'échantillon initial $\mathcal{D}$ et calibrer sur chacun d'eux un même modèle statistique (phase bootstrap) ; dans un second temps, agréger les M modèles obtenus, avec une moyenne par exemple, pour former le modèle final (phase d'agrégation). L'expérimentation montre que, utilisé pour calibrer un modèle statistique, le bagging aboutit en général à un meilleur modèle, c'est à dire à une erreur de prédiction plus faible, que l'approche classique qui consiste à calibrer une seule fois le modèle sur l'échantillon entier $\mathcal{D}$. Breiman souligne que l'amélioration obtenue est importante lorsque les modèles individuels calibrés sur les échantillon $\mathcal{D}_n$ ont une forte variance (i.e. forte sensibilité du modèle à l'échantillon $\mathcal{D}_n$, forte variabilité dans la procédure de calibrage du modèle), mais que toutefois, appliqué dans une situation où les modèles individuels varient peu, le bagging n'a pas d'effet positif et peut même conduire à une diminution sensible de la précision du modèle.

La combinaison des arbres CART et du bagging est proposée par Breiman en 1996. Breiman montre que les performances de l'algorithme CART sont grandement améliorées lorsque que ce dernier est associé au bagging. Les résultats de la procédure CART sont en effet très sensibles aux variations d'échantillon (Breiman et al. (1996)) car la modification de la coupure choisie en un noeud de l'arbre se répercute sur tous les noeuds enfants qui en découlent. Ces réflexions aboutissent au développement du concept de forêt aléatoire dans Breiman (2001).

L'Algorithme 2 décrit la procédure de forêt aléatoire. On peut voir que la forêt aléatoire correspond en fait au bagging d'arbres CART. Plus précisément, il s'agit d'une version avancée du bagging d'arbre CART, car différentes améliorations sont apportées par Breiman. Plutôt que de chercher, en chaque noeud, la coupure optimale parmi l'ensemble des coupures possibles, Breiman propose de tirer aléatoire (uniformément, sans remise) pour chaque noeud, un nombre $mtry$ de variables explicatives parmi lesquelles la coupure optimale est recherchée (lignes 9 à 13 dans l'Algorithme 1). Bien que cette technique rende chaque prédicteur individuel moins précis, Breiman s'est aperçu qu'associée au bagging cette modification apportait souvent une amélioration de l'estimateur agrégé de

la forêt aléatoire : plus précis, plus robuste aux observations extrêmes, et aussi plus rapide à calculer. Ce dernier point qui concerne les temps de calcul est particulièrement important lors de l'utilisation de l'algorithme en grande dimension (lorsque p est grand). D'autre part, Breiman souligne l'importance de ne pas élaguer les arbres individuels de la forêt aléatoire, c'est à dire de ne pas sélectionner un sous-arbre optimal à partir d'un arbre initialement construit comme cela est proposé dans l'algorithme CART. Enfin, il explore la piste des quantités *out-of-bag*, qui permettent de suivre en continu l'amélioration du modèle agrégé au fil de l'ajout de nouveaux prédicteurs individuels, et de mesurer l'importance de chaque variable explicative.

Algorithme 2 : Forêt aléatoire

Entrée : Données : $(V_i, X_i)_{i=1,..,n}$, $M \in \mathbb{N}^*$ the number of trees, $mtry \in \{1, \ldots, p\}$, $maxdepth \in \mathbb{N}^*$, $minsplit \in \mathbb{N}^*$.

Sortie : Ensemble des arbres RF, estimateur $\hat{m}_{RF}$.

1 Initialiser $RF = \{\}$ l'ensemble des arbre de la forêt aléatoire.
2 Initialiser $\hat{m}_{RF} = 0$ l'estimateur de la fonction de régression.
3 **pour** $j = 1, \ldots, M$ **faire**
4 Construire l'échantillon bootstrap $\mathcal{D}_{n,j}$ en tirant uniformément, avec remise, n observations dans l'ensemble $\mathcal{D}$.
5 Construire l'arbre CART $\mathcal{T}_j$ à partir de l'échantillon $\mathcal{D}_{n,j}$, en utilisant l'Algorithme 1, et en respectant les valeurs des arguments $mtry$, $maxdepth$ et $minsplit$.
6 Ajouter l'arbre T_j à l'ensemble RF.
7 **fin**
8 Construire l'estimateur $\hat{m}_{RF} = \frac{1}{M}\sum_{j=1}^{M} \hat{m}_j$ où chaque estimateur $\hat{m}_j$ est calculé à partir de l'arbre $\mathcal{T}_j$.

renvoyer : RF, $\hat{m}_{RF}$

Du fait de la censure des données de durée $(T_i)_{i=1,..,n}$ et des biais évoqués à la Section 1.2.4, les algorithmes décrits dans cette section ne permettent pas d'estimer $E[T|X = x]$. Toutefois, différentes adaptations de l'algorithme CART et de la forêt aléatoire ont été proposées pour répondre à ce problème. Nous présentons dans les parties suivantes les principales méthodes rencontrées dans la littérature.

1.4.3 Une première approche : L'algorithme RSF (Random Survival Forest)

L'algorithme RSF (*Random Survival Forest*) a été introduit par Ishwaran et al. (2008). Dans cet algorithme, un test *logrank* est utilisé en chaque noeud pour déterminer la

coupure optimale. A ce titre, la procédure décrite ci-dessous pour construire un arbre de régression en présence de censure reprend les idées développées par Ciampi et al. (1986), puis Segal (1988). Elle a ensuite été étudiée dans LeBlanc & Crowley (1993), puis par Hothorn, Hornik, & Zeileis (2006) dans le contexte des *conditional inference trees*.

L'algorithme RSF généralise le concept de forêt aléatoire au cas des données censurées à droite en permettant d'estimer la fonction de survie conditionnelles d'une durée T (i.e. $S_T(\cdot|X)$) à partir des données $(Y_i, \delta_i, X_i)_{i=1,\dots,n}$. Le principe de *bagging*, utilisé dans la forêt aléatoire, est employé de façon similaire dans RSF. À ce titre, l'algorithme 2 décrivant la forêt aléatoire est également valable pour RSF, et seule diffère la méthode employée pour construire les arbres individuels lors de l'appel à la procédure CART à la ligne 5 de l'algorithme.

Comme souligné par Breiman et al. (1984), il est nécessaire de définir trois éléments pour construire un arbre de régression de type CART : un *critère de coupure* (split criteria) pour sélectionner une coupure à chaque noeud intermédiaire, un *critère d'arrêt* pour déterminer si un noeud est terminal, et un *estimateur de feuille terminale* pour associer à chaque feuille de l'arbre une prédiction. Dans le cas de l'algorithme RSF, les critères d'arrêt utilisés sont le *minsplit* ou le *maxdepth* comme présentés pour la forêt aléatoire dans la partie précédente. Néanmoins, il est nécessaire d'adapter le critère de coupure et l'estimateur de feuille terminal du fait de la censure des données.

L'algorithme RSF, dont l'implémentation est détaillée dans Ishwaran & Kogalur (2007), utilise le test *logrank* pour déterminer la coupure choisie en chaque noeud. Dans les explications qui suivent, nous nous plaçons dans le cas général où les variables T et C ne sont pas supposées continues. Étant donné un noeud noté A, soient $t_1 < t_2 < \ldots < t_H$ les instants distincts auxquels un évènement a été observé pour une observation du noeud A. Pour une variable j et un réel u, on définit alors $A_l = A \cap \{X^j \leq u\}$ et $A_r = A \cap \{X^j > u\}$. Pour $i = 1, \dots, H$, soient $d_{i,A_l} = \sum_{j=1}^n \delta_j \mathbb{1}_{Y_j = t_i, X_j \in A_l}$ (resp. $d_{i,A} = \sum_{j=1}^n \delta_j \mathbb{1}_{Y_j = t_i, X_j \in A}$) le nombre d'évènements survenu à l'instant t_i pour une observation contenue dans A_l (resp. A), et $n_{i,A_l} = \sum_{j=1}^n \mathbb{1}_{Y_j \geq t_i, X_j \in A_l}$ (resp. $n_{i,A} = \sum_{j=1}^n \mathbb{1}_{Y_j \geq t_i, X_j \in A}$) le nombre d'observations à risque à l'instant t_i dans le noeud A_l (resp. A). La statistique *logrank* (Mantel (1966), Peto & Peto (1972)) est définie par

$$L^{RSF}(u, j, A, \mathcal{D}) = \frac{\sum_{i=1}^H \left(d_{i,A_l} - n_{i,A_l} \frac{d_{i,A}}{n_{i,A}} \right)}{\sqrt{\sum_{i=1}^H \frac{n_{i,A_l}}{n_{i,A}} \left(1 - \frac{n_{i,A_l}}{n_{i,A}}\right) \left(\frac{n_{i,A} - d_{i,A}}{n_{i,A} - 1}\right) d_{i,A}}}.$$

Elle permet, sous l'hypothèse H1, de tester si deux échantillons d'observations censurées sont des réalisations de lois (Y, δ) correspondant à une même fonction de survie S_T. Comme dans l'algorithme de forêt aléatoire initial, la sélection de coupure dans RSF suit les étapes des lignes 9 à 13 de l'algorithme CART, en remplaçant la minimisation du critère L par la maximisation du critère $|L^{RSF}|$. En effet, la valeur de $|L^{RSF}|$ est d'autant plus grande que les fonctions de survie des deux échantillons comparés sont distinctes.

Une fois un arbre $\mathcal{T}$ associé à une partition $(\mathcal{X}_k)_{k=1,\dots,K}$ construit, un estimateur de Nelson-Allen (Nelson (1969), Aalen (1976)) est utilisé dans chaque feuille terminale $\mathcal{X}_k$ pour estimer la fonction de risque cumulé interne à la feuille. Notons $t_{1,k} < \ldots < t_{H_k,k}$ les instants distincts auxquels surviennent un évènement pour une observation de $\mathcal{X}_k$, $(d_{i,k})_{i=1,\dots,H_k}$ les nombres d'évènements survenus à chaque instant, et $(n_{i,k})_{i=1,\dots,H_k}$ les nombres d'observations à risque à chaque instant. L'estimateur de Nelson-Aalen dans une feuille s'exprime alors

$$\hat{\Lambda}_k(t) = \sum_{i:\ t_{i,k} \leq t} \frac{d_{i,k}}{n_{i,k}}. \tag{1.38}$$

Chaque arbre $(\mathcal{T}_j)_{j=1,\dots,M}$ de la forêt fournit donc un estimateur de la fonction de risque cumulée conditionnelle

$$\hat{\Lambda}_j(t|X = x) = \sum_{k=1}^{K} \hat{\Lambda}_{k,j}(t) \mathbb{1}_{x \in \mathcal{X}_{k,j}}.$$

La sortie de l'algorithme RSF n'est donc pas $\hat{m}_{RF}$ comme calculé dans l'algorithme 2, mais un estimateur de la fonction de risque cumulé obtenu en faisant la moyenne des $(\hat{\Lambda}_j)_{j=1,\dots,M}$:

$$\hat{\Lambda}_{RSF}(t|X = x) = \frac{1}{M} \sum_{j=1}^{M} \hat{\Lambda}_j(t|X = x).$$

En utilisant la relation $S_T(t) = \exp(-\Lambda_T(t))$ et l'estimateur correspondant $\hat{S}_T(t) = \exp(-\hat{\Lambda}_T(t))$ (étudié dans Altshuler (1970)), on obtient également un estimateur de la fonction de survie conditionnelle $\hat{S}_{RSF}(t|X = x) = \exp(-\hat{\Lambda}_{RSF}(t|X = x))$. A la fois $\Lambda_T(\cdot|X = x)$ et $S_T(\cdot|X = x)$ caractérisent la loi de T sachant X, ainsi l'algorithme RSF répond bien au problème de régression en présence de censure.

L'algorithme RSF a été appliqué à de nombreux cas d'étude, comme dans Fantazzini & Figini (2009) où il est utilisé pour évaluer le risque de crédit des entreprises à partir de données financières. Rarement rencontré dans la littérature actuarielle, il est très populaire en biostatistique où il est utilisé pour identifier des facteurs de risque génétique (Chen & Ishwaran (2012), Schmid et al. (2016)) ou épidémiologique (Ishwaran et al.

(2014)). L'algorithme RSF est facile d'utilisation car il est implémenté dans le package R *randomForestSRC*, ainsi que dans le plus récent package R *ranger*.

Un point négatif de l'algorithme RSF est que le test *logrank* utilisé pour sélectionner les coupures en chaque noeud nécessite que l'hypothèse H1 soit vérifiée pour s'appliquer, et ainsi garantir asymptotiquement de sélectionner la coupure optimale. Dans un cas où H2 est vérifié sans que H1 ne le soit, Cui, Zhu, Zhou, & Kosorok (2017) ont mis en évidence le biais de sélection de coupure de RSF. Par ailleurs, le test *logrank* n'est pas adapté dans le cas où les fonctions de survie de deux échantillons se croisent, comme cela est souligné dans Logan et al. (2008).

1.4.4 Une seconde approche : Les Relative Risks Trees (RRT)

L'algorithme *Relative Risk Tree* (RRT) a été développé par LeBlanc & Crowley (1992). L'approche proposée repose sur la maximisation de la vraisemblance d'un arbre de régression, approximée en faisant une hypothèse de hasard proportionnel. Les forêts de *relative risk trees* ont été étudiées par Ishwaran et al. (2004), ainsi que Hothorn et al. (2004).

Pour décrire l'algorithme RRT, nous reprenons là encore les trois éléments *critère de coupure*, *critère d'arrêt* et *estimateur de feuille terminale*. Le méthode utilisée pour construire un arbre RRT consiste à maximiser la vraisemblance d'un arbre $\mathcal{T}$ associée à une partition en feuilles terminales $(\mathcal{X}_k)_{k=1,..,K}$. On note λ_k (resp. Λ_k) le taux de risque instantané (resp. taux de risque cumulé) de la loi $\mathcal{L}(T|X \in \mathcal{X}_k)$, alors la vraisemblance des données $(Y_i, \delta_i, X_i)_{i=1,...,n}$ s'exprime

$$\prod_{k=1}^{K} \prod_{i:X_i \in \mathcal{X}_k} \lambda_k(Y_i)^{\delta_i} e^{-\Lambda_k(Y_i)}. \tag{1.39}$$

En faisant l'hypothèse de hasard proportionnel $\lambda_k(t) = \lambda_0(t)\theta_k$ (i.e. $\Lambda_k = \Lambda_0\theta_k$) pour tout $k = 1, .., K$, avec λ_0 le taux de risque instantané de référence, la vraisemblance (1.39) devient

$$\prod_{k=1}^{K} \prod_{i:X_i \in \mathcal{X}_k} (\theta_k \lambda_0(Y_i))^{\delta_i} e^{-\theta_k \Lambda_0(Y_i)}. \tag{1.40}$$

L'estimateur du maximum de vraisemblance pour $(\theta_k)_{k=1,..,K}$ est alors donné par $\tilde{\theta}_k = \left(\sum_{i=1}^n \delta_i \mathbb{1}_{X_i \in \mathcal{X}_k}\right) / \left(\sum_{i=1}^n \Lambda_0(Y_i)\mathbb{1}_{X_i \in \mathcal{X}_k}\right),\ k = 1, .., K$. En remplaçant la fonction Λ_0 par

l'estimateur de Nelson-Aalen $\hat{\Lambda}_0(t) = \sum_{Y_i \leq t} \delta_i / \left(\sum_{i=1}^n \mathbb{1}_{Y_j \geq Y_i} \right)$, on obtient un estimateur

$$\hat{\theta}_k = \frac{\sum_{i=1}^n \delta_i \mathbb{1}_{X_i \in \mathcal{X}_k}}{\sum_{i=1}^n \hat{\Lambda}_0(Y_i) \mathbb{1}_{X_i \in \mathcal{X}_k}}, \ k = 1, .., K.$$

La quantité $\hat{\theta}_k$ s'interprète comme le ratio entre le nombre d'événements observés dans la feuille $\mathcal{X}_k$ et le nombre moyen d'événements qui serait attendus dans $\mathcal{X}_k$ si X n'avait pas d'influence (i.e. $\theta_k = 1$). Pour un noeud $A \subset \mathcal{X}$, notons donc
$\hat{\theta}_A = \left(\sum_{i=1}^n \delta_i \mathbb{1}_{X_i \in A} \right) / \left(\sum_{i=1}^n \hat{\Lambda}_0(Y_i) \mathbb{1}_{X_i \in A} \right)$ et définissons

$$R(A, \mathcal{D}) = 2 \sum_{i=1}^n \left(\delta_i \log \left(\frac{\delta_i}{\hat{\Lambda}_0(Y_i) \hat{\theta}_A} \right) - (\delta_i - \hat{\Lambda}_0(Y_i) \hat{\theta}_A) \right) \mathbb{1}_{X_i \in A},$$

appelé déviance du noeud A. Dans l'algorithme RRT, la coupure choisie au noeud A est celle qui maximise le critère

$$L^{RRT}(u, j, A, \mathcal{D}) = R(A) - (R(A_r) + R(A_l)),$$

avec comme dans les algorithme précédent, $A_l = A \cap \{X^j \leq u\}$ et $A_r = A \cap \{X^j > u\}$. Il y a en fait équivalence entre la minimisation de la somme des déviances $R(A_r) + R(A_l)$ et la maximisation de la vraisemblance de l'arbre (1.40), ainsi le critère L^{RTT} à une interprétation naturelle.

Dans l'article Ishwaran et al. (2004), les arbres de régression qui composent la forêt aléatoire sont donc construits en optimisant le critère L^{RRT}, et de manière classique pour un algorithme de forêt aléatoire, la segmentation d'un noeud s'arrête lorsque celui-ci contient moins de 20 observations (critère d'arrêt $minsplit = 20$).

Plusieurs estimateurs de feuille terminale ont été explorés dans la littérature conjointement à l'utilisation de RRT. Alors que LeBlanc & Crowley (1992) et Ishwaran et al. (2004) utilisent les estimateurs $\hat{\theta}_k$ des feuilles terminales pour estimer les lois conditionnelles $\mathcal{L}(T|X \in \mathcal{X}_k)$, Hothorn et al. (2004) utilisent des estimateurs Kaplan-Meier dans chacune des feuilles terminales. Dans nos applications numériques du Chapitre 4, la méthode utilisée avec RRT est la même que celle décrite pour l'algorithme RSF ci-dessus. Elle est donc très proche de celle de Hothorn et al. (2004), puisqu'elle repose sur l'utilisation de l'estimateur de Nelson-Aalen dans chaque feuille terminale.

L'algorithme RRT est la méthode implémentée dans le package R *rpart* pour traiter les données censurées à droite. Il est moins populaire que RSF, ainsi on trouve peu d'articles

où RRT est utilisé pour analyser des données. Notons toutefois que RRT apparaît dans l'article Cho & Hong (2008) pour une étude de données médicales.

Enfin le critère de coupure utilisé dans RRT s'applique à la fois sous les hypothèses H1 et H2. Une autre hypothèse importante qui est faite dans cet algorithme est l'hypothèse du hasard proportionnel.

1.4.5 Une troisième approche : L'utilisation de poids IPCW

Les poids *Inverse Probability of Censoring Weights* (IPCW) ont été introduits dans la Section 1.2.4 de ce manuscrit. Par ailleurs, la méthode de construction des arbres de régression à l'aide des poids IPCW, et particulièrement dans le cas conditionnel où seule l'hypothèse H2 est vérifiée, est l'objet du Chapitre 4 de cette thèse. Elle est donc décrite en détails dans ce chapitre. Ainsi, nous nous contentons ici de remettre cette méthode dans son contexte, et d'insister sur ses atouts et ses faiblesses.

Bien que le procédé consistant à utiliser l'inverse de la fonction de survie de la v.a.r. de censure C pour compenser le biais de censure apparaisse dans la littérature dès les années 1980 (e.g. Koul et al. (1981)), le concept de poids IPCW a été introduit par Van der Laan & Robins (2003), appliqué au cas des arbres de régression dans Molinaro et al. (2004), et enfin à la forêt aléatoire dans Hothorn, Bühlmann, et al. (2006). On peut qualifier cette méthode d'approche en deux étapes, car elle nécessite dans un premier temps de calculer les poids IPCW de chaque observation, puis dans un second temps d'utiliser un algorithme de régression en pondérant les observations avec les poids calculés précédemment. Les poids IPCW permettent donc, comme on l'a vu à la Section 1.2.4, de débiaiser un échantillon et de se ramener à un cas où les données ne seraient pas censurées. Par opposition, les algorithmes présentés dans les deux paragraphes précédents fonctionnent en une étape et traitent directement les données censurées. Une particularité de la méthode IPCW est qu'elle permet de généraliser aux données censurées l'usage de nombreux algorithmes de régression comme la machine à vecteurs de support dans Goldberg & Kosorok (2017), ou encore le gradient boosting dans Hothorn, Bühlmann, et al. (2006).

Les références aux poids IPCW sont nombreuses dans la littérature, surtout en biostatistique, et ne se limitent pas aux cas des arbres de régression ou de la forêt aléatoire. Pour ce qui concerne les applications dans le cadre des arbres de régression, on peut citer Molinaro et al. (2014) qui utilisent la méthode pour déterminer des groupes de risque à partir de facteurs génétiques, ou Lopez et al. (2016) en actuariat.

Le package R *pec* permet de calculer les poids IPCW non conditionnels à l'aide d'un estimateur de Kaplan-Meier, ou conditionnels en utilisant le modèle de Cox (1972).

Un avantage de la méthode IPCW par rapport aux précédentes méthodes est qu'elle généralise l'usage des algorithmes de régression existant au cas des données censurées. Ainsi, dans le cas limite où dans un échantillon (Y_i, δ_i, X_i) aucune observation n'est censurée ($\delta_i = 1, \forall i = 1, .., n$), les poids IPCW des observations valent $1/n$ et l'on est ramené au cas de régression en absence de censure. De plus, dans son approche conditionnelle, le principe IPCW nécessite seulement que l'hypothèse H2 soit satisfaite pour s'appliquer ; aucune autre hypothèse n'est nécessaire. Enfin, l'utilisation des poids IPCW conjointement à la méthode CART nous permet, dans notre cas d'application du Chapitre 4, d'optimiser la fonction de régression de manière à prédire $E[\phi(T)|X]$ pour une fonction ϕ donnée. Cela différencie la méthode IPCW des algorithmes RSF et RRT dont le but est d'estimer la fonction de survie conditionnelle $S_T(\cdot|X)$, sans possibilité d'optimiser un critère donné par une fonction ϕ particulière. Néanmoins, la méthode IPCW, en particulier dans le cas conditionnel, présente parfois des problèmes d'instabilité numérique. Ces problèmes, ainsi que les solutions possibles, sont discutés au Chapitre 4.

1.5 Les contributions de notre travail

Dans cette thèse, nos contributions sont organisées en trois chapitres qui peuvent être lus indépendamment les uns des autres. Le Chapitre 2 présente le problème, de nature assurantielle, que nous avons étudié et apporte des éléments de contexte. Au Chapitre 3, nous étudions une méthode, basée sur la notion de copule, qui permet d'estimer la dépendance entre deux durées successives conditionnellement à des variables explicatives. Le contenu de ce chapitre constitue un article qui a été soumis à la revue *Journal of Multivariate Analysis*. Enfin, le Chapitre 4 est un travail méthodologique qui porte sur l'adaptation de l'algorithme de forêt aléatoire au cas où la variable à prédire est une durée censurée à droite. Cette contribution a été soumise à la revue *Journal of the American Statistical Association*.

Chapitre 2

Dans ce chapitre, nous commençons par décrire le fonctionnement de l'assurance santé en France, ainsi que les aspects liés à la souscription et à la résiliation de contrats d'assurance complémentaire santé. Puis nous présentons la problématique de l'entreprise de courtage

d'assurances santé avec laquelle nous avons collaboré durant cette thèse. Il s'agit d'estimer la valeur client d'un prospect susceptible de souscrire un contrat d'assurance complémentaire santé. La dernière partie du chapitre est consacrée à la description et à l'exploration des données recueillies par le courtier. Une attention particulière est portée sur l'étude statistique de la durée de résiliation et de la durée avant effet des contrats d'assurance commercialisés par le courtier.

Chapitre 3

Le sujet abordé dans ce chapitre est motivé par le problème concret de la mesure de la dépendance entre la durée avant effet et la durée de résiliation pour un contrat d'assurance complémentaire santé. Nous proposons une méthode d'estimation semi-paramétrique de la copule conditionnelle liant ces deux durées, qui ont la particularité d'être des durées successives et d'être censurées à droite. Dans notre application, le vecteur de variables explicatives est unidimensionnel et correspond à l'âge de l'assuré. L'estimateur que nous utilisons est celui qui apparaît à l'équation (1.36). Il repose sur l'utilisation des poids IPCW pour compenser le biais induit par la censure. De plus, il exploite la situation de durées successives, qui fait que la censure bivariée portant sur les deux durées se réduit à une variable unidimensionnelle.

Sous certaines hypothèses (identifiabilité du modèle, régularité de la log-vraisemblance, intégrabilité de la log-vraisemblance aux extrémités du carré unité, vitesse de convergence du paramètre de fenêtre utilisé pour la fonction noyau), nous démontrons la convergence uniforme avec une vitesse optimale de l'estimateur proposé, vers la fonction $x \to \theta(x)$ qui correspond au vrai paramètre de copule. L'application numérique sur les données du courtier met en évidence une dépendance négative entre les deux durées, et une évolution non linéaire de cette dépendance avec l'âge : dépendance moins importante chez les 30-50 ans et plus importante chez les moins de 30 ans et les plus de 50 ans. Nous observons également que la dépendance est plus fortement négative pour les contrats d'assurance haut de gamme, en particulier dans les âges élevés.

Chapitre 4

Le dernier chapitre de notre travail est consacré à l'utilisation de l'algorithme de forêt aléatoire dans le but d'estimer la valeur client d'un prospect (voir Section 2.3.1 pour la présentation détaillée du cas d'application). Le problème mathématique sous-jacent est

celui de l'estimation de la fonction de régression $f(x) = E[\phi(T)|X = x]$ lorsque ϕ est une fonction connue, et T est une variable de durée censurée à droite. La méthode que nous étudions, évoquée dans la Section 1.4.5, utilise les poids IPCW pour supprimer le biais d'échantillon induit par la censure.

À travers des expériences réalisées sur des données simulées et sur des données réelles (les données du courtier), nous comparons les performances de notre approche avec celles de méthodes concurrentes (voir Sections 1.4.3 et 1.4.4), en prenant en compte l'influence du taux de censure, de la dépendance entre la variable de censure et les variables explicatives X, du mode d'estimation des poids IPCW (estimés conditionnellement à X ou non), ainsi que des paramètres utilisés pour la forêt aléatoire. Différentes manières d'exploiter les poids IPCW des observations dans l'algorithme de forêt aléatoire sont également comparées. Nous mettons en évidence l'intérêt d'estimer les poids IPCW conditionnellement aux covariables, ainsi que le fort impact du taux de censure sur le type d'algorithme réalisant les meilleures performances. Nous apportons également des réponses pratiques au problème de l'explosion des poids IPCW. Dans l'application à la prédiction de la valeur client, nous montrons que notre méthode améliore sensiblement la précision du modèle, mesurée avec l'erreur quadratique moyenne. Enfin, le travail de ce chapitre a mené au développement du package R *sword*, qui permet de tester les différents algorithmes utilisés dans notre étude.

Chapitre 2

Contexte de l'assurance complémentaire santé en France, et présentation de notre cas d'étude

2.1 L'assurance santé en France

2.1.1 La Sécurité sociale

Introduction

La Sécurité sociale a été créée en France en 1945, au lendemain de la seconde guerre mondiale, dans le but d'unifier les différentes formes d'assurance sociale qui existaient alors. Financé par une cotisation interprofessionnelle à taux unique, l'organisme devait garantir à tous, et notamment aux plus vulnérables (enfants, personnes âgées, mères au foyer) la protection de la santé, la sécurité matérielle, le repos et les loisirs (Gibaud (1986)). La structure était à l'origine organisée en quatre branches : maladie, famille, recouvrement, vieillesse ; et une cinquième branche portant sur la lutte contre la dépendance à été créée en 2004. Notre travail, qui porte sur l'analyse de données relatives à des contrats de complémentaire santé, concerne la branche maladie. Aussi, c'est à cette branche de la Sécurité sociale que nous nous intéressons dans la suite de cette section.

Les différents régimes de Sécurité sociale

L'affiliation à la Sécurité sociale est obligatoire pour toute personne résidant en France. Il existe plusieurs régimes de Sécurité sociale, dont les principaux sont : le régime général qui

concerne 92% de la population française en 2017 d'après le rapport de la Sécurité Sociale (Sécu (2018)), le régime social des indépendants (RSI, 4.2% de la population), et le régime agricole (1.8% de la population). Dans les départements du Bas-Rhin, du Haut-Rhin et de la Moselle, le régime général est géré indépendamment du régime général national. Il est appelé régime local Alsace-Moselle. D'autres régimes, dit spéciaux, existent également. Par exemple, ils peuvent concerner les salariés de grandes entreprises publiques comme la SNCF ou EDF.

L'Assurance maladie

La filiale de la Sécurité sociale consacrée à la maladie, appelée Caisse nationale de l'assurance maladie des travailleurs salariés (Cnam) ou plus simplement Assurance maladie, assure les dépenses de santé des résidents français. Chaque acte médical, comme la consultation d'un médecin généraliste (conventionné de secteur 1, de secteur 2, ou non conventionné), d'un dentiste, ou l'achat de lunettes de vue, est codifié selon une nomenclature et associé à une base de remboursement correspondante. Le montant remboursé à l'assuré, pour un acte médical donné, est alors égal à la base de remboursement multipliée par un taux de remboursement, qui dépend du type d'acte médical et du régime du Sécurité sociale de l'assuré. Depuis 2005 et pour certains actes médicaux seulement, une participation forfaitaire de un euro est déduite du montant remboursé. Prenons l'exemple d'une consultation chez un médecin généraliste conventionné de secteur 1. La base de remboursement est de 25 euros, le taux de remboursement de 70% et la participation forfaitaire s'applique, ainsi le montant remboursé par la sécurité sociale pour un tel acte est de $25 \times 0.7 - 1 = 16.5$ euros.

On remarque donc que dans de nombreux cas, tel que celui décrit ci-dessus, le remboursement de la Cnam est partiel et ne correspond pas à 100% de la base de remboursement. De plus, la prise en charge de l'Assurance maladie est forfaitaire et ne dépend pas des frais réels payés par l'assuré, qui sont supérieurs à la base de remboursement si des dépassements d'honoraire sont appliqués par le professionnel de santé. En général, une part conséquente des frais réels reste donc à la charge de l'assuré. En cas de dépenses de santé importantes à cause par exemple d'une maladie grave, d'un accident ou d'une hospitalisation, les sommes à la charge de l'assuré peuvent être très importantes, et un second mécanisme de protection est donc indispensable. Le système de l'assurance complémentaire santé, que nous présentons dans la section suivante, est une réponse à ce besoin.

2.1.2 L'assurance complémentaire santé

Le fonctionnement de l'assurance complémentaire santé en France

L'assurance complémentaire santé est un contrat d'assurance qui permet d'obtenir un remboursement des frais de santé, complémentaire à celui de l'Assurance maladie, moyennant le paiement d'une cotisation. Seule la part des dépenses non remboursée par l'Assurance maladie est éligible au remboursement pas la complémentaire santé, de sorte que la somme des remboursements ne peut excéder le coût des soins. De plus, le remboursement de la complémentaire santé peut être total ou partiel, en fonction du montant restant à charge, du type d'acte médical et des clauses du contrat. Dans l'exemple précédent de la consultation d'un médecin généraliste, les 30% de 25 euros (que l'on appelle souvent ticket modérateur), soit 7.5 euros, sont en général remboursés par la complémentaire santé, et la participation forfaitaire de un euro reste à la charge du patient.

Les organismes de complémentaire santé

Les contrats de complémentaire santé peuvent être commercialisés par trois types d'organismes: les mutuelles, les instituts de prévoyance, et les compagnies d'assurance. Les mutuelles, comme les instituts de prévoyance, sont des sociétés de personnes à but non lucratif, dans lesquelles les adhérents (ou sociétaires) ont une forte représentation au conseil d'administration. À l'opposé, les compagnies d'assurance sont des sociétés de capitaux à but lucratif, dans lesquelles le conseil d'administration est nommé par les actionnaires. Chaque type d'organisme obéit à des règles spécifiques : les mutuelles sont soumises au code de la mutualité, les instituts de prévoyance au code de la sécurité sociale, et les compagnie d'assurance au code des assurances.

L'apparition des organismes d'assurance santé précède la création de la Sécurité sociale. En effet, les premières pratiques mutualistes sont apparues au 19e siècle, avant que la Charte de la Mutualité (1898) ne définisse plus précisément le rôle et le mode de fonctionnement des mutuelles (Dreyfus (2011)). Comme indiqué dans Gibaud (1986, 2008), la création de la Sécurité sociale en 1945 marque un tournant dans l'organisation de la solidarité, avec le passage d'un modèle d'adhésion individuel et facultatif à un modèle collectif et obligatoire. À partir de 1945, le rôle des mutuelles devient complémentaire à celui de la Sécurité sociale.

Comparaison du système français avec les systèmes des autres pays de l'OCDE

D'après l'enquête de l'OCDE de Paris et al. (2010), tous les pays de l'OCDE disposent d'un système de financement des dépenses de santé qui prend en charge plus de 50% des dépenses totales. On trouve deux principaux modèles de financement parmi les pays de l'OCDE : un système d'assurance maladie à affiliation obligatoire où les prestations sont versées en contrepartie de cotisations, qui correspond au modèle français ou allemand ; et un système de service national de santé financé par l'impôt, que l'on trouve par exemple au Royaume-uni, en Irlande, ou au Canada. Dans les pays qui ont opté pour une assurance maladie obligatoire, le système d'assurance est parfois privé. C'est le cas par exemple en Suisse, aux Pays-Bas ou au États-Unis depuis la réforme *Obamacare* de 2014. Dans ce cas de figure, l'État contrôle et régule les entreprises privées pour garantir la solidarité du système de protection.

Dans la plupart des pays de l'OCDE, des systèmes complémentaires d'assurance privée existent. Cependant, ils sont plus ou moins développés selon les pays.

Aux États-Unis, la part des dépenses courantes de santé dans le PIB atteint environ 17% en 2016. Ceci constitue le taux le plus élevé parmi les pays de l'OCDE. En France, ce taux est de 12%, soit l'un des taux les plus élevé de l'Union européenne. Notons que l'agrégat utilisé pour comparer les dépenses de santé des différents pays est la dépense courante de santé au sens international (DSCi). Celle-ci inclut, contrairement à la CSBM qui est définie dans la section suivante, les dépenses de soins pour les maladies de longue durée, les subventions de l'état au système de soin, le financement de la prévention institutionnelle, ainsi que les coûts de gestion.

2.1.3 Le marché de l'assurance santé en France

D'après la publication de la Direction de la recherche, des études, de l'évaluation et des statistiques (Drees (2018)), la consommation de soins et biens médicaux (CSBM) en France s'élève en 2017 à 199.3 milliards d'euros, soit 8.7% du PIB français. Les données disponibles depuis les années 1950 (Drees (2017)) montrent que la CSBM n'a cessé d'augmenter depuis cette période, progressant plus vite que le PIB (la CSBM représentait environ 1% du PIB en 1952). Le marché des dépenses de santé est donc soutenu par une croissance régulière, qui a perduré ces dernières années : +21% entre 2008 et 2017. L'augmentation des dépenses entre 2008 et 2017 s'explique notamment par l'allongement de la durée de vie, les progrès médicaux, et l'évolution des pratiques de consommation

médicale. En 2017, la CSBM est composée à 46.6% par les soins hospitaliers, à 26.8% par les soins de ville (consultation de médecin généraliste, dentiste, etc.), à 16.3% par les médicaments, à 7.8% par d'autres biens médicaux (optique, prothèse par exemple), et à 2.5% par des frais liés au transport sanitaire.

Durant les dix dernières années, la part de la CSBM financée par les ménages a légèrement baissé (baisse régulière de 9.4% en 2008 à 7.5% en 2017), contre-balancée par une hausse des remboursements de l'Assurance maladie (76.8% en 2008 et 77.8% en 2017) et des organismes complémentaires (12.3% en 2008, 13.2% en 2017). Le reste des dépenses (environ 1%) est financé par l'état au titre de la CMU-C (Couverture maladie universelle complémentaire). Le faible reste à charge pour les ménages français s'explique par le taux élevé de personnes bénéficiant d'une complémentaire santé (96% de la population en 2010), qui a beaucoup progressé depuis les années 1950 (31% de personnes couvertes en 1960, 69% en 1980, 96% en 2010).

Parmi les organismes complémentaires, la part des remboursements attribuée aux mutuelles est en baisse (57.9% en 2008 et 50.8% en 2017), au profit des sociétés d'assurance (24.0% en 2008, 29.3% en 2017) et des instituts de prévoyance (18.2% en 2008, 19.9% en 2017). Cela s'explique en partie par l'Accord national interprofessionnel (ANI) de 2013, qui a plutôt profité aux entreprises d'assurance au détriment des mutuelles.

2.2 La souscription et la résiliation de contrats d'assurance santé

2.2.1 La souscription

Les types de contrats

Les contrats d'assurance complémentaire santé sont soit collectifs, soit individuels. Les contrats collectifs sont le plus souvent souscrits par les entreprises, au bénéfice de leurs salariés. En effet, depuis janvier 2016 et l'entrée en application de l'Accord national interprofessionnel (ANI), toutes les entreprises privées ont l'obligation de proposer à leurs salariés d'adhérer à un contrat collectif de complémentaire santé. Cette adhésion est en principe obligatoire, mais divers cas de dispense existent (cas où le salarié bénéficie déjà d'une complémentaire santé par exemple). Si le contrat le prévoit, les ayants droit de l'assuré peuvent également bénéficier de la protection d'entreprise. En cas d'adhésion à un contrat collectif d'entreprise, l'employeur participe au paiement de la cotisation à hauteur

d'au moins 50%. Parfois, en complément d'un contrat à adhésion obligatoire, l'employeur propose un contrat "surcomplémentaire" à adhésion facultative, qui comprend des options de garanties supplémentaires.

Les contrats d'assurance complémentaire santé peuvent aussi être souscrits à titre individuel. Ce type de contrats s'adresse principalement aux personnes qui ne sont pas bénéficiaires (y compris en tant qu'ayant droit) d'un contrat collectif d'entreprise, c'est à dire les retraités, les fonctionnaires, les travailleurs indépendants, les chômeurs ou les étudiants. Les ayants droit du titulaire d'un contrat individuel peuvent également bénéficier de la protection en général. Aussi, certains salariés du privé qui souhaitent souscrire un contrat surcomplémentaire peuvent se tourner vers un contrat individuel.

Le rapport de la Drees (2018) indique qu'en 2016 et 2017, environ la moitié des contrats d'assurance complémentaire santé souscrits sont des contrats collectifs, et l'autre moitié des contrats individuels.

Les canaux de distribution

Selon qu'un contrat de complémentaire santé est individuel ou collectif, le mode de souscription est différent. Pour un contrat collectif, le contrat est conclu entre l'organisme complémentaire et l'employeur, et les démarches d'adhésion sont simplifiées pour l'employé. Le cas que nous étudions dans cette thèse concerne la distribution de contrats d'assurance complémentaire santé individuels et c'est donc ce sujet que nous approfondissons dans la suite de cette section.

Les organismes complémentaires disposent en général d'un réseau de distribution exclusif, chargé de vendre les contrats propres à l'entité. Ce réseau est formé des agences de proximité qui emploient des salariés de l'organisation, de la vente en direct par internet ou par téléphone, des agents généraux d'assurance qui sont des travailleurs indépendants mandatés par l'organisme complémentaire pour faire l'intermédiaire entre les clients et l'organisme, ou encore des filiales de gestion privée. D'une manière générale, on trouve davantage d'agences physiques chez les mutuelles, alors que les entreprises d'assurance ont souvent privilégié le système basé sur les agents généraux.

Les contrats sont également distribués par le biais d'un réseau non exclusif, constitué par les entreprises de courtage, les conseillers indépendants, les comparateurs d'assurance, etc. On parle de réseau non exclusif car ces intermédiaires peuvent vendre les contrats de plusieurs entités. Selon les cas, ils peuvent prendre en charge la gestion du contrat d'assurance (e.g. courtier gestionnaire), ou non.

Lorsqu'un intermédiaire (agent général d'assurance, courtier, conseiller indépendant, comparateur, etc.) intervient dans la vente d'un contrat d'assurance, il perçoit une commission de la part de l'organisme complémentaire. Ce mode de rémunération est classique en assurance.

La directive sur la distribution des assurances (DDA) votée au parlement européen en 2016 et entrée en vigueur en octobre 2018 en France a pour buts principaux de renforcer la protection des souscripteurs d'assurance et d'harmoniser les règles applicables aux distributeurs de produits d'assurance. Tous les types d'assurances (santé, Iard, épargne, etc.) ainsi que tous les distributeurs d'assurances (y compris le réseau exclusif des organismes) sont concernés. Le texte comporte cinq axes de conformité en matière de distribution d'assurances : la capacité professionnelle, le devoir de conseil, l'information et la transparence, la rémunération et les conflits d'intérêt, ainsi que la gouvernance et la surveillance des produits (ACPR (2018)). Comme tous les autres acteurs de l'assurance, les organismes de complémentaire santé et les intermédiaires distributeurs doivent désormais respecter cette réglementation.

À la Section 2.3, nous présentons notre cas d'étude qui porte sur l'optimisation du ciblage client pour le courtage de produits d'assurance complémentaire santé. La composante résiliation, que nous présentons dans la section suivante, est l'objet précis de notre travail.

2.2.2 La résiliation

Les aspects réglementaires

La résiliation d'un contrat d'assurance complémentaire santé est régie par le Code des assurances ou le Code de la mutualité selon l'organisme complémentaire concerné. Nous n'abordons ici que les règles relatives aux contrats individuels.

La plupart des contrats d'assurance complémentaire santé prévoient une période d'engagement d'un an. Au-delà de la première année, le contrat est renouvelé tacitement d'année en année, sauf en cas de refus de l'une ou l'autre des parties. L'assuré a donc la possibilité de résilier son contrat complémentaire à la date de l'échéance annuelle du contrat, sans motif ni frais de résiliation. Pour cela, il doit signaler à l'organisme complémentaire son intention au moins deux mois avant l'échéance du contrat. À cet effet, la loi Chatel de 2005 oblige l'assureur à transmettre à l'assuré un avis d'échéance annuel qui mentionne la date limite jusqu'à laquelle l'assuré peut faire sa demande de

résiliation (Mallet-Bricout (2005)). L'avis d'échéance doit être transmis au plus tard 15 jours avant la date limite de résiliation. Dans certains cas particuliers, l'assuré a la possibilité de demander une résiliation de contrat hors période de renouvellement. Il s'agit par exemple d'un changement de situation (changement de domicile, de situation matrimoniale, de profession, départ en retraite, etc.) modifiant les risques de l'assuré, de l'adhésion à contrat collectif d'entreprise, ou d'une résiliation en réponse à une augmentation injustifiée de la prime annuelle d'assurance (sous certaines conditions seulement ; ce motif de résiliation ne s'applique pas aux mutuelles). Dans un tel cas de figure, aucun délai ne s'applique et la résiliation prend effet dans le mois suivant la demande de l'assuré. L'organisme complémentaire rembourse alors la part des cotisations versées par l'assuré pour la période suivant la résiliation. Mentionnons également que dans le cadre de la vente à distance par internet ou par téléphone, l'assuré dispose d'un délais de rétractation de 14 jours pendant lequel il peut renoncer à son contrat sans justification ni frais.

L'organisme complémentaire dispose également d'un droit de résiliation du contrat d'assurance complémentaire, sans motif ni indemnisation, à l'échéance du contrat. Il doit pour cela respecter un délai de deux mois de préavis, avant la date d'échéance annuelle du contrat. En cas non paiement de la cotisation par l'assuré, ou de fausse déclaration de l'assuré, l'organisme complémentaire a le droit de résilier le contrat à tout moment. Précisons toutefois que l'assureur ne peut pas exclure un assuré ou majorer les cotisations d'un assuré au cas par cas, mais seulement modifier annuellement les cotisations de l'ensemble des assurés du contrat.

Les règles de résiliation des contrats de complémentaire santé sont amenées à évoluer dans les prochaines années car une proposition de loi visant à aligner les règles de résiliation des contrats d'assurance santé avec celles s'appliquant aux contrats d'assurance habitation et auto (régis par la loi Hamon depuis 2014) a été adoptée par le parlement français en mars 2019. Cette loi, qui devrait entrer en vigueur avant décembre 2020, prévoit la possibilité pour les assurés de résilier leur contrat de complémentaire santé à tout moment et sans frais une fois passée la première année d'engagement.

La modélisation de la résiliation de contrats d'assurance

Anticiper, à l'échelle d'un contrat individuel ou d'un portefeuille de contrats, le comportement de résiliation des assurés est important en assurance. En assurance vie, on parle de rachat de contrat lorsque l'assuré décide de clôturer son contrat et de récupérer l'argent épargné. Ce sujet est historiquement la première application de l'étude des résiliations

en assurance, et l'article de Cummins (1973) donne plusieurs références sur les premiers modèles de rachat utilisés en assurance. Aujourd'hui, la modélisation des rachats est toujours centrale car le risque de rachat en assurance vie constitue l'un des modules de risque dans la formule standard de Solvabilité II. La modélisation des résiliations est également importante en assurance non-vie, par exemple pour optimiser une stratégie de rétention de clients ou bien pour évaluer la valeur / rentabilité d'un portefeuille d'assurance non-vie.

Du point de vue statistique, on rencontre principalement deux approches dans la littérature pour modéliser les résiliations de contrats en assurance. D'un côté, les modèles où l'on cherche à prédire la probabilité de résiliation future sur une période donnée, par exemple durant l'année qui suit. Et d'un autre côté, les modèles de résiliation basés sur l'utilisation des modèles de durée, où l'on cherche à prédire la fonction de survie des contrats, sans fixer d'horizon préalable. Une autre distinction importante est la différence entre la modélisation des résiliations à l'échelle individuelle qui vise à prédire la probabilité qu'un contrat particulier soit résilié, et la modélisation à l'échelle agrégée où l'on s'intéresse au taux de résiliation global sur un portefeuille de contrats.

Dans le cas où on modélise la probabilité de résiliation sur une période donnée, les modèles utilisés sont des modèles de régression (e.g. modèle linéaire généralisé (GLM)) ou de classification (e.g. régression logistique). Ainsi, Renshaw & Haberman (1986) et Outreville (1990) cherchent à déterminer quels sont les principaux facteurs de rachat en assurance vie en utilisant des GLM sur des données agrégées, en s'intéressant à la prédiction du taux de rachat sur une année. Le même sujet est traité par Kim (2005), qui utilise une fonction de lien logit et des variables économiques comme variables explicatives. A l'échelle individuelle, Milhaud et al. (2011) utilisent l'algorithme CART et la régression logistique pour segmenter un portefeuille d'assurance vie en plusieurs classes de risque de rachat. En assurance auto, Dutang (2012) étudie les déterminants de la résiliation de contrats avec une approche par élasticité des prix. Également en assurance non-vie, Günther et al. (2014) prédisent la probabilité mensuelle de résiliation à partir de données de contrats d'assurance auto, habitation et santé, en utilisant un modèle logistique. Lorsqu'une approche par discrétisation du temps est utilisée, comme dans les exemples ci-dessus, les modèles sont généralement calibrés plusieurs fois, sur des pas de temps successifs. De plus les variables explicatives, comme par exemple le niveau de chômage, sont susceptibles d'évoluer dans le temps. On parle alors de variables explicatives qui varient dans le temps.

Une approche naturelle pour prendre en compte la temporalité du problème, ainsi que

l'éventuelle présence de censure dans les données (présente ou non selon la manière dont les données ont été recueillies), est d'utiliser les modèles de durée. Ce type de méthodes a été utilisé par Van den Poel & Lariviere (2004a,b) dans le but d'analyser l'impact de campagnes de rétention de clients sur les résiliations. Dans le contexte du CRM (customer relationship management) et de la rétention client, Guillén et al. (2012) se sont intéressés aux résiliations des assurés possédant plusieurs contrats dans une même compagnie ; évaluation du risque de résiliation pour les contrats restants lorsqu'un assuré résilie l'un des contrats qu'il possède avec la compagnie. Pour cela, ils ont utilisé un modèle de Cox avec prise en compte des variables explicatives qui varient dans le temps, et avec des coefficients de régression également variables dans le temps. Aussi, Milhaud & Dutang (2018) ont utilisé les modèles à risques compétitifs en assurance vie pour évaluer le risque lié aux rachats.

À notre connaissance, il existe peu de travaux sur la modélisation des résiliations de contrat d'assurance santé. Notons néanmoins l'article de Günther et al. (2014) déjà mentionné ci-dessus ainsi que le mémoire d'actuariat de Labit Hardy (2012) qui traite spécifiquement de ce sujet.

2.3 Présentation de notre cas d'étude

2.3.1 Le contexte

Le courtage de contrats d'assurance complémentaire santé

L'ensemble des applications que nous réalisons dans cette thèse, aux Chapitres 3 et 4, concernent la problématique suivante. Nous considérons un courtier dont l'activité est de vendre des contrats d'assurance complémentaire santé individuels pour le compte de différents organismes complémentaires (mutuelles, compagnies d'assurance) que nous appelons les partenaires assureurs du courtier. Pour toucher les clients désireux de souscrire une complémentaire santé, ou d'en changer, le courtier acquiert des informations sur des clients potentiels appelés prospects. Un prospect est par exemple un particulier qui a utilisé un comparateur d'assurance santé en ligne, ou qui a signalé sur le site du courtier sa volonté d'être appelé par un conseiller. L'acquisition de prospects est donc possible par plusieurs canaux d'acquisition, qui sont le plus souvent payants (e.g. comparateur d'assurance, bannière de publicité en ligne, référencement payant sur les moteur de recherche...), mais parfois gratuits (e.g. site internet du courtier consulté sans

référencement payant). Les conseillers clientèle qui travaillent pour le courtier contactent ensuite les prospects par téléphone pour leur proposer une complémentaire santé adaptée à leurs besoins.

Le commissionnement

Pour chaque contrat vendu, le partenaire assureur verse au courtier une commission qui constitue la rémunération du courtier. Le mode de calcul de la commission versée par l'organisme complémentaire peut varier d'un partenaire à un autre, toutefois cette variabilité est faible et dans la suite, on retiendra la formule de calcul *type* suivante :

$$Com_{tot} = Com_{prec} + Com_{lin},$$

où Com_{tot} est la commission totale, Com_{prec} est appelé le précompte, et Com_{lin} le linéaire. Le précompte est la commission versée au courtier par le partenaire assureur à la signature du contrat, qui correspond à 50% de la prime d'assurance annuelle du contrat. En cas de résiliation du contrat par l'assuré durant la première année, une partie du précompte perçu par le courtier est reversée au partenaire. Cette part est calculée au prorata temporis des primes perçues par le partenaire pour la première année d'assurance (e.g. $4/12 \approx 33\%$ de précompte reversé si l'assuré a payé 8 mois de cotisation). Le linéaire correspond à la commission versée au courtier après la première période de 12 mois ; il est égal à 10% des primes d'assurance encaissées par le partenaire à partir du 13e mois. Le montant de la commission totale dépend donc de la prime d'assurance annuelle, ainsi que de la durée pendant laquelle l'assuré conserve le contrat, c'est à dire la durée de résiliation du contrat.

Le scoring de prospects

Avec le mécanisme de rémunération décrit dans la section précédente, on comprend qu'il est important pour le courtier de conclure un maximum de contrats, mais également de conclure des contrats qui rapportent une commission élevée. C'est dans cette optique que le courtier avec lequel nous avons travaillé attribue à chaque prospect, dès son acquisition, un score de rentabilité qui permet de classer les prospects par ordre de priorité. Sur un marché concurrentiel, il est en effet important d'identifier rapidement les clients les plus rentables, pour maximiser les chances de conclure un contrat avec eux. Le score calculé correspond à un estimateur de la valeur client du prospect, c'est à dire à un estimateur de l'espérance des commissions probables générées par le prospect et actualisées à la date

d'acquisition. En notant avec un chapeau les grandeur estimées, l'estimateur de la valeur client se décompose

$$\widehat{value} = \widehat{p}_{sub} \cdot \widehat{pr} \cdot \widehat{f}_{ew} \cdot \widehat{f}_{ch},$$

avec p_{sub} la probabilité que le prospect souscrive un contrat avec le courtier (taux de transformation), pr la prime d'assurance du prospect en cas de souscription, f_{ew} la probabilité que le contrat ne soit pas sans effet en cas de signature (*early withdrawal*), et f_{ch} le facteur de résiliation (*churn factor*) exprimé en unité de prime, qui tient compte de la durée de résiliation du contrat et de l'actualisation des flux de commission. À propos de f_{ew}, on a en effet vu à la Section 2.2.2 qu'en cas de vente à distance l'assuré dispose d'un délais de rétractation de 14 jours. De plus, il arrive que l'assuré ne paye jamais la première cotisation de son contrat d'assurance, ce qui correspond également à un contrat sans effet.

Les quatre facteurs qui apparaissent dans le score sont estimés séparément. L'estimation du taux de transformation est complexe car elle est dynamique, fortement impactée par les offres concurrentes (e.g. promotion temporaire d'un concurrent sur le marché) et le canal d'acquisition du prospect. La prime dépend du profil du prospect et de ses besoins en termes de complémentaire santé. On peut la prédire avec une bonne précision. Le taux de rétractation est en moyenne d'environ 20%, et on distingue des variations en fonction du contrat souscrit et du profil de l'assuré. Dans notre travail, nous nous focalisons sur l'étude du facteur de résiliation et donc sur la modélisation de la durée de résiliation des contrats. En notant T la durée de résiliation d'un contrat, on a $f_{ch} = \phi_{ch}(T)$ avec ϕ_{ch} une fonction déterministe représentée à la Fig. 2.1. Les données dont nous disposons pour mener notre étude sont présentées dans les sections suivantes.

2.3.2 Les données

Le périmètre de l'étude

Nous étudions la base des données recueillies par le courtier depuis 2009. Le courtage de contrats d'assurance complémentaire santé constitue l'activité principale du courtier, et nous ne considérons dans l'étude que les données relatives à l'assurance santé. De plus, la date de début d'observation choisie est le 1er janvier 2010 et la date de fin d'observation le 15 janvier 2016. Après le traitement de la base visant à sélectionner les observations à prendre en compte dans l'étude, la base de données compte 229 245 observations. La date

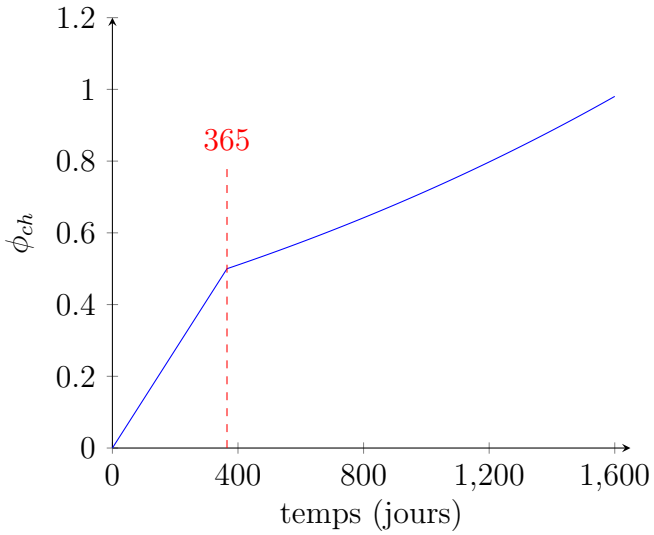

Fig. 2.1: Le facteur de résiliation en fonction de la durée de résiliation T.

Le taux de commissionnement se décompose en deux parties : le précompte égal à 50% de la prime pendant la première année, et le linéaire égal à 10% de la prime (éventuellement revalorisée) pour les années suivantes. Les flux sont actualisés avec un taux d'intérêt annuel de 8%.

d'effet d'un contrat correspond à la date à partir de laquelle un contrat préalablement signé devient actif (et l'assuré devient couvert). Le nombre annuel de prises d'effet de contrats, enregistrées par le courtier entre 2010 et 2016, est représenté sur la la Fig. 2.2. On observe sur ce graphique que l'activité du courtier est en forte croissance depuis 2010, malgré un léger plafonnement en 2015. Notons que du fait de l'arrêt de l'observation le 15/01/2016, seule une partie des contrats signés en 2016 sont représentés ici ce qui explique le faible nombre de prises d'effet en 2016.

Les caractéristiques du portefeuille de contrats d'assurance : description des variables explicatives disponibles

Lors de l'acquisition d'un prospect, on dispose d'un nombre réduit d'informations concernant le client potentiel. Nous avons résumé ces informations en six variables, qui seront les critères sur lesquels nous baserons l'estimation de la valeur client (plus spécifiquement du facteur de résiliation). Il s'agit du nombre d'enfants qui seraient bénéficiaires du contrat en plus de l'assuré, de l'âge de l'assuré lors de l'acquisition du prospect, du régime de Sécurité sociale du prospect, de la gamme (niveau de garantie) de produit d'assurance désirée par le prospect, de sa zone géographique ainsi que de son état civil. Chacune de ces variables est qualitative et présente plusieurs modalités. Les différentes modalités

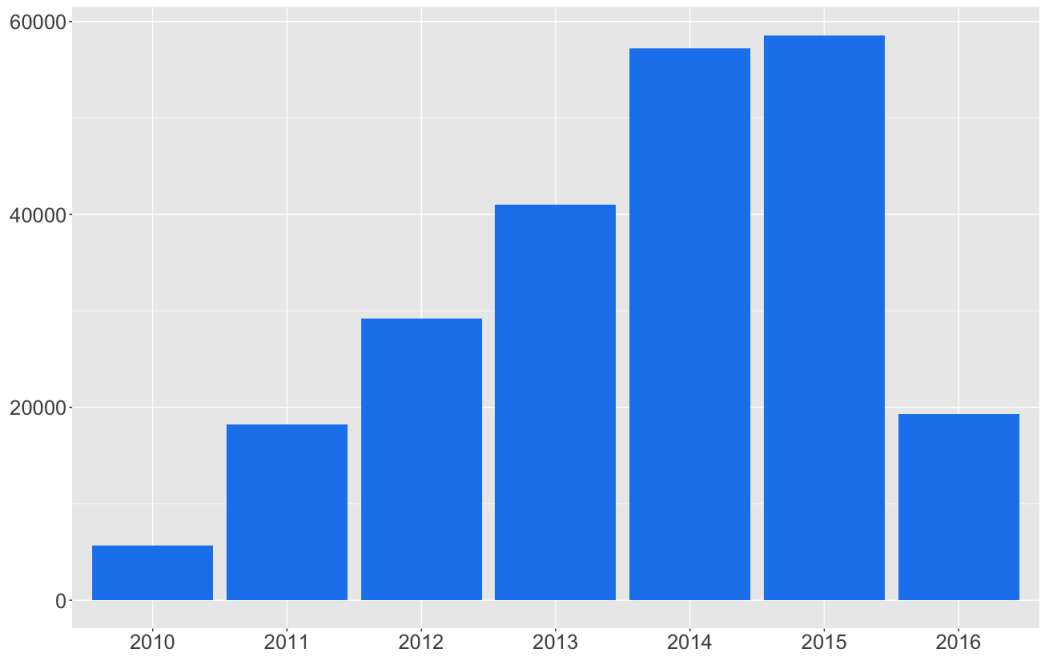

Fig. 2.2: Nombre de prises d'effet de contrats par an entre 2010 et 2016.

des variables ainsi que les effectifs de chaque modalité sont représentés sur la Fig. 2.3. La variable *Nombre d'adultes assurés* (qui vaut 1 pour une personne seule et 2 pour un couple) est également disponible dans les données, toutefois elle n'est pas bien renseignée, c'est pourquoi nous ne l'utilisons pas comme variable explicative dans les modèles de prédiction et qu'elle n'apparaît pas dans les graphiques présentés dans cette section. On peut toutefois estimer que 72% des contrats du portefeuille sont associés à une personne seule. Ces statistiques montrent que les prospects sont principalement des adultes seuls et sans enfant. Selon leurs âges, ces personnes sont le plus souvent salariées, au chômage ou retraitées. La légère sur-représentation de la tranche d'âge des 60-69 ans s'explique par l'entrée en retraite et la perte du contrat complémentaire santé collectif d'entreprise, compensée par la souscription d'un contrat individuel. De façon surprenante en regard de nos remarques Section 2.2.1, une part très importante des adhérants sont salariés (52.8% du nombre total d'adhérents, dont 48.0% sont au régime général, 1.6% au régime Alsace-Moselle, et 3.5% sont fonctionnaires).

La Fig. 2.4 permet de visualiser l'évolution de la typologie des contrats conclus en fonction de l'année d'effet du contrat. On peut noter la très forte baisse du nombre de contrats individuels souscrits par les salariés en 2016, par rapport aux années précédentes. Cela correspond à l'entrée en application de l'ANI (voir Section 2.2.1). Aussi, l'année 2015 a vu une nette augmentation du nombre de souscriptions des 60-69 ans, souvent retraités

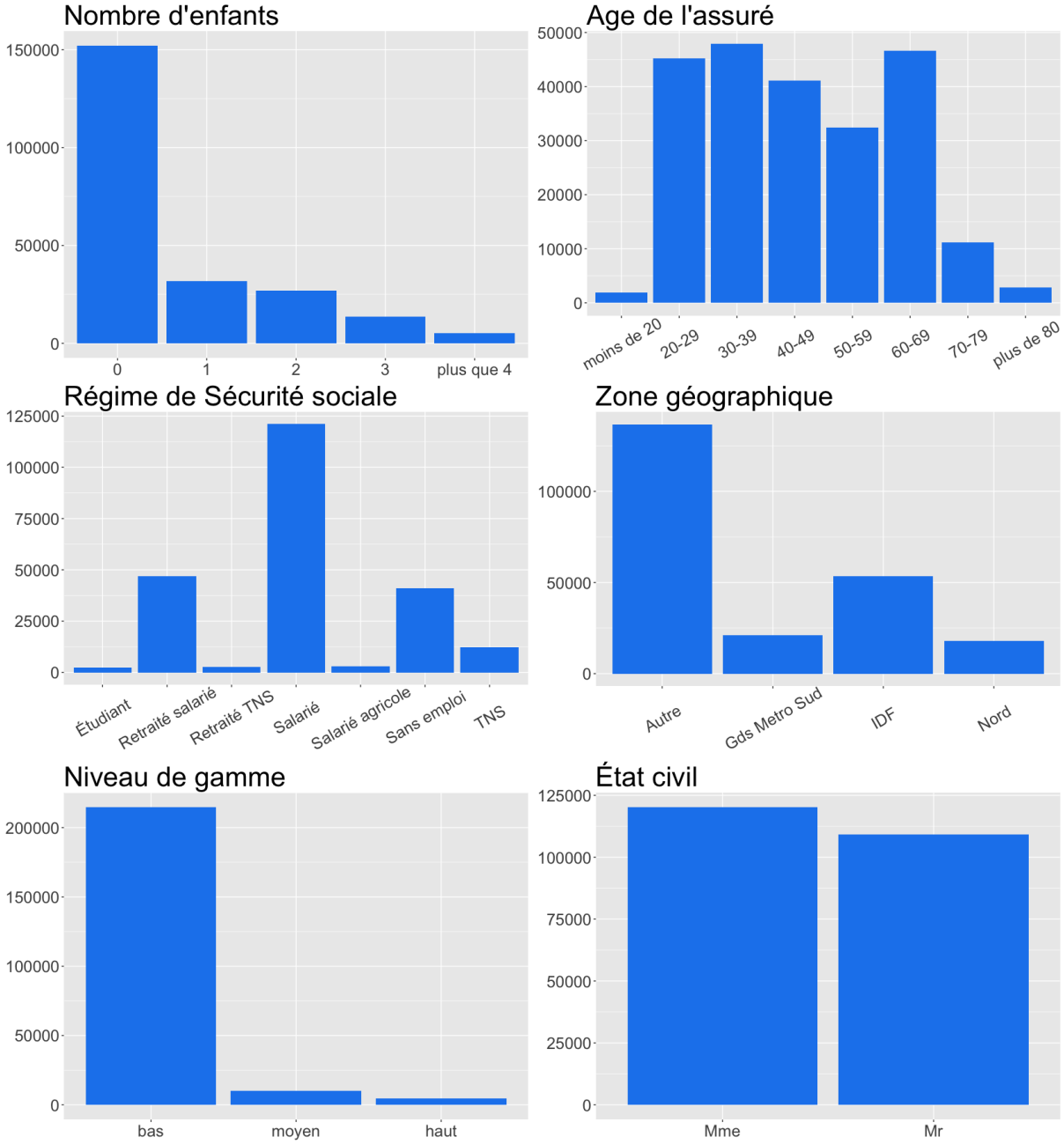

Fig. 2.3: Effectif de chaque modalité pour les variables explicatives utilisées dans l'étude.

TNS : Travailleur non salarié
IDF : Île de France

après avoir occupé un emploi salarié. Cette augmentation, qui semble être confirmée par les chiffres de début 2016, correspond à un choix stratégique du courtier. En effet, les primes d'assurance pour les personnes âgées sont plus importantes, et donnent lieu à de meilleures commissions. Enfin, le rapport homme-femme, qui reste proche de 50%, semble s'inverser au profit d'une part légèrement plus importante d'assurés masculins dans les nouvelles prises d'effet.

2.3.3 Statistiques sur la durée de résiliation

La saisonnalité des résiliations de contrat

La Fig. 2.5 présente le nombre de résiliations enregistrées, par jour, pour l'ensemble des contrats de la base de données. Ce graphique permet d'observer que l'activité du courtier présente une double saisonnalité, avec une périodicité annuelle associée à un pic important de résiliations le 31 décembre de chaque année, mais aussi une périodicité mensuelle avec des résiliations qui interviennent le dernier jour de chaque mois. Lorsque nous avons abordé les aspects réglementaires relatifs à la résiliation d'un contrat d'assurance complémentaire santé à la Section 2.2.2, nous avons vu que la résiliation est en général possible à l'échéance annuelle du contrat. Ceci est vrai pour la première année, toutefois la deuxième année est souvent raccourcie pour terminer le 31 décembre. Les échéances annuelles suivantes sont alors fixées au 31 décembre de chaque année. Ceci explique que l'on observe de nombreuses résiliations le 31 décembre de chaque année. Les résiliations qui ne se produisent pas le 31 décembre peuvent correspondre à des non renouvellements à l'issue de la première année, ou à des résiliations hors période de renouvellement.

La Fig. 2.5 explique également que nous ayons choisi le 15 janvier 2016 comme date de fin d'observation. En effet, à cette date l'ensemble des résiliations survenues à la fin de l'année 2015 sont passées et on obtient un estimateur de la fonction de survie des contrats qui décroît plus rapidement qu'en choisissant une date de fin d'observation à un autre moment de l'année. Le choix fait est donc conservatif et permet de ne pas surestimer le facteur de résiliation f_{ch}. Par ailleurs, les résiliations datées d'un jour donné peuvent être saisies dans la base de données seulement plusieurs mois plus tard. Par exemple, la Fig. 2.6 donne l'histogramme des dates de saisie pour les résiliations datées du 31/12/2014, ainsi que le nombre cumulé de résiliations saisies. On observe que la saisie des résiliations datées du 31/12/2014 est souvent anticipée (dans 75% des cas environ), mais aussi qu'il faut attendre fin juillet pour que 98.7% des résiliations datées du 31/12/2014 soient saisies.

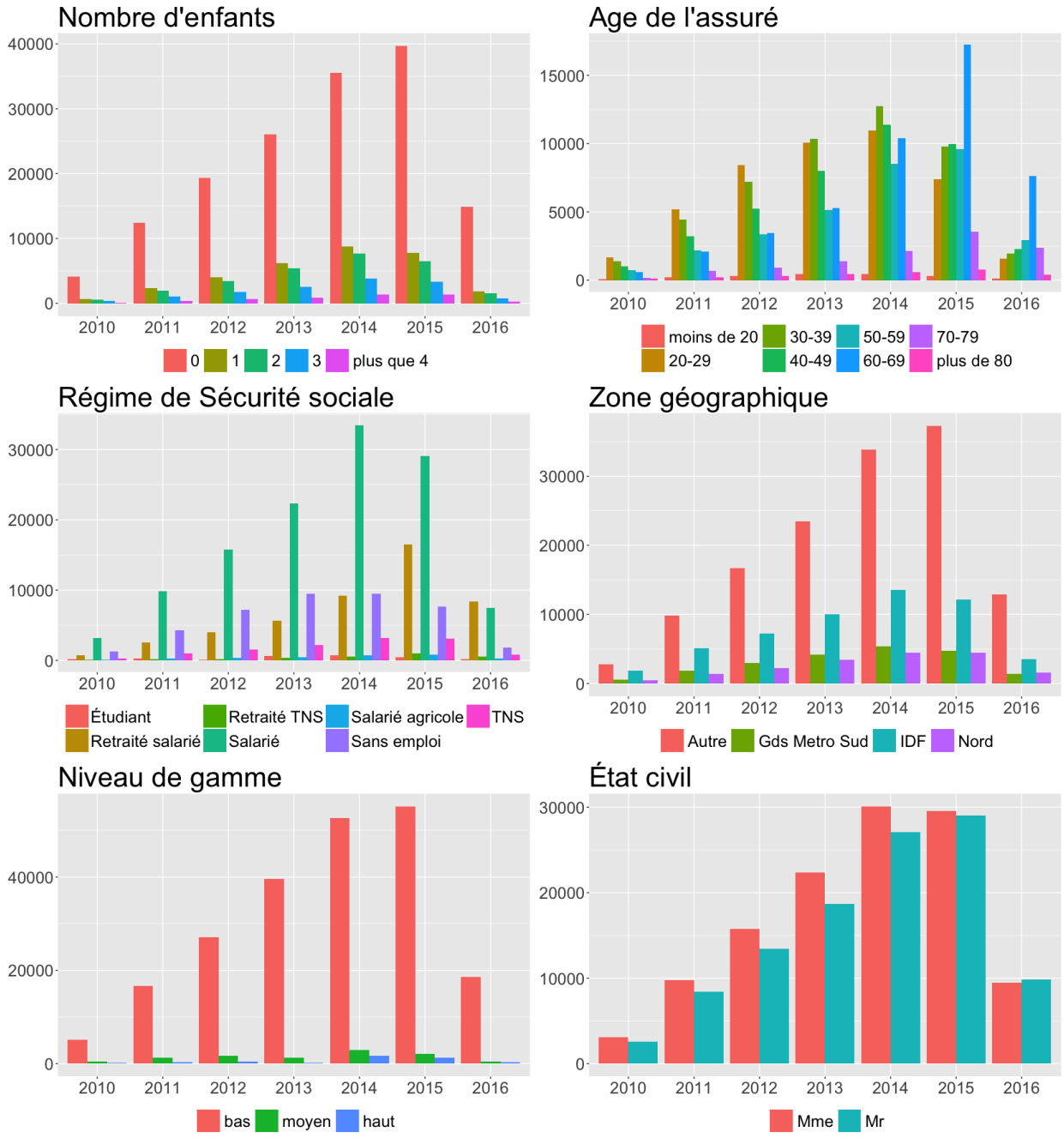

Fig. 2.4: Évolution de l'effectif de chaque modalité en fonction de l'année d'effet.

La base de données dont nous disposons ayant été extraite le 31/07/2016, on peut penser que presque toutes les résiliations qui sont intervenues fin 2015 sont reportées dans la base. Il est donc cohérent d'utiliser le 15/01/2016 comme date de fin d'observation.

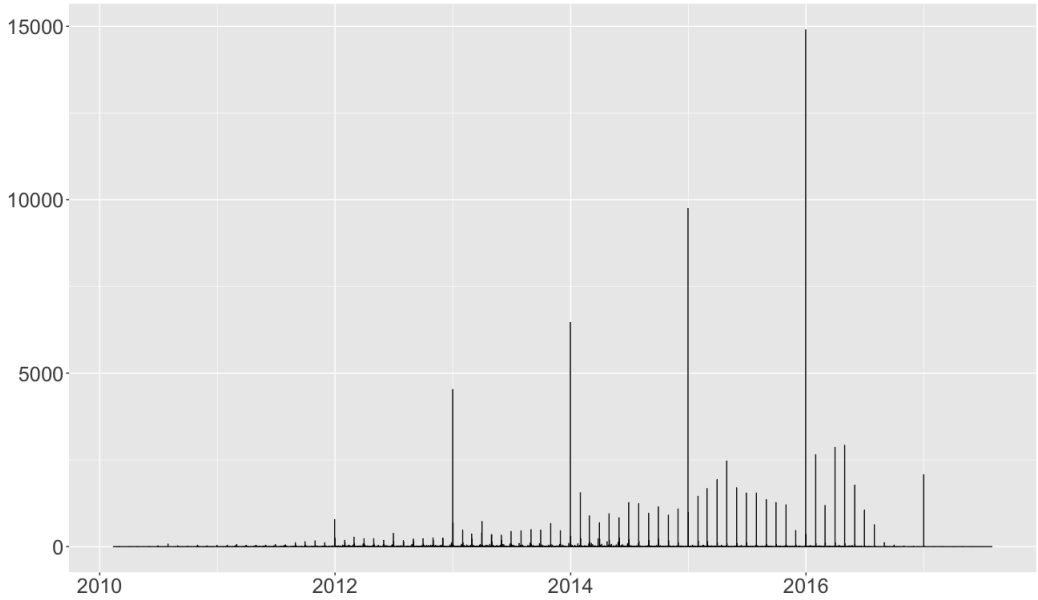

Fig. 2.5: Nombre de résiliations en fonction du jour de résiliation.

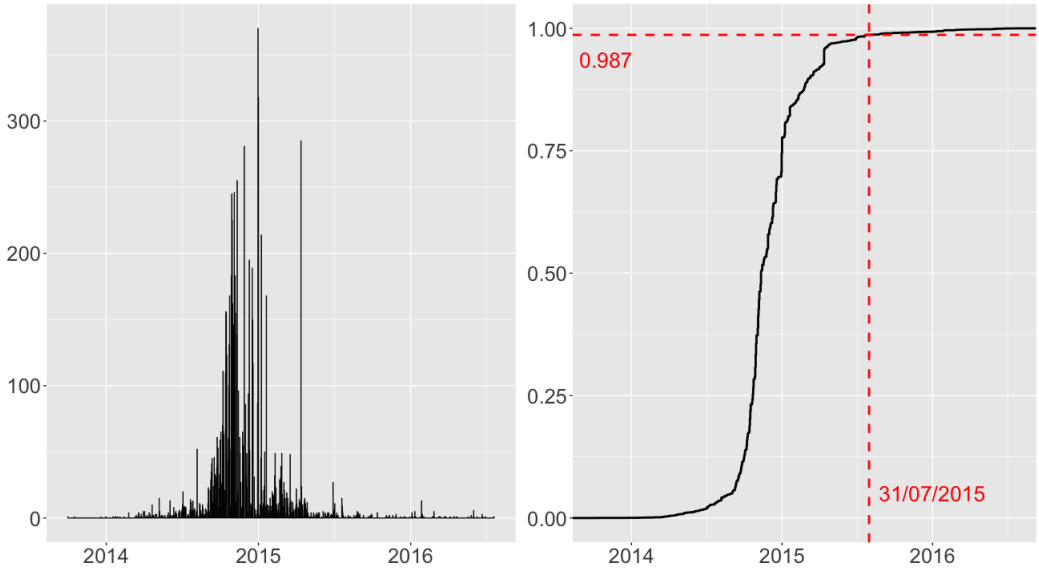

Fig. 2.6: À gauche : histogramme du nombre de saisies (par jour) pour les contrat résiliés le 31/12/2014. À droite : proportion cumulée de saisies opérées pour les contrat résiliés le 31/12/2014. La date de saisie est indiquée en abscisse.

La fonction de survie globale du portefeuille

L'estimateur de Kaplan-Meier de la fonction de survie de la durée de résiliation, calculé sur l'ensemble du portefeuille, est représenté sur la Fig. 2.7. Les sauts observés tous les 365 jours sont dus à la présence de pics de résiliations, que nous avons mis en évidence à la section précédente. Le taux de censure pour la durée de résiliation est de 54.9%, alors que la durée de résiliation médiane est de 2 ans. Le taux de censure élevé s'explique par la croissance de l'activité que nous avons constatée à la Fig. 2.2, qui induit qu'une partie importante du portefeuille de contrats est constituée de contrats récents.

L'impact des variables explicatives et de l'année d'effet sur la durée de résiliation

L'effet des variables explicatives présentées à la Section 2.3.2 sur la durée de résiliation est illustré sur la Fig. 2.8. On observe que les assurés qui ne possèdent pas d'enfants ont tendance à résilier moins rapidement leurs contrats. Par ailleurs, la propension à résilier son contrat croît avec le nombre d'enfants. De même, les assurés âgés (retraités) résilient moins vite que les autres. À l'opposé, la catégorie des sans-emplois présente un taux de résiliation légèrement plus élevé que la moyenne. Les autres variables explicatives représentées n'ont pas un impact marginal important sur la durée de résiliation, toutefois elle sont utilisées dans l'étude du Chapitre 4 de manière à prendre en compte leurs

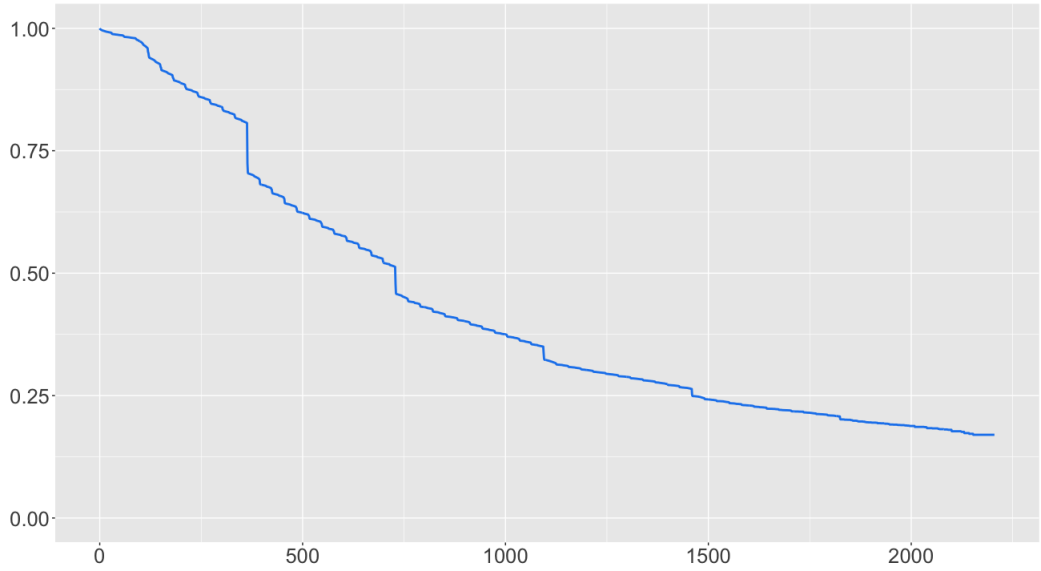

Fig. 2.7: Estimateur de Kaplan-Meier de la fonction de survie de la durée de résiliation (exprimée en jours). Ensemble du portefeuille.

interactions avec les autres variables.

La Fig. 2.9 montre que l'impact de l'année d'effet du contrat sur la durée de résiliation est faible, même si l'historique disponible pour les années récentes est court. Il ne semble donc pas y avoir de tendance longue à l'augmentation ou à la baisse des résiliations.

Au Chapitre 4, nous utilisons une méthode de forêt aléatoire pour prédire le facteur de résiliation $f_{ch} = \phi_{ch}(T)$, à partir des six variables explicatives présentées ici.

2.3.4 Statistiques sur la durée avant effet

Dans cette section, ainsi qu'au Chapitre 3, la base de données étudiée concerne les contrats dont la date d'acquisition est antérieure au 15 janvier 2016. Elle est donc sensiblement différente de la base utilisée à la section précédente et au Chapitre 4, où seuls les contrats prenant effet avant le 15 janvier 2016 étaient considérés. Finalement, la base de données étudiée ici contient 224 897 observations, en prenant en compte le fait que la date d'acquisition n'est pas renseignée pour certains contrats.

L'impact de la saisonnalité des prises d'effet de contrats

La Fig. 2.10 représente le nombre de prises d'effet de contrats par jour tout le long de la fenêtre d'étude. Comme pour les résiliations, les prises d'effet de contrats se concentrent en certains jours de l'année ; environ 37% des contrats prennent effet un 1er janvier. La majorité des autres prises d'effet ont lieu le premier jour d'un mois autre que janvier, ou moins souvent le 15e jour du mois.

Nous appelons *durée avant effet* d'un contrat la durée qui sépare la date de souscription du contrat et la date de prise d'effet du contrat. Le taux de censure de la durée avant effet, qui correspond à la proportion de contrats qui prennent effet après le 15 janvier 2016, est de 1.7% dans les données étudiées. Cette proportion est faible étant donné que la durée avant effet n'excède pas un mois en général et que la majorité des contrats présents dans la base ont plus d'un an d'ancienneté. Elle serait plus importante si nous avions choisi une date de fin d'observation antérieure au 1er janvier 2016. Sur la Fig. 2.11, nous avons représenté la fonction de survie de la durée avant effet et de la durée de résiliation d'un contrat selon que la prise d'effet du contrat se produit un 1er janvier ou bien un autre jour de l'année. On observe une différence nette entre les deux fonctions de survie, qui s'explique par le fait que les contrats qui prennent effet le 1er janvier ont le plus souvent été conclus de manière anticipée, alors que ce n'est pas le cas pour les contrats qui prennent effet à une autre date. Remarquons que la valeur médiane pour la

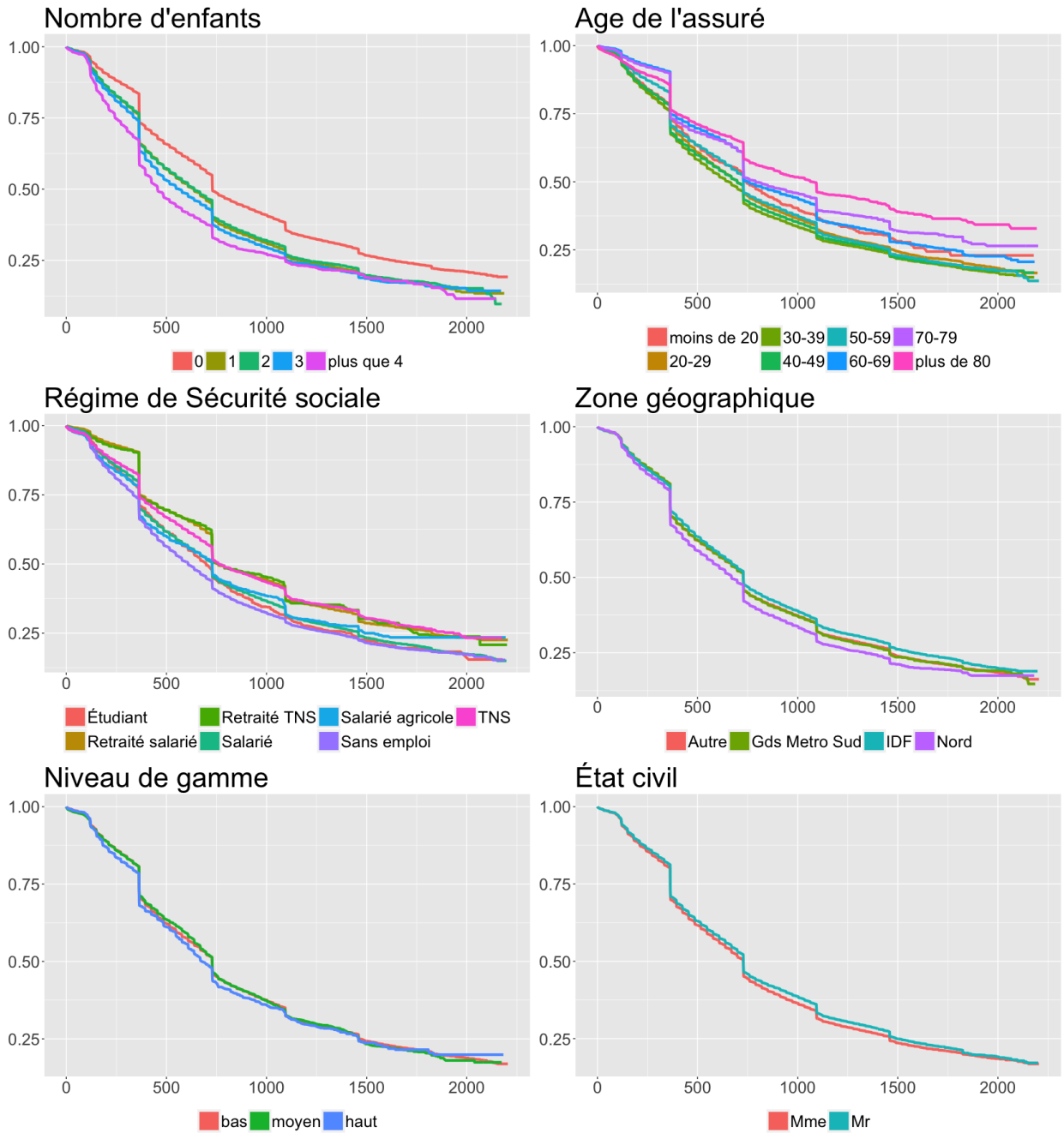

Fig. 2.8: Estimateur de Kaplan-Meier de la fonction de survie de la durée de résiliation pour chaque modalité.

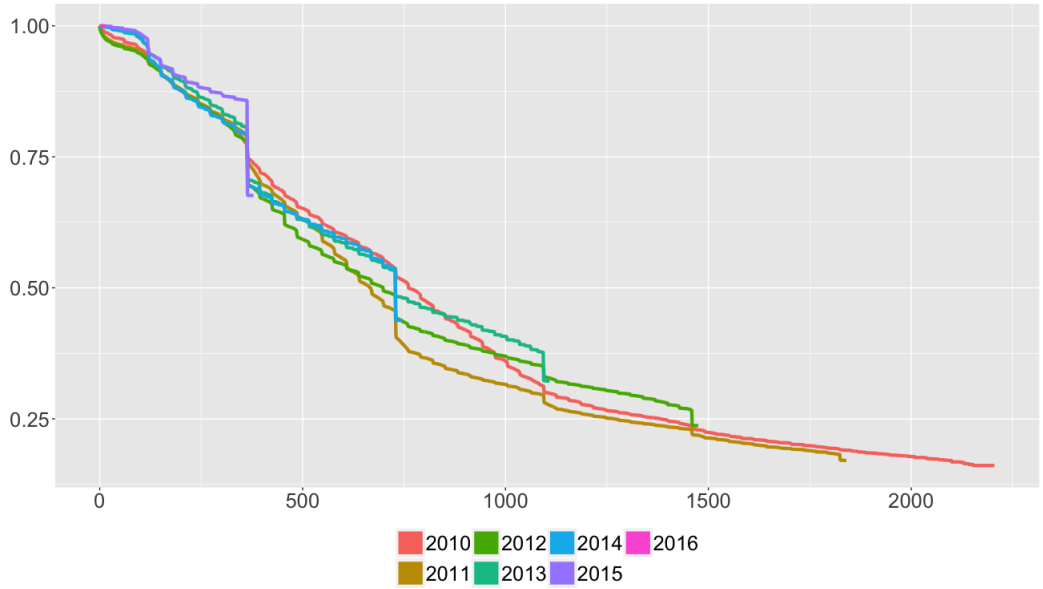

Fig. 2.9: Estimateur de Kaplan-Meier de la fonction de survie de la durée de résiliation par année de prise d'effet.

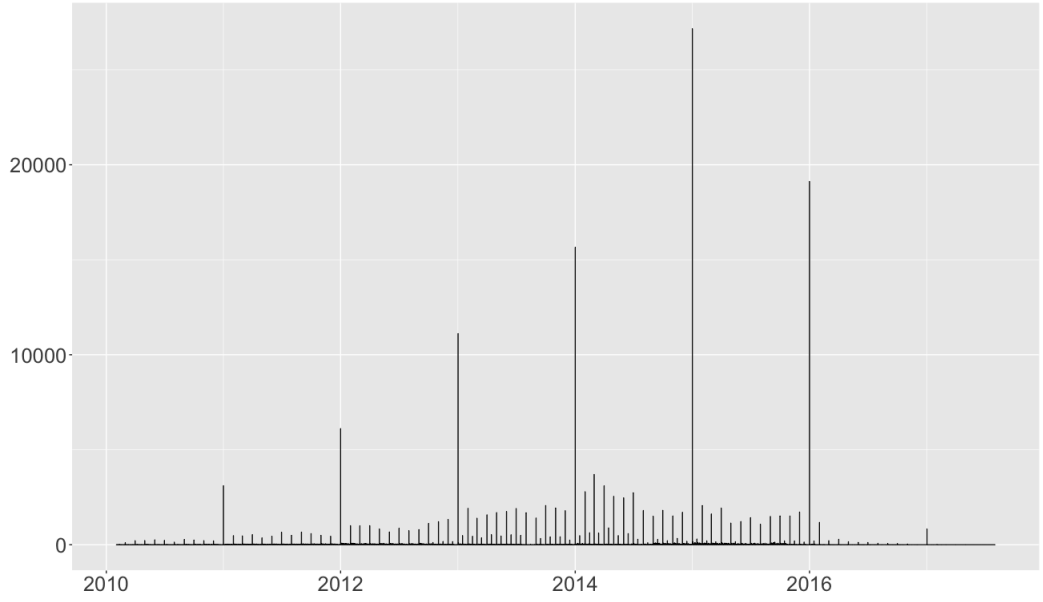

Fig. 2.10: Nombre de prises d'effet par jour.

durée avant effet est de 98 jours en cas de prise d'effet le 1er janvier, et de 8 jours en cas de prise d'effet un autre jour de l'année.

Pour l'étude menée au Chapitre 3, qui porte sur la mesure de la dépendance entre la durée avant effet et la durée de résiliation d'un contrat, ainsi que pour les statistiques descriptives présentées dans la section suivante, nous avons choisi de considérer uniquement

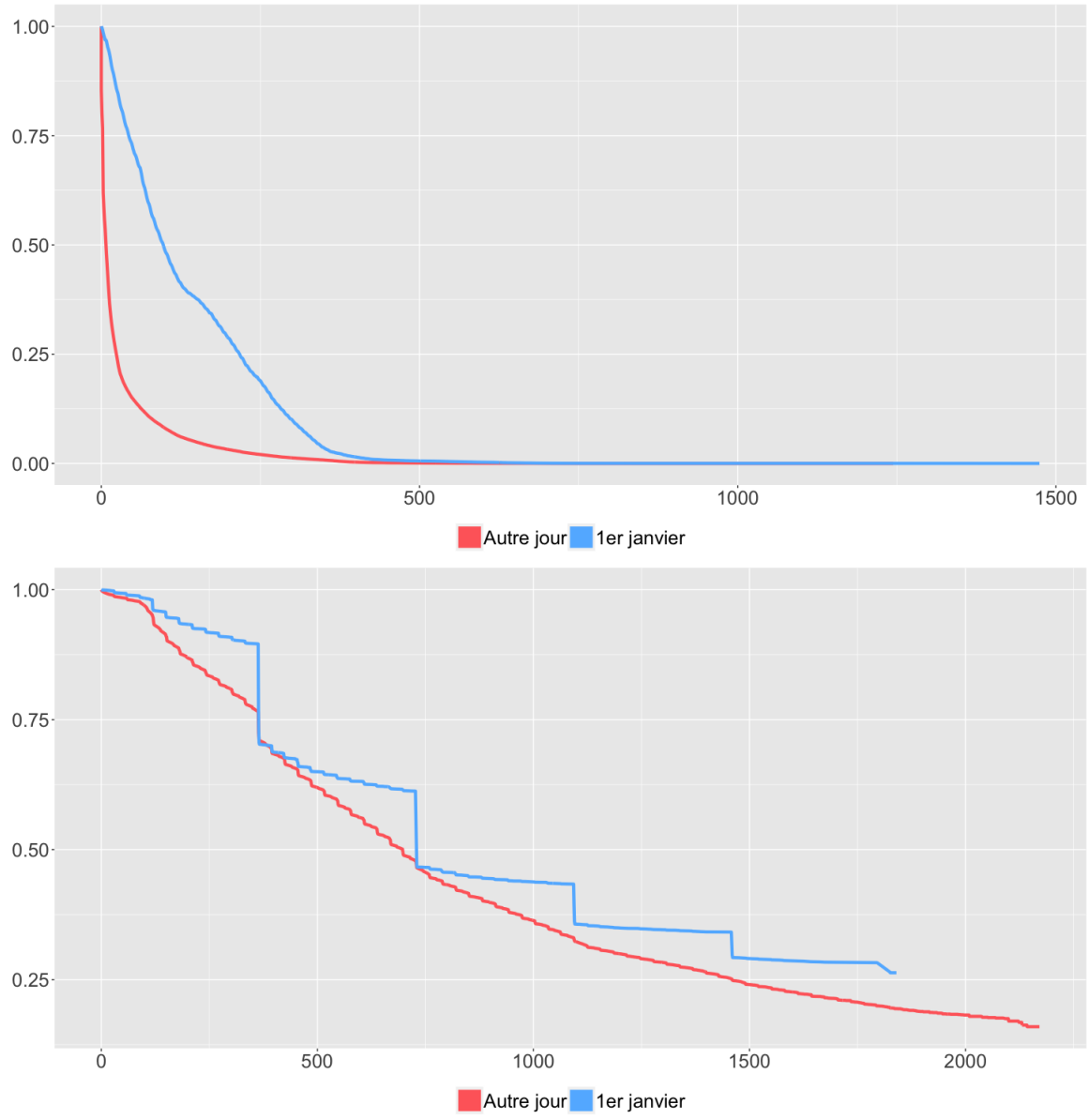

Fig. 2.11: En haut : Fonctions de survie de la durée avant effet pour les contrats qui prennent effet le 1er janvier, et les autres. En bas : même chose mais pour la durée de résiliation

les contrats qui prennent effet le 1er janvier. En effet, les phénomènes observés sont plus importants sur cette partie de la base de données.

L'impact des variables explicatives sur la durée avant effet

Comme pour la durée de résiliation, les variables explicatives peuvent avoir un impact sur la fonction de survie de la durée avant effet. La Fig. 2.12 donne les différentes fonctions de survie obtenues pour chaque modalité. Notons que, comme nous l'avons signalé dans la section précédente, nous considérons ici seulement les contrats qui prennent effet un 1er janvier. On observe que la durée avant effet est réduite pour les assurés jeunes, de moins de 30 ans, dont une partie sont des étudiants. De plus, les contrats haut de gamme présentent une durée avant effet légèrement plus longue. Sur la Fig. 2.13 on a une nouvelle fois représenté l'effet des variables explicatives sur la durée de résiliation, mais en considérant ici seulement les contrats qui ont pris effet un 1er janvier.

Au Chapitre 3, nous présentons une méthode statistique basée sur l'utilisation des copules qui permet d'étudier la dépendance conditionnelle entre la durée avant effet et la durée de résiliation, et plus généralement entre deux durées successives potentiellement censurées à droite. La dépendance est étudiée conditionnellement à l'âge de l'assuré. Remarquons que pour le besoin des représentations graphiques dans ce chapitre, la variable *Âge de l'assuré* a été découpée en plusieurs modalités. Toutefois nous disposons de l'information complète (continue) concernant l'âge des assurés, dont la densité est représentée à la Fig. 2.14. L'information sur l'âge de l'assuré est donc traitée comme variable continue au Chapitre 3.

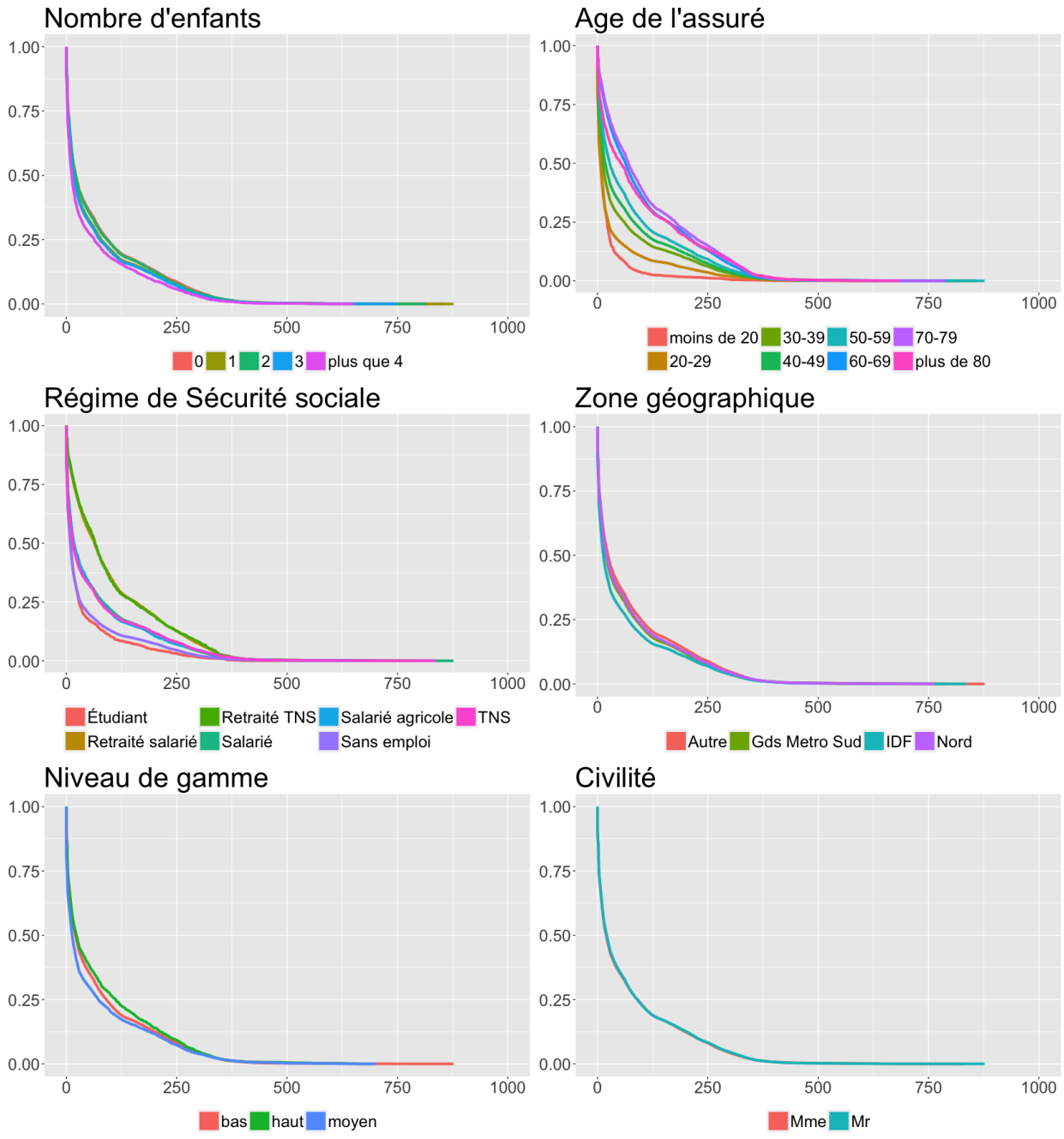

Fig. 2.12: Estimateur de Kaplan-Meier de la fonction de survie de la durée avant effet pour chaque modalité. Base des contrats prenant effet le 1er janvier.

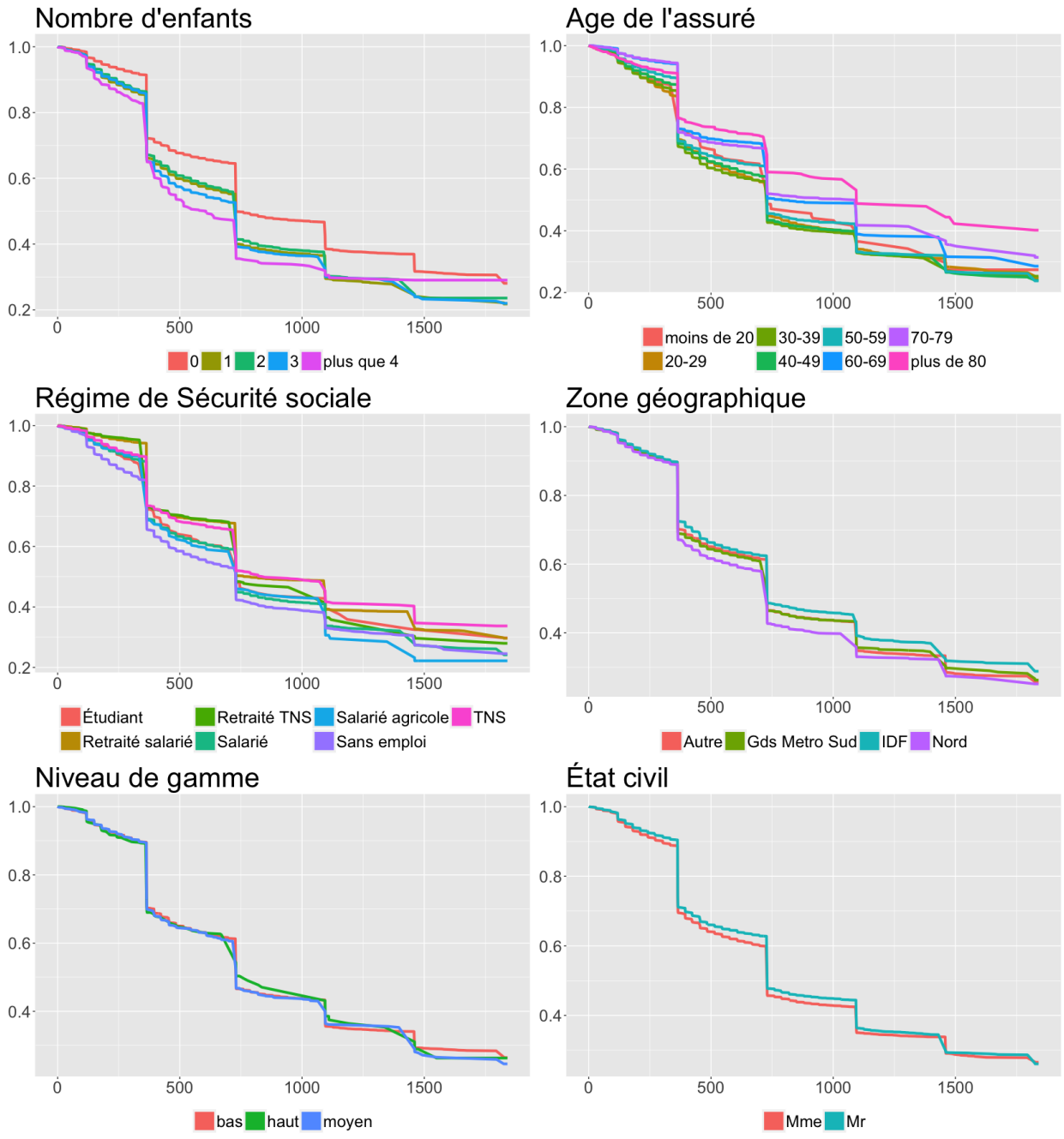

Fig. 2.13: Estimateur de Kaplan-Meier de la fonction de survie de la durée de résiliation pour chaque modalité. Base des contrats prenant effet le 1er janvier.

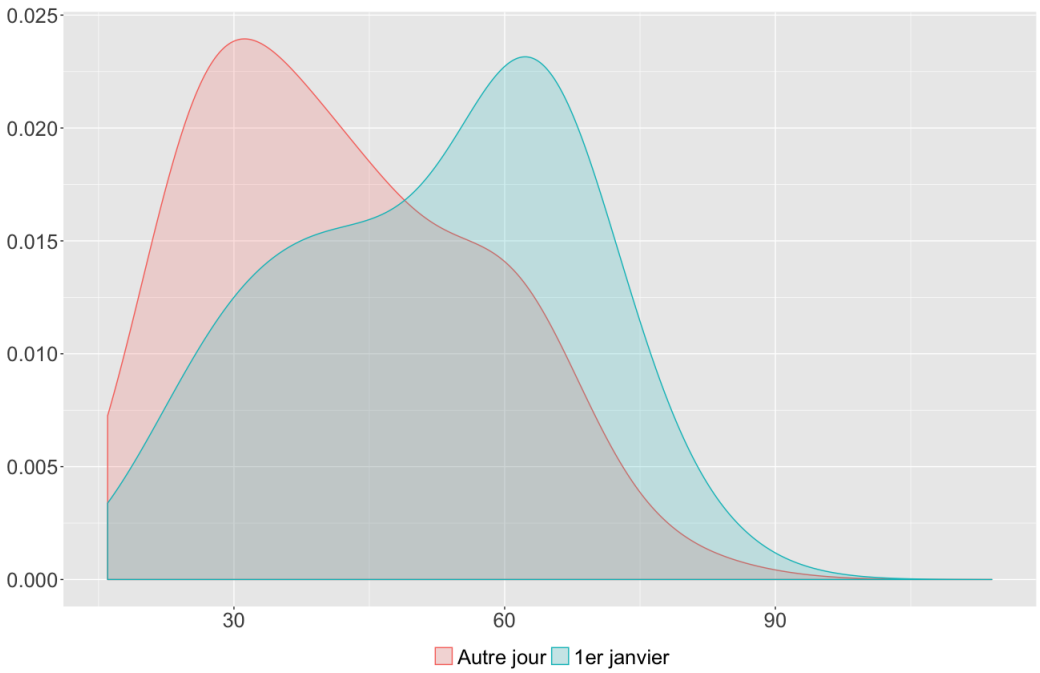

Fig. 2.14: Estimateur à noyau de la densité de la variable *Âge de l'assuré*, pour les contrats qui prennent effet le 1er janvier, et les autres.

Chapitre 3

A nonparametric conditional copula model for successive duration times, with application to insurance subscription

We consider two dependent random times T and U, that correspond to two successive events. This setting is motivated by an application to insurance subscription, where a potential dependence exists between a time before effectiveness of the contract T, and a time U before its termination by the policyholder. The setting also extends to various types of applications involving two duration variables with some hierarchical link between the events. Indeed, since a contract can be terminated only after it becomes effective, data are subject to a particular type of censoring, where the variable U is systematically censored when the variable T is. In this framework, a nonparametric conditional copula model is considered, in the spirit of Gijbels et al. (2011). The uniform consistency of the conditional association parameter is obtained under conditions of dependence structure and of censoring mechanism. A simulation study and a real data application show the practical behavior of the method.

3.1 Introduction

In this paper, we consider the estimation of a conditional copula function of a couple of duration variables, in a framework where the two durations are observed successively. In

numerous practical situations, one can be interested in the occurrence of two events that happen in a time succession. The example we have in mind comes from the study of the termination of insurance contracts. From the subscription, the owner may have some random delay T before the contract to be effective, while the termination of the contract occurs at a time $T+U$, with $U \geq 0$. Similar situations occur in biostatistics, where T can be an infection time and U the time before the individual is cured (see for example Wang & Wells (1998), and Meira-Machado et al. (2016) for a detailed review of the applications and techniques in this field). This paper aims to study the dependence structure between two such times T and U, in presence of covariates $X \in \mathbb{R}^d$ that may impact the joint distribution.

Copula theory is a quite popular way to deal with dependence, due to Sklar's Theorem Sklar (1959) which states that the joint distribution function $F(t,u) = \mathbb{P}(T \leq t, U \leq u)$ of a bivariate vector (T,U) can be written as

$$F(t,u) = \mathfrak{C}(F_T(t), F_U(u)),$$

where $F_T(t) = \mathbb{P}(T \leq t)$, $F_U(u) = \mathbb{P}(U \leq u)$, and $\mathfrak{C}$ is a copula function (that is a distribution function over $[0,1]^2$ with uniform margins), this decomposition being unique when the margins are continuous. Hence Sklar's Theorem ensures a separation between the marginal behaviors of T and U (defined by F_T and F_U), and the dependence structure, entirely contained in the function $\mathfrak{C}$. Conditional copulas are required when one focus on the influence of covariates X on this dependence structure (see e.g. Gijbels et al. (2011), Veraverbeke et al. (2011), Derumigny & Fermanian (2017)).

When dealing with duration variables, a supplementary difficulty is caused by the censoring phenomenon. In the situation we describe, a unique censoring variable C is involved, representing the time before the end of the statistical study. Indeed, if $C < T+U$, the policyholder did not stay under observation long enough to observe the whole phenomenon we are interested in. Copulas under censoring have been studied, for instance by Lakhal-Chaieb (2010), Gribkova & Lopez (2015) and Geerdens et al. (2018). In this paper, we correct the effects of the censoring by using appropriate weights that allow our conditional copula estimator to be asymptotically consistent.

The rest of the paper is organized as follows. In Section 3.2, we define the observations and the methodology to estimate conditional copulas under censoring. Then, the Section 3.3 is devoted to the presentation of asymptotic results while we investigate the finite sample behavior of the procedure in a simulation study and a real data analysis, presented

in Section 3.4. Technical arguments are presented in the Appendix section.

3.2 Observations and Methodology

3.2.1 Model

We consider i.i.d. replications $(T_i, U_i, X_i, C_i)_{1\leq i\leq n}$ of a random vector (T, U, X, C) and we aim to study the dependence structure between $T \in \mathbb{R}$ and $U \in \mathbb{R}$. The random variable $X \in \mathbb{R}^d$ is a vector of covariates that may have an impact on this dependence structure (and also on the marginal distributions of T and U), and $C \in \mathbb{R}$ is a censoring variable.

The variables T and U are not always observed, due to the presence of the censoring. Instead of (T_i, U_i), one observes

$$\begin{cases} Y_i &= \min(T_i, C_i), \\ Z_i &= \min(U_i, C_i - T_i), \\ \eta_i &= \mathbf{1}_{T_i\leq C_i}, \\ \gamma_i &= \mathbf{1}_{T_i+U_i\leq C_i}. \end{cases}$$

The covariates X_i are assumed to be fully observed (not subject to censoring). For the realization of the censoring C_i, two cases exist. In the application we have in mind (see Section 3.2.2 and Section 3.4.2), C_i is known for all observations. In a more general situation, C_i may not be known. In this last case, the statistical methodology that we develop is a little bit more delicate as we will see in the following.

The following identifiability assumption is required in order to estimate the distribution of (T, U, X) from the observations.

Assumption 0.1 *Assume that C is independent from (T, U, X).*

Let $F(t, u|x) = \mathbb{P}(T \leq t, U \leq u|X = x)$ be the conditional distribution function of (T, U) given $X = x$, and $F_T(t|x) = \mathbb{P}(T \leq t|X = x)$ (resp. $F_U(u|x) = \mathbb{P}(U \leq u|X = x)$) be the conditional distribution function of T (resp. U), where all distribution functions are assumed to be continuous. We also define $\tau_{T+U}(x) = \inf\{z : \mathbb{P}(Y + Z \geq z|x) = 0\}$. Clearly, the distribution of (T, U, X) can not be estimated (at least nonparametrically) for values of (t, u, x) such that $t + u \geq \tau_{T+U}(x)$, since it is impossible to observe non-censored events in this part of the distribution. Sklar's Theorem ensures that $F(t, u|x) = \mathfrak{C}^{(x)}(F_T(t|x), F_U(u|x))$, where $\mathfrak{C}^{(x)}$ denotes the copula of the distribution of (T, U) conditionally on $X = x$. In the following, we assume that the copula $\mathfrak{C}^{(x)}$ stays in the same

parametric copula family for all x, but with its association parameter allowed to depend on x. This is summarized in Assumption 0.2 below.

Assumption 0.2 *Let $\mathcal{C} = \{\mathfrak{C}_\theta : \theta \in \Theta\}$, with Θ a compact subset of $\mathbb{R}^k$, be a parametric family of copula functions. Assume that, for all x in the support of the random vector X, there exists $\theta(x) \in \Theta$ such that*

$$\mathfrak{C}^{(x)} = \mathfrak{C}_{\theta(x)}.$$

Our aim is to retrieve the function $\theta(x)$ from our observations.

3.2.2 Motivation of this model

Our method applies to a problem which arises in the field of insurance subscription. The data we consider (described in Section 3.4.2) belongs to an insurance broker who wants to have information about the quality of the underwriters who sell the insurance contracts. A first indicator would be the volume of sales per underwriter, but a crucial issue is to have insight in the quality of the contracts that have been subscribed. One element to appreciate this quality is the time the consumer keeps his contract, before terminating it and starting another contract with a different insurer. In our framework, the lifetime of the contract is the variable U, that is the difference between the date of termination of the contract and the date of effect. The date of effect of the policy is usually not the same as the date of subscription. We denote by T the time between the date of subscription and the date of effect.

It seems obvious that the two durations T and U should not be independent. The knowledge of their dependence structure is a precious indicator to develop sales strategies and to evaluate the turnover in an insurance portfolio. Additionally, many variables on the customer are usually available, and these variables may have an impact on the dependence structure. This motivates the use of conditional copulas to model the dependence between T and U.

3.2.3 Conditional copula estimation

Let $M(x, \theta) = E[\log c_\theta(F_T(T|X), F_U(U|X))|X = x]$, where $c_\theta(a, b) = \partial^2_{a,b}\mathfrak{C}_\theta(a, b)$ denotes the copula density associated with copula function $\mathfrak{C}_\theta$. We have, by definition of $\theta(x)$,

$$\theta(x) = \arg\max_{\theta\in\Theta} M(x, \theta).$$

To ensure identifiability of the model, we assume that for all x in the support of X, $\theta(x)$ is the unique maximum of $M(x,\theta)$. The idea of our procedure is to estimate the function M, and then to perform its maximization in order to estimate $\theta(x)$.

In an ideal situation, first consider that F_T and F_U are exactly known. If we had observed the complete data $(T_i, U_i, X_i)_{1\leq i\leq n}$, we could have estimated $M(x,\theta)$ thanks to a Nadaraya-Watson estimator (Watson (1964) and Nadaraya (1964)) such as

$$\sum_{i=1}^{n} w_{i,n}(x) \log c_\theta(F_T(T_i|X_i), F_U(U_i|X_i)),$$

where

$$w_{i,n}(x) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)}, \tag{3.1}$$

and K is a kernel function (i.e. a positive and symmetric real valued function such that $\int K(u)du = 1$). However, this is impossible in our case due to the presence of censoring. If we consider nevertheless a function ϕ such that $E[|\phi(T,U,X)|] < \infty$ and $\phi(t,u,x) = 0$ for $t + u \geq \tau_{U+T}(x)$, then under Assumption 0.1, elementary computations show that

$$E\left[\frac{\delta\phi(Y,Z,X)}{S_C(Y+Z)}\middle| X\right] = E\left[\phi(T,U,X)|X\right], \tag{3.2}$$

with $S_C(t) = \mathbb{P}(C > t)$, and $\delta = \eta\gamma$.

As a consequence of equation (3.2), we see that if the function S_C were known, the function $M(x,\theta)$ could be estimated by the kernel estimator

$$\sum_{i=1}^{n} w_{i,n}(x) \frac{\delta_i \log c_\theta(F_T(Y_i|X_i), F_U(Z_i|X_i))}{S_C(Y_i + Z_i)},$$

with $w_{i,n}(x)$ as in (3.1). The kernel function K that we consider in this article is assumed to be a symmetric positive and bounded function, with $K(u) = 0$ for $\|u\| \geq 1$, $\int K(u)du = 1$ and $\int \|u\|^2 K(u)du < \infty$.

In practice, the function S_C is not known. However, it can be estimated, at least in the two situations that we mention in Section 3.2.1.

- **First case: C_i is observed for all individuals.** In this situation, we can estimate

$S_C(t)$ with the empirical survival function, i.e.

$$\hat{S}_C^{(1)}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{C_i > t}. \tag{3.3}$$

- **Second case: Some C_i are unobserved due to the censoring.** In this second situation, the function S_C can be estimated by a Kaplan-Meier estimator, see Kaplan & Meier (1958a). Put more precisely, consider

$$\begin{cases} Y_i' &= \min(C_i, T_i + U_i), \\ \delta_i' &= \mathbf{1}_{C_i \leq T_i + U_i}. \end{cases} \tag{3.4}$$

  In our framework, the variables (Y_i', δ_i') are observed. Moreover, C_i is independent from $T_i + U_i$ from Assumption 0.1. As a consequence, the survival function S_C can be consistently estimated by

$$\hat{S}_C^{(2)}(t) = \prod_{Y_i' \leq t} \left(1 - \frac{\delta_i'}{\sum_{j=1}^{n} \mathbf{1}_{Y_j' \geq Y_i'}}\right),$$

  assuming that there are no ties among the $(Y_i')_{1 \leq i \leq n}$.

Additionally, the margins F_T and F_U may not be known in practice. Several techniques may be used to estimate them, as it will be discussed in Section 3.2.4. To state the results in the most general form, we define $A_i = F_T(Y_i|X_i)$ and $B_i = F_U(Z_i|X_i)$, and consider that we have at our disposal pseudo-observations $(\hat{A}_i, \hat{B}_i)_{1 \leq i \leq n}$.

This leads to our final estimator of $\theta(x)$, that is

$$\hat{\theta}_h(x) = \arg\max_{\theta \in \Theta} M_{n,h}(x, \theta), \tag{3.5}$$

where

$$M_{n,h}(x,\theta) = \frac{1}{nh^d}\sum_{i=1}^{n} W_{i,n} K\left(\frac{X_i - x}{h}\right) \log c_\theta(\hat{A}_i, \hat{B}_i)\hat{w}_i(\nu_n), \tag{3.6}$$

with $W_{i,n} = \delta_i / \hat{S}_C^{(j)}(Y_i + Z_i)$ for $j = 1, 2$ depending on our ability to observe C_i or not, ν_n a sequence tending to zero which will be defined later on (see Section 3.3.2), and introducing

a trimming function $\hat{w}_i$ defined for a sequence η_n, as

$$\hat{w}_i(\eta_n) = \mathbf{1}_{\min(\hat{A}_i, \hat{B}_i, 1-\hat{A}_i, 1-\hat{B}_i) \geq \eta_n}.$$

The presence of the trimming is required to prevent the procedure from an erratic behavior when the pseudo-observations are close to the border of the unit square. The weights $W_{i,n}$ may be seen as an approximation of $W_i = \delta_i / S_C(Y_i + Z_i)$, which, according to (3.2), would be the way we would correct the presence of the censoring if we knew exactly its distribution S_C. Let us observe that $M_{n,h}(x, \theta)$ is an estimator of $M_f(x, \theta) = M(x, \theta) f_X(x)$, where f_X denotes the density of the random vector $X \in \mathbb{R}^d$. Maximizing M or M_f with respect to θ is of course equivalent.

The practical performance of our nonparametric estimator $\hat{\theta}_h(x)$ depends on an appropriate choice of the bandwidth h. In Section 3.4.2, we present a data-driven method to select an appropriate h empirically.

3.2.4 Estimation of the margins

Sklar's Theorem allows to separate the marginal distributions from the dependence structure. Therefore we do not wish to specify the way the margins are estimated since various approaches may be used, possibly different from a margin to another. We will assume that this estimation has been performed separately, in a preliminary step. A parametric or semiparametric model can be put on the margins. If one only focuses on the dependence structure, a nonparametric estimation of the margins can be performed, for example using kernel estimation. In this case, one may use

$$\hat{F}_T(t|x) = \sum_{i=1}^{n} W_{i,n} \frac{K\left(\frac{X_i - x}{h'}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h'}\right)} \mathbf{1}_{Y_i \leq t}, \tag{3.7}$$

with a similar definition for $\hat{F}_U$ (replacing Y_i by Z_i). Let us note that the bandwidth h' can be different from the bandwidth used to estimate the copula parameter. Moreover, different bandwidths may be used for each of the margins. The method of equation (3.7) is the one that we use in the real data application of Section 3.4.2, whereas for the simulated data (Section 3.4.1) we use a Kaplan-Meier estimator without kernel since the marginal distributions are assumed to be independent of X.

The results that we propose in the following are valid under the condition that $\hat{A}_i = \hat{F}_T(Y_i|X_i)$ and $\hat{B}_i = \hat{F}_U(Z_i|X_i)$ are close to $A_i = F_T(Y_i|X_i)$ and $B_i = F_U(Z_i|X_i)$, but do

not impose a particular method to compute the pseudo-observations. The conditions that we require (see Assumption 0.8) are valid for a large number of estimation techniques and at least hold for the estimator (3.7).

3.3 Uniform rate of convergence for $\hat{\theta}_h(x)$

As usual in nonparametric estimation, the error $\hat{\theta}_h(x)-\theta(x)$ can be decomposed into some bias term, and a stochastic term that corresponds to the fluctuations of $\hat{\theta}_h(x)$ around its central value. The most delicate convergence result to obtain is Theorem 3.2, which deals with the stochastic term, while Theorem 3.1 only studies the deterministic term, that can be handled by standard results in approximation theory.

3.3.1 Bias term

Let

$$\theta_h^*(x) = \arg\max_{\theta\in\Theta} \frac{1}{h^d} E\left[K\left(\frac{X-x}{h}\right)\log c_\theta(F_T(T|X), F_U(U|X))\right].$$

The difference between $\theta_h^*(x)$ and $\theta(x)$ represents the bias of the method, where the empirical mean in $M_{n,h}$ has been replaced by its limit value (we will show this convergence in the Appendix section). The aim of this section is to determine the uniform rate of convergence of this bias term on a set $\mathcal{X}$, which is assumed to be compact and strictly included in the support of the random vector X.

We use the notation $c(a,b|x) := c_{\theta(x)}(a,b)$ to denote the the conditional copula density given $X = x$. Also, let $\phi(a,b,\theta) = \log c_\theta(a,b)$, $\dot{\phi}(a,b,\theta) = \nabla_\theta \log c_\theta(a,b)$ and $\ddot{\phi}(a,b,\theta) = \nabla_\theta^2 \log c_\theta(a,b)$.

Assumptions 0.3 to 0.5 are required to obtain the convergence of the bias term. The first two can be understood as regularity assumptions on the model when x varies.

Assumption 0.3 *Assume that the function $(a,b,\theta) \to \phi(a,b,\theta)$ is twice continuously differentiable with respect to θ, and that for all $\theta \in \Theta, x \in \mathcal{X}$,*

$$\{|1+\phi(a,b,\theta)| + \|\ddot{\phi}(a,b,\theta)\|\}c(a,b|x) \le \Lambda_1(a,b), \tag{3.8}$$

with $\int \Lambda_1(a,b)dadb < \infty$.

Assumption 0.4 *Assume that the function* $(a, b, x) \to \mathfrak{c}(a, b|x) f_X(x)$ *is twice continuously differentiable with respect to* x*, and that for all* $\theta \in \Theta, x \in \mathcal{X}$*,*

$$\|\dot{\phi}(a, b, \theta)\| \cdot \|\nabla_x^2 \{c(a, b|x) f_X(x)\}\| \leq \Lambda_2(a, b), \tag{3.9}$$

with $\int \Lambda_2(a, b) da db < \infty$*.*

The next assumption is required to ensure that the maximization problem is locally quadratically close to the true value $\theta(x)$.

Assumption 0.5 *Assume that there exists some* $c_0 > 0$ *such that, for all* $x \in \mathcal{X}$*, we have*

$$\forall v \in \mathbb{R}^d, \langle E[\ddot{\phi}(F_T(T|X), F_U(U|X), \theta(x))|X = x] \cdot v, v\rangle \leq -c_0\|v\|^2 \leq 0,$$

where $\langle v, w\rangle$ *denotes the scalar product of two vectors* v *and* w *in* $\mathbb{R}^d$*.*

Moreover, assume that the density of X *denoted by* f_X *is such that* $f_X(x) \geq c_0$ *for all* $x \in \mathcal{X}$*.*

We now can state our result about the bias term.

Theorem 3.1 *Under Assumptions 0.1 to 0.5,*

$$\sup_{x \in \mathcal{X}} \|\theta_h^*(x) - \theta(x)\| = O(h^2).$$

This h^2 rate is classical when dealing with kernel smoothing. This rate could of course be improved by strengthening the regularity of the conditional copula function, and by considering a higher order kernel (that is a function K such that $\int u^j K(u) du = 0$ for all $j \leq k$, with k larger than 1).

The proof of Theorem 3.1 is dealt with in the Appendix section (see Section 3.6.1).

3.3.2 Stochastic term

This section presents the main theoretical result of the paper, which shows the uniform convergence of the stochastic term. The convergence rate involves four terms as given in Theorem 3.2 below; a traditional rate for kernel smoothing estimators ($n^{-1/2}h^{-d/2}[\log n]^{1/2}$), and additional terms that may become preponderant if the estimation of the margins is performed at a slow rate or if the copula density and its derivatives behave too wildly

close to the frontier of the unit square. We first begin with the assumptions required to obtain the result.

Since $\hat{\theta}_h(x)$ can be seen as a conditional version of the semiparametric copula estimator proposed by Tsukahara (2005) and Genest et al. (1995b), the conditions required to obtain the convergence of the stochastic term are basically the same as in these two papers, with some modifications imposed by the use of smoothing and because of the censoring.

We remind that a function $r : (0,1) \to (0,\infty)$ is called $u-$shaped if r is symmetric about $1/2$ and decreasing on $(0,1/2]$. For such a $u-$shaped function r, and for $0 < \zeta < 1$, define

$$r_\zeta(t) = r(\zeta t)\mathbf{1}_{0<t\leq 1/2} + r(1-\zeta(1-t))\mathbf{1}_{1/2<t<1}.$$

A $u-$shaped function is called a reproducing $u-$shaped function if it verifies that $r_\zeta \leq M_\zeta r$ for all $\zeta > 0$ in a neighborhood of 0, with M_ζ a finite constant. In the following we note $\mathcal{R}$ the set of reproducing $u-$shaped functions.

Assumptions 0.6 and 0.7 are close to assumptions A.1 to A.5 present in Tsukahara (2005). They ensure that the modulus of continuity of ϕ satisfies some integrability conditions, and that the derivatives of ϕ are dominated by $u-$shaped functions in order to control the explosion of these derivatives close to the border of the unit square. As shown in Tsukahara (2005), these conditions are satisfied by a large number of copula families.

Due to the censoring, a term $S_C(T+U)$ appears at the denominator. A similar assumption is present for example in Gill (1983), Stute (1995) or Gribkova & Lopez (2015). In case of heavy censoring, that is if S_C decreases too fast, the integrability conditions in Assumptions 0.6 and 0.7 may not hold. This is a classical issue in survival analysis: in such a situation, the right-tails of the distributions of T and U are rarely observed since the censoring variable tends to take small values. A solution is then to restrain the study of the distribution of T and U conditionally on $T+U \leq \tau$, where τ is a fixed bound, strictly included in the support of the variable $T+U$, though this introduces an asymptotic bias.

Assumption 0.6 *Assume that*

$$|\log c_\theta(a,b) - \log c_{\theta'}(a,b)| \leq R(a,b)\|\theta - \theta'\|,$$

and that, for some $p > 2$ *and some* $\theta_0 \in \Theta$,

$$\sup_{x\in\mathcal{X}} E\left[\frac{|\log c_{\theta_0}(F_T(T|X),F_U(U|X))|^p+[R(F_T(T|X),F_U(U|X))]^p}{[S_C(T+U)]^{p-1}}\Big|X=x\right]<\infty. \quad (3.10)$$

Also assume that, for some $p'>1$,

$$\sup_{x\in\mathcal{X}} E\left[\frac{|\log c_{\theta_0}(F_T(T|X),F_U(U|X))|^{p'}+[R(F_T(T|X),F_U(U|X))]^{p'}}{S_C(T+U)^{2p'-1}}\Big|X=x\right]<\infty. \quad (3.11)$$

Finally, assume that

$$\liminf_{n\to\infty}\ nh^{\frac{d}{1-2/p}}[\log n]^{-1}>0, \quad (3.12)$$

and that

$$\liminf_{n\to\infty} nh^{2d}>0. \quad (3.13)$$

Assumption 0.7 *Assume that there exist functions* r_j *and* $\tilde{r}_j$ *in* $\mathcal{R}$, $j=1,2$, *such that*

$$\begin{aligned}\|\dot{\phi}(a,b,\theta)\| &\le r_1(a)r_2(b),\\ |\partial_a\phi(a,b,\theta)|+\|\partial_a\dot{\phi}(a,b,\theta)\| &\le \tilde{r}_1(a)r_2(b),\\ |\partial_b\phi(a,b,\theta)|+\|\partial_b\dot{\phi}(a,b,\theta)\| &\le \tilde{r}_2(b)r_1(a).\end{aligned}$$

Assume further that for some $p''>1$, *we have*

$$\sup_{x\in\mathcal{X}} E\left[\left(\frac{\tilde{r}_1(F_T(T|X)^{p''}r_2(F_U(U|X)^{p''}+\tilde{r}_2(F_U(U|X)^{p''}r_1(F_T(T|X)^{p''}}{S_C(T+U)^{p''-1}}\right)\Big|X=x\right]<\infty.$$

The next assumption concerns the estimation of the margins. The results we provide may hold for different strategies of estimation of the margins (nonparametric, semiparametric or parametric). We only require a rate of consistency for the margins, a comparison of this rate with the speed ν_n, and the condition (3.14) below which involves the rate of convergence of the margins and p''.

Assumption 0.8 *Assume that*

$$\sup_{1\le i\le n}|\hat{A}_i-A_i|+|\hat{B}_i-B_i|=O_P(\varepsilon_n),$$

with $\varepsilon_n=o(\nu_n)$, *and*

$$\lim_{\beta\searrow 1}\limsup_{n\to\infty}\ \varepsilon_n n^{1/p''-1/(2\beta)}=0, \quad (3.14)$$

where p'' *is defined in Assumption 0.7.*

We now can state the main result of this section, which is proven in Section 3.6.7.

Theorem 3.2 *Under Assumptions 0.1 to 0.8, we have for any $\beta > 1$,*

$$\sup_{x\in\mathcal{X}} \|\hat{\theta}_h(x) - \theta_h^*(x)\| = O_P(n^{-1/2}h^{-d/2}[\log n]^{1/2} + \nu_n^{1-1/p} + \varepsilon_n n^{\max(1/p''-1/(2\beta),0)} + n^{\max(1/p'-1/(2\beta),0)-1/2}).$$

The rate of convergence of the stochastic term can be decomposed in four parts. The rate $n^{-1/2}h^{-d/2}[\log n]^{1/2}$ corresponds to the standard convergence rate for kernel estimators when the margins and the censoring mechanism are known. The second term, $\nu_n^{1-1/p}$ is caused by the trimming function. The third term comes from the estimation of the margins, while the last one is caused by the estimation of the censoring distribution. Let us discuss the impact of each of these three terms, and of the parameters p, p' and p'' related to the moment conditions of Assumptions 0.6 and 0.7.

- Trimming term: let us recall that the term ν_n can be chosen arbitrarily close to the rate ε_n. At worse, that is if the largest value of p that we can take is close to 2, a rough bound of this rate is $\varepsilon_n^{1/2}$. This rate can be considerably improved if p is large, which happens if the copula density and its derivatives do not explode too fast close to the boundary of the unit square, and if the censoring is not too heavy.

- Estimation of the margins: the rate ε_n is potentially deteriorated if $p'' \leq 2$. Indeed the quantity $n^{\max(1/p''-1/(2\beta),0)}$ is equal to one if $p'' > 2$.

- Estimation of the censoring distribution: as for the previous term, if $p' > 2$, this term is $n^{-1/2}$ and becomes negligible.

Clearly, these three additional terms only disappear if the estimation of the margins is performed at a sufficiently fast rate, and if the explosion of the copula density and its derivatives (combined to the strength of the censoring) is controlled.

3.4 Experiments of the method using data

In the following part, the method developed in Section 3.2 to estimate the conditional dependence parameter of a copula is illustrated numerically.

Four families of parametric copulas (Gaussian, Clayton, Gumbel and Frank) are tested to model the dependence between T and U :

- the Gaussian copula family $\mathcal{C}^1 = \{\mathfrak{C}^1_\theta : \theta \in [-1;1]\}$, where $\mathfrak{C}^1_\theta(a,b) = g_\theta\left(g^{-1}(a), g^{-1}(b)\right)$, with g_θ is the cumulative distribution function of a bivariate Gaussian vector (V_1, V_2) – with mean $E(V_1, V_2) = (0,0)$, marginal variances $\mathrm{Var}(V_1) = \mathrm{Var}(V_2) = 1$ and covariance $\mathrm{Cov}(V_1, V_2) = \theta$ – and g^{-1} is the inverse cumulative distribution function of a standard normal random variable.
- the Clayton copula family $\mathcal{C}^2 = \{\mathfrak{C}^2_\theta : \theta > 0\}$, with $\mathfrak{C}^2_\theta(a,b) = \left(a^{-\theta} + b^{-\theta} - 1\right)^{-1/\theta}$
- the Gumbel copula family $\mathcal{C}^3 = \{\mathfrak{C}^3_\theta : \theta \geq 1\}$, with

$$\mathfrak{C}^3_\theta(a,b) = \exp\left[-\left((-\log(a))^\theta + (-\log(b))^\theta\right)^{1/\theta}\right]$$

- the Frank copula family $\mathcal{C}^4 = \{\mathfrak{C}^4_\theta : \theta \in \mathbb{R}\backslash\{0\}\}$ with

$$\mathfrak{C}^4_\theta(a,b) = -\frac{1}{\theta}\log\left[1 + \frac{(\exp(-\theta a) - 1)(\exp(-\theta b) - 1)}{\exp(-\theta) - 1}\right]$$

3.4.1 Simulated data

Data setting

We first consider simulated data for the model testing. The covariate vector X is taken as one dimensional, with uniform law on $[0,1]$. On the other hand, the successive times T and U are independent of X and follow log-normal distributions : $\log(T) \sim \mathcal{N}(0,1)$ (resp. $\log(U) \sim \mathcal{N}(0,1)$).

Given a sample $(T_i, U_i)_{1\leq i\leq n}$, let $n_c = \sum_{i=1}^n \sum_{j=1}^n \mathbf{1}_{T_i<T_j, U_i<U_j}$ be the number of concordant pairs in the sample, and $n_d = \sum_{i=1}^n \sum_{j=1}^n \mathbf{1}_{T_i<T_j, U_i>U_j}$ be the number of discordant pairs. Then the Kendall tau Kendall (1938) between T and U expresses as

$$\tau_n = \frac{n_c - n_d}{n_c + n_d}. \tag{3.15}$$

where we can remark that $n_c + n_d = n(n-1)/2$. Its expected value is given by $\tau = 2E\left[\mathbf{1}_{(\tilde{T}_1 - \tilde{T}_2)(\tilde{U}_1 - \tilde{U}_2) > 0}\right] - 1$ where $(\tilde{T}_1, \tilde{U}_1)$ (resp. $(\tilde{T}_2, \tilde{U}_2)$) follows the same law as (T,U) and $(\tilde{T}_1, \tilde{U}_1)$ is independent of $(\tilde{T}_2, \tilde{U}_2)$.

The dependence between T and U in the simulations is set using the Kendall τ through the relation :

$$\log(\tau/1-\tau) = a + b \cdot X \tag{3.16}$$

with $a = -3$ and $b = 4$. Indeed, for any of the four copula families we consider, there is a bijection between the copula parameter θ and the expected value of the Kendall tau Nelsen (2007). Then, making the assumption that the dependence between T and U belongs to some copula family, it is enough to set a value for τ to specify the conditional copula between T and U (see Section 3.4.1).

Let $q = P(T + U > C)$ the censoring rate of the variable $T + U$. The influence of q on the results is studied in the experiments. To do so, the distribution of the censoring variable C, whose log is assumed to follow an exponential distribution, is adjusted so that the desired censoring rate is achieved among the simulated data ($q = 0.3$ or $q = 0.5$).

Description of the experiments

Each simulated dataset consists of $n = 1000$ observations. For each copula family $(\mathcal{C}^l)_{l=1,...,4}$, we estimate the copula parameter θ at five different x values, which correspond to five distinct values $\tau(x)$ and $\theta^l(x)$ that are summarized in Tab. 3.1. The marginal laws of T and U are estimated with the Kaplan-Meier estimator, as well as the survival function of the censoring S_C, used to compute the weights $W_{i,n}$ (see equation (3.6)). Also, we use a quadratic kernel : $K(u) = 15/16 \cdot (1 - u^2)^2 \cdot \mathbf{1}_{|u|\leq 1}$ to localize the estimation in x neighborhoods, and different candidate values for the bandwidth parameter h (see Fig. 3.1). For each bandwidth h, this results in estimators $(\hat{\theta}^l_h(x))_{l=1,..4}$ of the conditional copula parameters at the point x, for the copula families $(\mathcal{C}^l)_{l=1,..,4}$. We can compare them with the exact parameters $(\theta^l(x))_{l=1,..4}$ by computing the quadratic errors at each x value : $\forall\ l = 1, \ldots, 4,\ \epsilon^l_h(x) = (\hat{\theta}^l_h(x) - \theta^l(x))^2$. Then the error of a copula model $\mathcal{C}^l$ with bandwidth h is taken as the average of $\epsilon^l_h(x)$ over the five x values: $\forall\ l = 1, \ldots, 4,\ \epsilon^l_h = 1/5 \sum_{x\in\{x\ values\}} \epsilon^l_h(x)$.

Results

The results of the simulations are shown in Fig. 3.1. We represent the mean values of ϵ^l_h, computed over 100 i.i.d. replications of the above procedure of data simulation and copula fitting, as a function of h. The error is split into a bias part and a variance part, which are known to form an additive decomposition of the total error, and that we also represent in Fig. 3.1.

All the six graphics present the same pattern in a u-shape for the total error. For small h, the bias is low but the high variance of the estimator leads to a bad precision of the estimator overall. The situation is reversed when h takes big values, with low variance

and high bias for the estimator. The optimal value for the bandwidth then has to be taken among middle values of h.

As expected we also remark that the increase of the rate of censoring deteriorates the precision of the estimation.

Although we have shown in Section 3.3 that the estimation procedure we propose is asymptotically consistent, it was important to verify that the method behaves well with finite samples. In this regard, the results of Fig. 3.1 show that, for the four copula families, reasonable errors could be achieved with our method. In the four cases, a value of h equal to 0.15 is a good compromise between the bias and the variance, so that the total error is close from its minimum.

x	0.10	0.30	0.50	0.70	0.90
Kendall tau : $\tau(x)$	0.07	0.14	0.27	0.45	0.65
Gaussian : $\theta^1(x)$	0.11	0.22	0.41	0.65	0.85
Clayton : $\theta^2(x)$	0.15	0.33	0.74	1.64	3.64
Gumbel : $\theta^3(x)$	1.07	1.17	1.37	1.82	2.82
Frank : $\theta^4(x)$	0.62	1.30	2.57	4.90	9.29

Tab. 3.1: Values of the Kendall tau and of the exact copula parameters at the five different x points

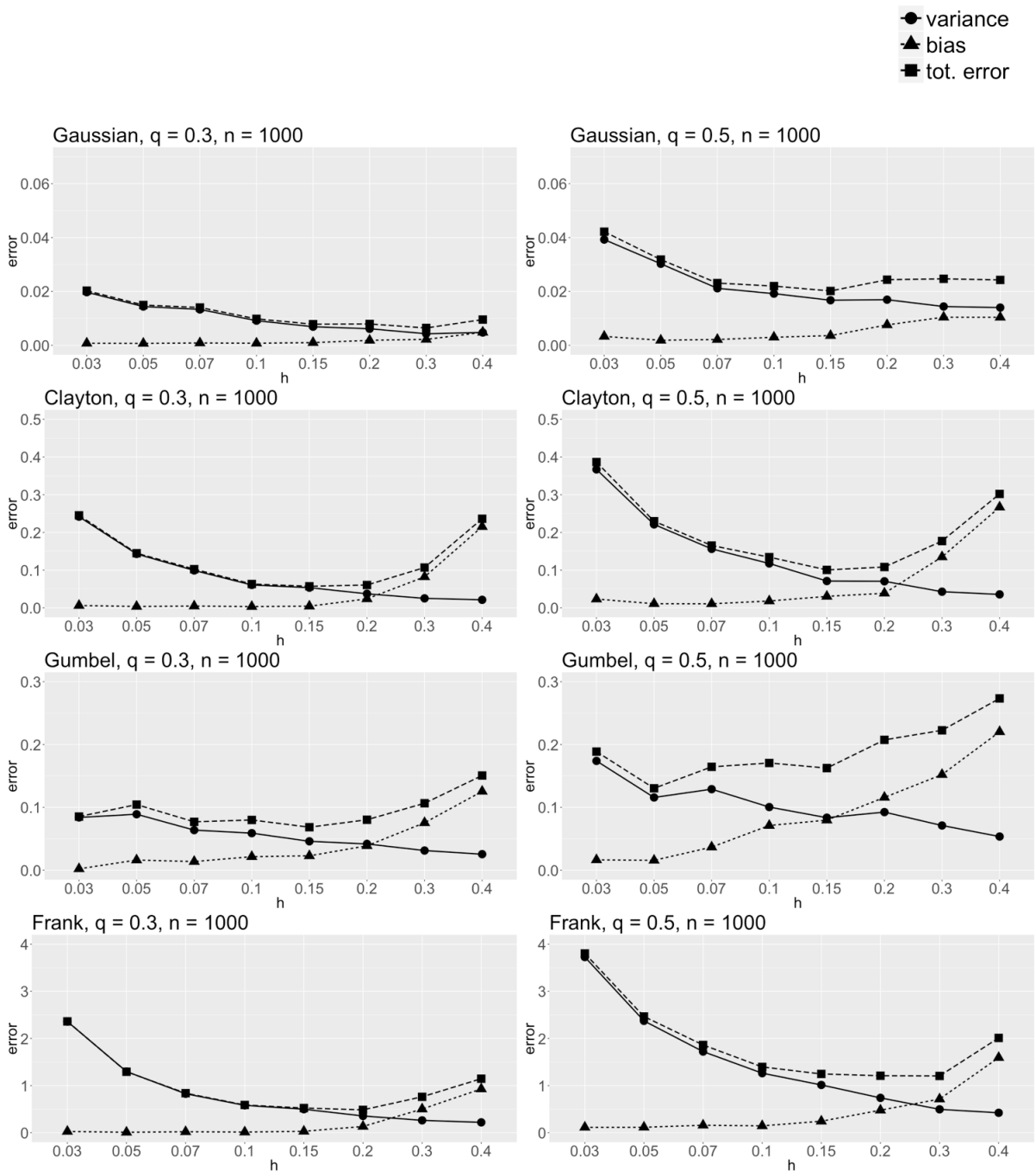

Fig. 3.1: For each copula family $\mathcal{C}^l$ and each rate of censoring $q \in \{0.3, 0.5\}$, mean values of ϵ_h^l (100 repetitions) as a function of h, and their decompositions into bias and variance terms. Size of the simulated datasets: n = 1000.

3.4.2 Application on real data

Description of the data

We have applied our estimation methodology to data provided by a broker of health insurance contracts. In the context of the study, the time T corresponds to the effective time of a contract (i.e. the duration between the date of subscription and the date of effect of the contract), whereas U is the termination time of the contract (i.e. the duration between the date of effect and the date of termination of the contract). As the churn of a contract holder impacts the commission received by a broker for this contract, it is important from the broker's point of view to understand the dependence between the successive times T and U, and especially to measure it given some characteristics X of the contract holder. Indeed, evaluating such dependence allows to fairly compare two underwriters that have been selling insurance products to customers with different characteristics, by weighting the performances of the underwriters with a value factor (that is a factor which gives the expected value of a given customer). This study of the dependence should take into account the censoring that is present in the data, which is due to the fact that any contract may stop to be under observation before T, or U, occurs (e.g. due to the end of the study, or the end of the observation period). In the following, we tackle the problem of the estimation of the dependence between T and U, given the age of the contract holder at the subscription.

The dataset that we study has 224897 entries, recorded from October 1^{st} 2009 to January 15^{th} 2016. We observe that the dataset can be split into two parts according to whether a contract date of effect is on January 1, or not. Indeed, contracts coming into effect on January 1 represent more than one third of the database, and are generally associated with longer delays before the date of effect. For those contracts, the dependence between T and U is stronger. We focus on this part of the database in the following application.

Methodology in the experiments

For the modelling of the dependence between T and U conditionally on X, which is in our situation the age of the contract holder at the subscription, we analyze the results given by the Gaussian, Clayton, Gumbel and Frank parametric copula families (see Section 3.4).

The estimation is made according to the equation (3.5). We use a quadratic kernel as in the simulated data experiments. The survival function of the censoring is estimated

thanks to the estimator of the equation (3.3). Indeed, in this application the censoring variable is the age of the contract (from its subscription to the end of the observation period), hence it is always observed since the perimeter of the study corresponds to contracts that have been subscribed. The pseudo observations $\hat{A}_i$ and $\hat{B}_i$ are derived using local Kaplan-Meier estimators of the marginal distributions of T and U, whose formula are given at equation (3.7).

Assuming no prior on the direction of the dependence between T and U, our fitting procedure for copula needs to be adapted to cases where the sign of the dependence changes as the age of the policyholder varies. This requirement is not problematic for the Gaussian copula and the Frank copula, which may model positive and negative dependences, but the Clayton copula and the Gumbel copula can only model positive dependences. Moreover, the Gaussian copula and the Frank copula are symmetric (i.e. the copula densities satisfy $c_\theta(a,b) = c_\theta(1-b,1-a)$), whereas the Clayton copula and the Gumbel copula are not. Hence, for the latter two copulas, we need to fit the copula four times to cover all possible dependence relations between T and U: we successively fit the copula on the pseudo-observations $(\hat{A}_i, \hat{B}_i)$, $(1-\hat{A}_i, \hat{B}_i)$, $(\hat{A}_i, 1-\hat{B}_i)$ and $(1-\hat{A}_i, 1-\hat{B}_i)$ where $\hat{A}_i$ and $\hat{B}_i$ are defined in Section 3.2.4. This gives four candidate maximums of the criteria $M_{n,h}$ (equation (3.6)), from which we can select the highest maximum. The dependence between T and U is then positive if the maximum corresponds to the case $(\hat{A}_i, \hat{B}_i)$ or $(1-\hat{A}_i, 1-\hat{B}_i)$, and negative otherwise.

In the following numerical experiments, we use a train-test approach with 100 repetitions. For each of the 100 iterations, two non-overlapping subsamples $\mathcal{D}_{tr}$ (train) and $\mathcal{D}_{te}$ (test), of size 10000 ($n_{tr} = n_{te} = 10000$), are drawn from the initial dataset. On the training set, and for each age x from 20 to 80, we apply our method to compute $(\hat{\theta}^l_{h,tr}(x))_{l=1,..4}$, the estimates of the conditional copula parameters corresponding to the copula families $(\mathcal{C}^l)_{l=1,..,4}$ and to the values $h = 1, 5, 10, 20, 40$. On the test set, we estimate the conditional Kendall tau between T and U, at each age x. This is done using the following kernel estimator of the Kendall tau : for all test observations i, let $W_{i,n}(x) = W_{i,n}K((X_i - x)/h_1)$, where $W_{i,n}$ is defined at equation (3.6), $(X_i)_{i=1,..,n_{te}}$ denotes the age values and $h_1 = 1$. Then let $n_c^{(w,x)} = \sum_{i=1}^{n_{te}} \sum_{j=1}^{n_{te}} W_{i,n}(x) W_{j,n}(x) \mathbf{1}_{T_i<T_j, U_i<U_j}$ and $n_d^{(w,x)} = \sum_{i=1}^{n_{te}} \sum_{j=1}^{n_{te}} W_{i,n}(x) W_{j,n}(x) \mathbf{1}_{T_i<T_j, U_i>U_j}$; we can estimate the conditional Kendall tau at the age x by $\hat{\tau}^w_{h_1,te}(x) = (n_c^{(w,x)} - n_d^{(w,x)})/(n_c^{(w,x)} + n_d^{(w,x)})$. The conditional Kendall tau is also estimated on the training set $(\hat{\tau}^w_{h_1,tr}(x))$, using the same method.

To compare the results of the different copula families and to identify the optimal value

for the bandwidth parameter h, we compute train and test errors based on Kendall tau estimates. Thanks to the one to one relations between the parameter θ and the Kendall tau for the copula families $(\mathcal{C}^l)_{l=1,..,4}$ (see Section 3.4.1), we deduce from the estimators $(\hat{\theta}^l_{h,tr}(x))_{l=1,..4}$ train estimates of the conditional Kendall tau $(\hat{\tau}^l_{h,tr}(x))_{l=1,..4}$. Then we define the test (resp. train) error at a point x as $\epsilon^l_{h,te}(x) = (\hat{\tau}^l_{h,tr}(x) - \hat{\tau}^w_{h_1,te}(x))^2$ (resp. $\epsilon^l_{h,tr}(x) = (\hat{\tau}^l_{h,tr}(x) - \hat{\tau}^w_{h_1,tr}(x))^2$), and thereafter the test (resp. train) error of the copula model as the aggregated error over all x values : $\epsilon^l_{h,te} = \sum_{x=20}^{80} \bar{w}_{x,te}\epsilon^l_{h,te}(x)$ (resp. $\epsilon^l_{h,tr} = \sum_{x=20}^{80} \bar{w}_{x,tr}\epsilon^l_{h,tr}(x)$), with $\bar{w}_{x,te} = w_{x,te}/\sum_{x=20}^{80} w_{x,te}$ and $w_{x,te} = \sum_{i\in\mathcal{D}_{te}} K((X_i - x)/h_1)$ (resp. $\bar{w}_{x,tr} = w_{x,tr}/\sum_{x=20}^{80} w_{x,tr}$ and $w_{x,tr} = \sum_{i\in\mathcal{D}_{tr}} K((X_i - x)/h_1)$).

Results

We represent in Fig. 3.2 box plots of the train and test square root errors $((\epsilon^l_{h,tr})^{1/2})_{l=1,..,4}$ and $((\epsilon^l_{h,te})^{1/2})_{l=1,..,4}$ measured over the 100 iterations, for each copula family and each bandwidth value h. The results show that the Frank copula achieves the lowest error (both train and test) on the data, and should be privileged to model the dependence between the two durations. Also, we observe that the test error is generally minimal for $h = 20$, which indicates that 20 years is the appropriate time scale to observe trends in the conditional dependence.

On Fig. 3.3, we show the average values of the conditional copula parameters for the fitted copulas, as well as 95% confidence intervals for the exact parameters. The graphic for the Frank copula indicates that the strength of the dependence between T and U decreases between the ages 20 and 40, and then increases from 40 to 65 (i.e. until the age of retirement), before it starts to decrease again. This means that young adults (20-30) and seniors (55-75) are more likely than other age categories to have their decision to terminate their contract impacted by the effective time of the contract, in the sens that a long effective time causes an higher probability of rapid termination of the contract.

The Fig. 3.4 shows that a different evolution of the dependence is observed for the contracts associated with higher product's level (i.e. better guarantees). These contracts exhibit a strong negative dependence overall, which becomes even stronger for the people aged over 60.

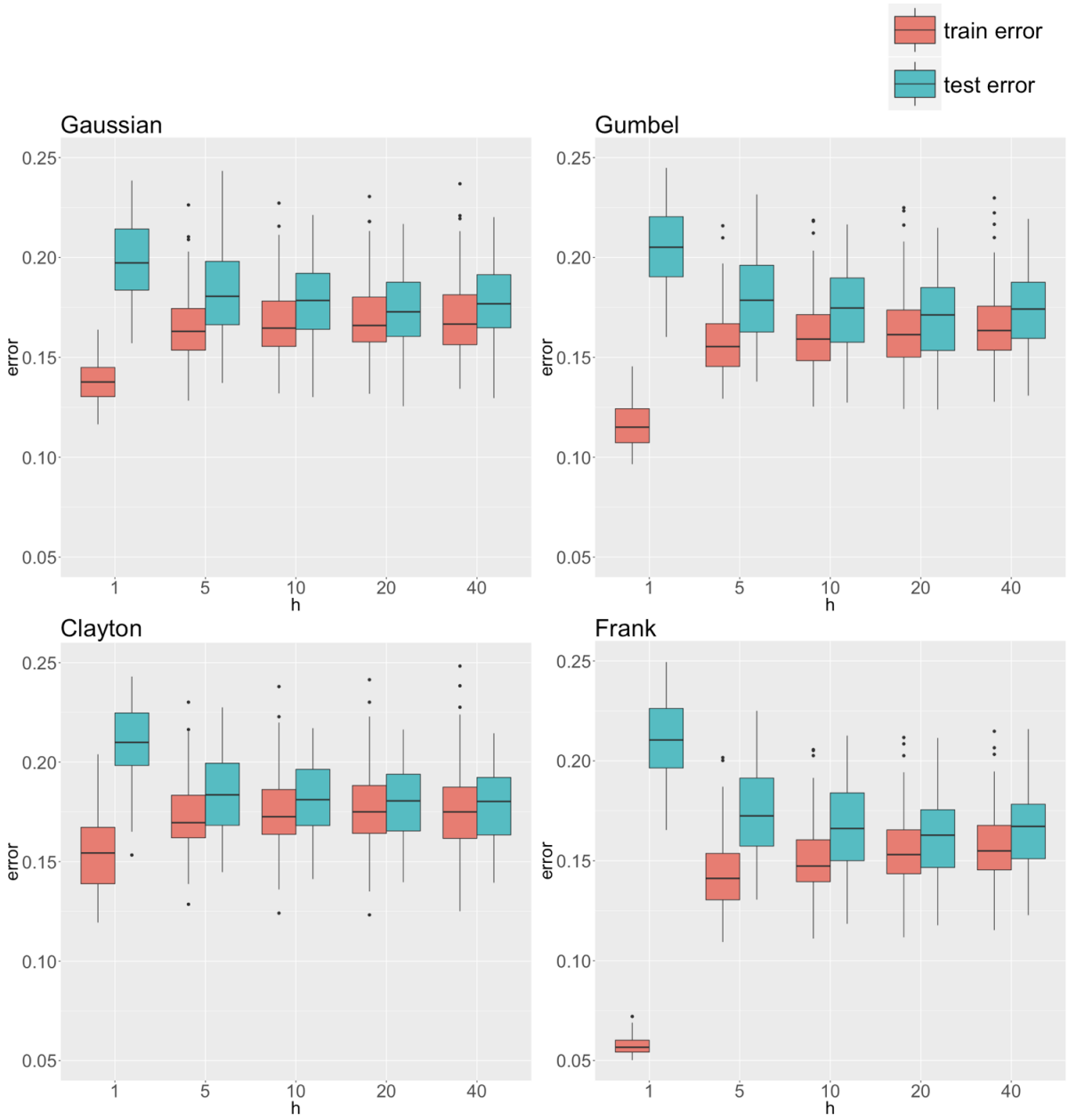

Fig. 3.2: For each copula family and each bandwidth value h, box plot of the train and test square root errors $((\epsilon^l_{h,tr})^{1/2})_{l=1,..,4}$ and $((\epsilon^l_{h,te})^{1/2})_{l=1,..,4}$ ($n = 10000$, 100 repetitions).

3.5 Conclusion

In this paper, we proposed a methodology for estimating a conditional copula function under random censoring, when the two variables linked through the copula are successive times. The model is semiparametric, since we assume that the conditional copula does not leave a parametric family, but with a nonparametric assumption on the dependence of the

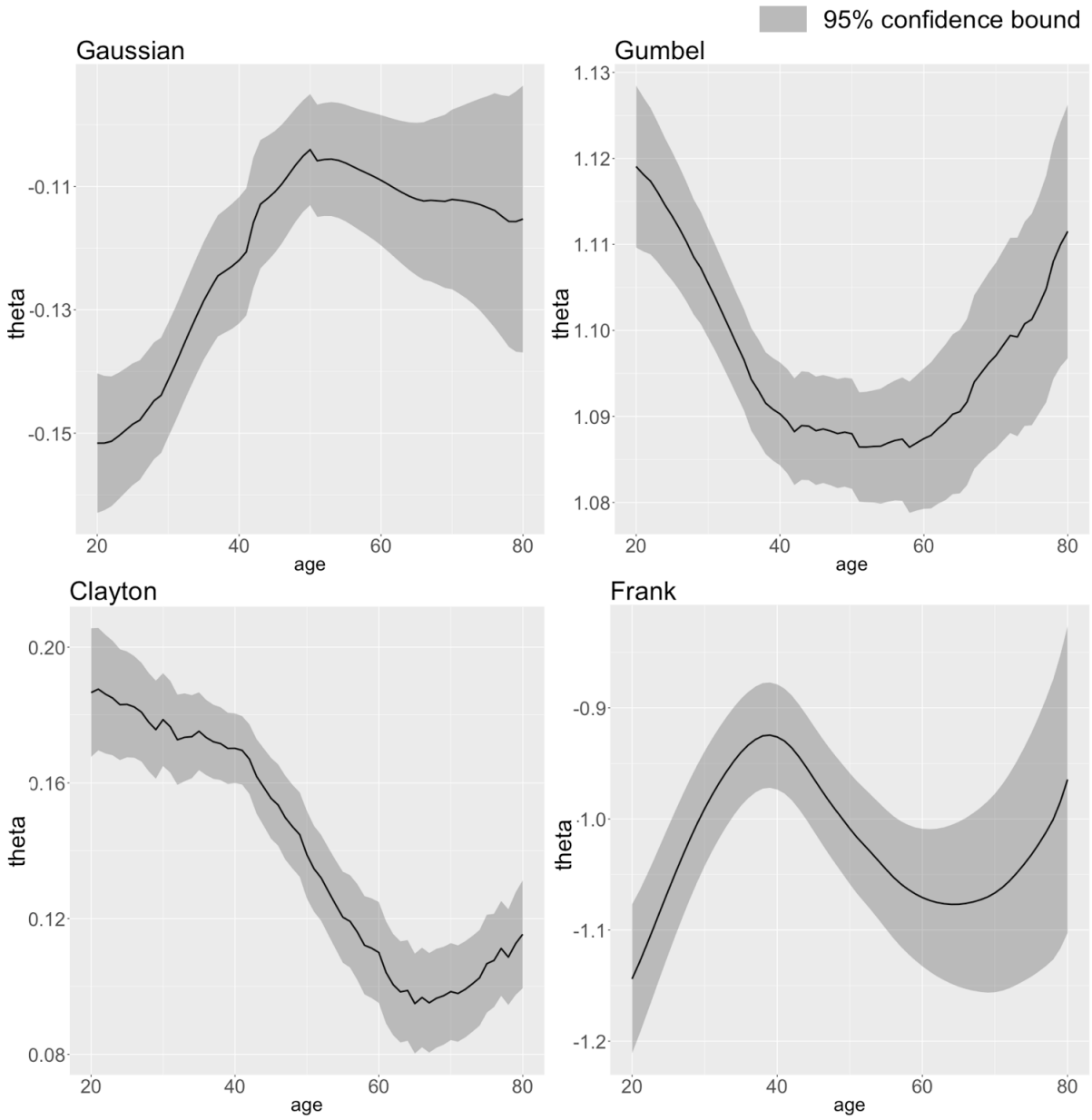

Fig. 3.3: For each copula family, mean value of the conditional copula parameter as a function of the age x ($h = 20$, $n = 10000$, 100 repetitions). As we notice in Section 3.4.2, the Gumbel copula and the Clayton copula don't vary in the same direction as the Gaussian copula and the Frank copula.

association parameter on the covariates. From a numerical point of view, the procedure is simple, since it relies on a weighted log-likelihood approach. The kernel smoothing approach can be extended to local linear modeling, as in Gijbels et al. (2011) in presence of complete data. Let us mention that our results hold with only standard conditions on the estimation of the margins, giving a relative freedom to practitioners on how they

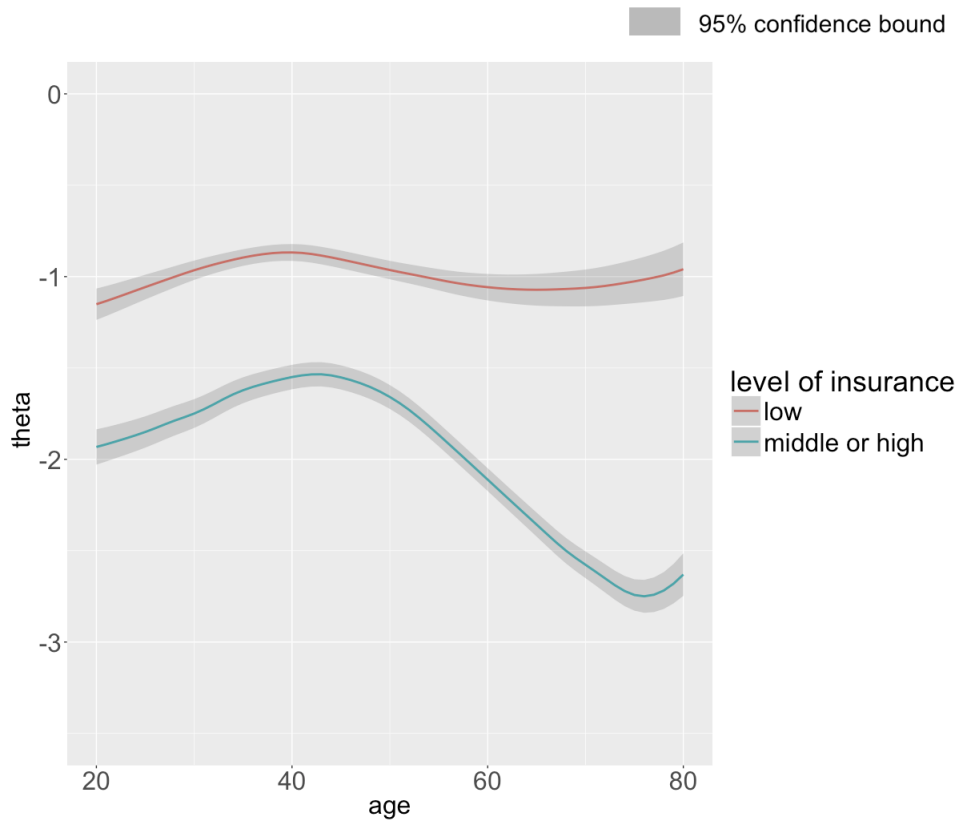

Fig. 3.4: Impact of the variable *level of insurance* on the conditional dependence between T and U, given the age of the prospect (Frank copula, $h = 20$, 100 repetition).

want to perform this estimation. Moreover, we provide conditions on the censoring which allow to understand the behavior of the method even in the tail of the distribution (that is near the right and upper corner of the unit square when looking at the copula). This indication is precious since under random censoring an important question is to control the behavior of the method near the right tail.

Code

The code used for producing the results is available at the address : `github.com/YohannLeFaou/copula-successive-duration-times`.

Acknowledgements

We would like to thank the company Forsides that supports us in this project, and the company Santiane who provided the data that served for the experiments.

3.6 Appendix: technical results

This appendix section gathers the proofs of Theorems 3.1 (Section 3.6.1) and 3.2 (Sections 3.6.2 to 3.6.7). The auxiliary results required to prove Theorem 3.2 consist of replacing the weights $W_{i,n}$ by W_i in the criterion $M_{n,h}$ (Section 3.6.3), dealing with the trimming function (Section 3.6.4), replacing the pseudo-observations by their limit values (Section 3.6.5), and providing a control of the stochastic term uniformly in x (Section 3.6.6).

The uniform consistency result is then obtained by applying the results of Einmahl & Mason (2005) on kernel smoothing (Section 3.6.7).

3.6.1 Proof of Theorem 3.1 (Bias term)

Let

$$M_f^{(c)}(x,\theta) \;=\; \frac{1}{h^d} E\left[K\left(\frac{X-x}{h}\right)\log c_\theta(F_T(T|X),F_U(U|X))\right].$$

First observe that

$$\int K(v)\phi(F_T(t|x+hv),F_U(u|x+hv),\theta)dF(t,u|x+hv)f_X(x+hv)dv = \int K(v)\phi(a,b,\theta)\mathfrak{c}(a,b|x+hv)f_X(x+hv)dvdadb.$$

The right-hand side converges towards

$$\int \phi(a,b,\theta)\mathfrak{c}(a,b|x)dadbf_X(x) = M(x,\theta)f_X(x) \tag{3.17}$$

as $h \to 0$, uniformly in θ and x, from Lebesgue's dominated convergence theorem and Assumption 0.3. Thus,

$$\sup_{x\in\mathcal{X}} \|\theta_h^*(x)-\theta(x)\| \underset{h\to 0}{=} o(1). \tag{3.18}$$

Next, we use a Taylor expansion to show the speed of convergence of $\sup_x \|\theta_h^*(x)-\theta(x)\|$. For all $j=1,\ldots,k$, let $\nabla_{\theta_j} M_f^{(c)}(x,\theta)$ the j^{th} component of the gradient $\nabla_\theta M_f^{(c)}(x,\theta)$. Since $\nabla_\theta M_f^{(c)}(x,\theta_h^*(x))=0$, we have

$$\forall j=1,\ldots,k,\ \langle \nabla_\theta\nabla_{\theta_j} M_f^{(c)}(x,\tilde{\theta}_h^j(x)),\theta_h^*(x)-\theta(x)\rangle = -\nabla_{\theta_j} M_f^{(c)}(x,\theta(x)), \tag{3.19}$$

with $\forall j,\ \tilde{\theta}_h^j(x) \in [\theta_h^*(x); \theta(x)]$. For the left hand side of (3.19) we have,

$$\begin{aligned}\|\nabla_\theta \nabla_{\theta_j} M_f^{(c)}(x, \tilde{\theta}_h^j(x)) - \nabla_\theta \nabla_{\theta_j} M(x, \theta(x)) f_X(x)\| \\ \leq \|\nabla_\theta \nabla_{\theta_j} M_f^{(c)}(x, \tilde{\theta}_h^j(x)) - \nabla_\theta \nabla_{\theta_j} M_f^{(c)}(x, \theta(x))\| \\ + \|\nabla_\theta \nabla_{\theta_j} M_f^{(c)}(x, \theta(x)) - \nabla_\theta \nabla_{\theta_j} M(x, \theta(x)) f_X(x)\|.\end{aligned}$$

Clearly $\sup_x \|\nabla_\theta \nabla_{\theta_j} M_f^{(c)}(x, \tilde{\theta}_h^j(x)) - \nabla_\theta \nabla_{\theta_j} M_f^{(c)}(x, \theta(x))\| = o(1)$ by (3.18) and the smoothness condition of ϕ in Assumption 0.3.

Moreover, $\sup_x \|\nabla_\theta \nabla_{\theta_j} M_f^{(c)}(x, \theta(x)) - \nabla_\theta \nabla_{\theta_j} M(x, \theta(x)) f_X(x)\| = o(1)$ using the same kind of development as in (3.17) (applied to $\nabla_\theta \nabla_{\theta_j} \phi(a, b, \theta)$ instead of $\phi(a, b, \theta)$), and Lebesgue's theorem.

Hence, one gets

$$\sup_x \|\nabla_\theta \nabla_{\theta_j} M_f^{(c)}(x, \tilde{\theta}_h^j(x)) - \nabla_\theta \nabla_{\theta_j} M(x, \theta(x)) f_X(x)\| \underset{h \to 0}{=} o(1),$$

so that using Assumption 0.5 and combining the left side of the k equations of (3.19), we have for h sufficiently small,

$$\forall x \in \mathcal{X},\ \sum_{j=1}^{k} \left| \langle \nabla_\theta \nabla_{\theta_j} M_f^{(c)}(x, \tilde{\theta}_h^j(x)), \theta_h^*(x) - \theta(x) \rangle \right| \geq c_0' \|\theta_h^*(x) - \theta(x)\|, \tag{3.20}$$

with $c_0' > 0$ a given constant.

Moreover, we have for the right hand side of (3.19),

$$\forall j,\ \sup_{x \in \mathcal{X}} \left| \nabla_{\theta_j} M_f^{(c)}(x, \theta(x)) \right| = O(h^2). \tag{3.21}$$

Indeed, a second order Taylor expansion leads to

$$\nabla_\theta M_f^{(c)}(x, \theta(x)) = \frac{h^2}{2} \int K(v) \dot{\phi}(a, b, \theta(x)) \langle \nabla_x^2 \{\mathfrak{c}(a, b|\tilde{x}) f_X(\tilde{x})\} \cdot v, v \rangle dv da db,$$

for some $\tilde{x}$ between x and $x + hv$, and the right-hand side is $O(h^2)$ uniformly in x thanks to Assumption 0.4.

Combining equations (3.19), (3.20) and (3.21) then leads to

$$\sup_x \|\theta_h^*(x) - \theta(x)\| \underset{h \to 0}{=} O(h^2).$$

3.6.2 Consistency of the Stochastic term

Before showing the convergence rate of $\hat{\theta}_h(x)$, we first show its uniform consistency in Proposition 3.1, by looking at its difference with the bias term $\theta_h^*(x)$.

Proposition 3.1 *Under the assumptions of Theorem 3.2,*

$$\sup_x \|\hat{\theta}_h(x) - \theta_h^*(x)\| \underset{n\to\infty}{=} o_P(1). \tag{3.22}$$

Proof. To show (3.22), first decompose

$$|M_{n,h}(x,\theta) - M_f^{(c)}(x,\theta)| \leq |M_{n,h}(x,\theta) - M_{n,h}^*(x,\theta)| + |M_{n,h}^*(x,\theta) - M_f^{(c)}(x,\theta)|,$$

where

$$M_{n,h}^*(x,\theta) = \frac{1}{nh^d}\sum_{i=1}^n W_{i,n} K\left(\frac{X_i - x}{h}\right) \log c_\theta(A_i, B_i)\hat{\omega}_i(\nu_n). \tag{3.23}$$

Let $\beta_0 \in]1; p'/(2-p')[$. From Lemma 3.1 and Lemma 3.2,

$$M_{n,h}^*(x,\theta) = \frac{1}{nh^d}\sum_{i=1}^n W_i K\left(\frac{X_i - x}{h}\right) \log c_\theta(A_i, B_i) + O_P(n^{\max(1/p' - 1/(2\beta_0),0) - 1/2}) + o_P(1),$$

uniformly in x and θ. Indeed, remark that the convergence rate in Lemma 3.2 is $o_P(1)$. Moreover, β_0 satisfies $1/p' - 1/(2\beta_0) - 1/2 < 0$, so that $n^{1/p' - 1/(2\beta_0) - 1/2} = o(1)$.

Next, from equation (3.14), let $\beta_1 > 1$ such that

$$\lim_{\delta \searrow 0} \limsup_{n\to\infty} \ \varepsilon_n n^{1/p'' - 1/(2\beta_1)} = 0. \tag{3.24}$$

From Lemma 3.3 applied with β_1, we get

$$\sup_{x\in\mathcal{X}, \theta\in\Theta} |M_{n,h}(x,\theta) - M_{n,h}^*(x,\theta)| = O_P(\varepsilon_n n^{\max(1/p'' - 1/(2\beta_1),0)})$$

so that $\sup_{x\in\mathcal{X},\theta\in\Theta} |M_{n,h}(x,\theta) - M_{n,h}^*(x,\theta)| = o_P(1)$ thanks to equation (3.24).

Then, Theorem 4 of Einmahl & Mason (2005) applies using Assumption 0.6 to show that $\sup_{x\in\mathcal{X},\theta\in\Theta} |\frac{1}{nh^d}\sum_{i=1}^n W_i K\left(\frac{X_i - x}{h}\right) \log c_\theta(A_i, B_i) - M_f^{(c)}(x,\theta)| = o_P(1)$, which concludes the proof. ■

3.6.3 Estimation of S_C

Lemma 3.1 below shows that, provided that an integrability condition holds, the weights $W_{i,n}$ (relying on the estimation of $\hat{S}_C$ the survival function of the censoring) are asymptotically equivalent to the weights W_i.

Lemma 3.1 *Let $\mathcal{F}$ denote a class of functions ψ such that, $\forall \psi \in \mathcal{F}$, $|\psi(y,z,x)| \leq \Psi(y,z,x)$. Assume that for some $p' > 1$,*

$$\sup_{x\in\mathcal{X}} E\left[\frac{\Psi(T,U,X)^{p'}}{S_C(T+U)^{2p'-1}}\Big|X=x\right] < \infty, \tag{3.25}$$

and that $\liminf_{n\to\infty} nh^{2d} > 0$.
Then, for any $\beta > 1$,

$$\sup_{x\in\mathcal{X},\psi\in\mathcal{F}}\left|\frac{1}{nh^d}\sum_{i=1}^{n}(W_{i,n}-W_i)\psi(Y_i,Z_i,X_i)K\left(\frac{X_i-x}{h}\right)\hat{\omega}_i(\nu_n)\right| = O_P\left(n^{\max(1/p'-1/(2\beta),0)-1/2}\right). \tag{3.26}$$

Proof. Let $\mathfrak{S}_{(1)} \leq \mathfrak{S}_{(2)} \leq ... \leq \mathfrak{S}_{(n)}$ denote the order statistics of $\mathfrak{S}_i = Y_i + Z_i$. Observe that

$$\begin{aligned}\left|\frac{1}{nh^d}\sum_{i=1}^{n}(W_{i,n}-W_i)\psi(Y_i,Z_i,X_i)K\left(\frac{X_i-x}{h}\right)\hat{\omega}_i(\nu_n)\right| &\\ \leq \sup_{t\leq\mathfrak{S}_{(n)}}\left|\hat{S}_C(t)-S_C(t)\right|\cdot\sup_{t\leq\mathfrak{S}_{(n)}}\left|\frac{S_C(t)}{\hat{S}_C(t)}\right| &\\ \times\frac{1}{nh^d}\sum_{i=1}^{n}\frac{W_i|\Psi(Y_i,Z_i,X_i)|K\left(\frac{X_i-x}{h}\right)}{S_C(Y_i+Z_i)}.&\end{aligned}$$

First notice that $\sup_{t\leq\mathfrak{S}_{(n)}}|\hat{S}_C(t)-S_C(t)| = O_P(n^{-1/2})$ and $\sup_{t\leq\mathfrak{S}_{(n)}} S_C(t)\hat{S}_C(t)^{-1} = O_P(1)$ (see Shorack & Wellner (2009) when $\hat{S}_C$ is the empirical survival function (3.3), and Gill (1983) when $\hat{S}_C$ is the Kaplan-Meier estimator (3.4)).

Then, since

$$E\left[\frac{W_i^{p'}\Psi(Y_i,Z_i,X_i)^{p'}}{S_C(Y_i+Z_i)^{p'}}\right] = E\left[\frac{\Psi(T_i,U_i,X_i)^{p'}}{S_C(T_i+U_i)^{2p'-1}}\right],$$

we get from Lemma 3.4 (ii) and Lemma 3.5 (ii) that for $\beta > 1$,

$$\sup_{x\in\mathcal{X}}\left|\frac{1}{nh^d}\sum_{i=1}^{n}\frac{W_i\Psi(Y_i,Z_i,X_i)K\left(\frac{X_i-x}{h}\right)}{S_C(Y_i+Z_i)}\right| = O_P\left(n^{\max(1/p'-1/(2\beta),0)}\right).$$

This concludes the proof. ■

3.6.4 Trimming function

To control the potential erratic behavior of $\hat{A}_i$ and $\hat{B}_i$ close to the border of the unit square, we introduced some trimming $\hat{\omega}_i(\nu_n)$. Lemma 3.2 then shows the consistency of this trimming approach.

Lemma 3.2 *Let $\mathcal{F}$ denote a class of functions ψ such that $\forall \psi \in \mathcal{F}$, $|\psi(y,z,x)| \leq \Psi(y,z,x)$. Assume that, for some $p > 2$,*

$$\sup_{x \in \mathcal{X}} E\left[\frac{\Psi(T,U,X)^p}{S_C(T+U)^{p-1}}\Big| X = x\right] < \infty, \tag{3.27}$$

with $\liminf_{n\to\infty} nh^{\frac{d}{1-2/p}}[\log n]^{-1} > 0$, and let

$$\eta_n := \sup_{x \in \mathcal{X}} E\left[\frac{\Psi(T,U,X)^p}{S_C(T+U)^{p-1}}(1 - w^*(2\nu_n))\Big| X = x\right].$$

Then

$$\sup_{x\in\mathcal{X},\psi\in\mathcal{F}}\left|\frac{1}{nh^d}\sum_{i=1}^n W_i K\left(\frac{X_i - x}{h}\right)\psi(Y_i,Z_i,X_i)(1-\hat{w}_i(\nu_n))\right| = O_P(\nu_n^{1-1/p} + n^{-1/2}h^{-d/2}[\log n]^{1/2}\eta_n^{1/p}).$$

Proof.
Let $E_n(M) = \{\sup_{1\leq i\leq n} |\hat{A}_i - A_i| + |\hat{B}_i - B_i| \leq M\varepsilon_n\}$. On $E_n(M)$ and for n large enough, we have $1 - \hat{\omega}_i(\nu_n) \leq 1 - w_i(2\nu_n)$, with

$$w_i(2\nu_n) = \mathbf{1}_{\min(A_i,B_i,1-A_i,1-B_i)\geq 2\nu_n}. \tag{3.28}$$

Additionally, defining for a sequence η_n

$$w_i^*(\eta_n) = \mathbf{1}_{\min(F_T(T_i|X_i),F_U(U_i|X_i),1-F_T(T_i|X_i),1-F_U(U_i|X_i))\geq\eta_n}, \tag{3.29}$$

we first note that

$$E\left[W_i^p\Psi(Y_i,Z_i,X_i)^p(1-w_i(2\nu_n))|X_i = x\right] = E\left[\frac{\Psi(T_i,U_i,X_i)^p}{S_C(T_i+U_i)^{p-1}}(1-w_i^*(2\nu_n))\Big| X_i = x\right], \tag{3.30}$$

so that

$$\sup_{x\in\mathcal{X}} E\left[W_i^p \Psi(Y_i, Z_i, X_i)^p (1 - w_i(2\nu_n))|X_i = x\right] = O(\eta_n). \tag{3.31}$$

Second, we have

$$\begin{aligned} E\left[W_i \Psi(Y_i, Z_i, X_i)(1 - w_i(2\nu_n))|X_i = x\right] &= E\left[\Psi(T_i, U_i, X_i)(1 - w_i^*(2\nu_n))|X_i = x\right] \\ &\leq \sup_{x\in\mathcal{X}} E[\Psi(T_i, U_i, X_i)^p|X_i = x]^{1/p} \times \\ &\quad \sup_{x\in\mathcal{X}} E[(1 - w_i^*(2\nu_n))|X_i = x]^{1-1/p} \end{aligned}$$

with

$$\begin{aligned} E[(1 - w_i^*(2\nu_n))|X_i = x] &\leq 4P(T_i \in [0; 2\nu_n[, U_i \in [0; 1]|X_i = x) \\ &\leq 8\nu_n. \end{aligned}$$

We can then apply the Lemma 3.4 (i) to prove that

$$\begin{aligned} \sup_{x\in\mathcal{X},\psi\in\mathcal{F}} \left| \frac{1}{nh^d} \sum_{i=1}^n W_i K\left(\frac{X_i - x}{h}\right) \Psi(Y_i, Z_i, X_i)(1 - w_i(2\nu_n))) \right| \\ = O_P(\nu_n^{1-1/p} + n^{-1/2} h^{-d/2} [\log n]^{1/2} \eta_n^{1/p}). \end{aligned}$$

To conclude, we observe that $\limsup_{M\to\infty} \lim_{n\to\infty} \mathbb{P}(E_n(M)) = 1$ from Assumption 0.8. ■

3.6.5 Pseudo-observations

The aim of this section is to show that the pseudo-observations $\hat{A}_i$ and $\hat{B}_i$ can be asymptotically replaced by A_i and B_i.

Lemma 3.3 *Let $M_{n,h}^*(x, \theta)$ as defined in (3.23). Then under Assumptions 0.7 and 0.8, and the supplementary condition $\liminf_{n\to\infty} nh^{2d} > 0$, we have:*

For any $\beta > 1$,

$$\sup_{x\in\mathcal{X},\theta\in\Theta} |M_{n,h}(x, \theta) - M_{n,h}^*(x, \theta)| = O_P\left(\varepsilon_n n^{\max(1/p'' - 1/(2\beta), 0)}\right).$$

Proof. We have, from a Taylor expansion and Assumption 0.7,

$$\begin{aligned}|M_{n,h}(x,\theta) - M^*_{n,h}(x,\theta)| &= \frac{1}{nh^d}\sum_{i=1}^n W_{i,n} K\left(\frac{X_i - x}{h}\right)(\log c_\theta(\hat{A}_i, \hat{B}_i) - \log c_\theta(A_i, B_i))\hat{w}_i(\nu_n) \\ &\leq \frac{1}{nh^d}\sum_{i=1}^n W_{i,n} K\left(\frac{X_i - x}{h}\right)\tilde{r}_1(\tilde{A}_i) r_2(\tilde{B}_i)|\hat{A}_i - A_i|\hat{\omega}_i(\nu_n) \\ &\quad + \frac{1}{nh^d}\sum_{i=1}^n W_{i,n} K\left(\frac{X_i - x}{h}\right) r_1(\tilde{A}_i)\tilde{r}_2(\tilde{B}_i)|\hat{B}_i - B_i|\hat{\omega}_i(\nu_n),\end{aligned}$$

where for all $i = 1, \ldots, n$, $\tilde{A}_i \in [A_i, \hat{A}_i]$ (resp. $\tilde{B}_i \in [B_i, \hat{B}_i]$).

To control the two terms in this last expression, the problems arise when $\tilde{A}_i$ and/or $\tilde{B}_i$ are close to 1 or 0. We explain how to study the case $\tilde{A}_i$ close to 0 since the other ones are similar. Therefore, we consider the case where both $\hat{A}_i$ and A_i are less than $1/2$.

If $\hat{A}_i \geq A_i$, then $\tilde{r}_1(\tilde{A}_i) \leq \tilde{r}_1(A_i)$ and $r_1(\tilde{A}_i) \leq r_1(A_i)$ by Assumption 0.7. To treat the case $\hat{A}_i \leq A_i$, consider that we are on the event $E_n(M) = \{\sup_{1\leq i\leq n} |\hat{A}_i - A_i| + |\hat{B}_i - B_i| \leq M\varepsilon_n\}$. Then, when $\hat{\omega}_i(\nu_n) = 1$ and for n large enough, $\hat{A}_i \geq A_i/2$ (indeed, note that $A_i \leq \nu_n + M\varepsilon_n$ when $\hat{\omega}_i(\nu_n) = 1$). Hence, from $\hat{A}_i \geq A_i/2$, $\tilde{r}_1(\tilde{A}_i) \leq C\tilde{r}_1(A_i)$ and $r_1(\tilde{A}_i) \leq Cr_1(A_i)$ for some constant C using the reproducibility property of the u-shaped functions in Assumption 0.7.

Then, on $E_n(M)$ and for n large enough we have that $|M_{n,h}(x,\theta) - M^*_{n,h}(x,\theta)|$ is bounded by

$$\mathcal{T}_1 := \frac{C}{nh^d}\sum_{i=1}^n W_{i,n} K\left(\frac{X_i - x}{h}\right)\left(\tilde{r}_1(A_i) r_2(B_i)|\hat{A}_i - A_i| + r_1(A_i)\tilde{r}_2(B_i)|\hat{B}_i - B_i|\right).$$

Moreover, noting that $W_{i,n} \leq W_i \cdot \sup_{t\leq \mathfrak{S}_{(n)}} \left|\frac{S_C(t)}{\hat{S}_C(t)}\right|$, and that $|\hat{A}_i - A_i| \leq M\varepsilon_n$ on $E_n(M)$, we have

$$\mathcal{T}_1 \leq \frac{CM\epsilon_n}{nh^d} \cdot \sup_{t\leq \mathfrak{S}_{(n)}} \left|\frac{S_C(t)}{\hat{S}_C(t)}\right| \cdot \sum_{i=1}^n W_i K\left(\frac{X_i - x}{h}\right)(\tilde{r}_1(A_i) r_2(B_i) + r_1(A_i)\tilde{r}_2(B_i)),$$

with $\sup_{t\leq \mathfrak{S}_{(n)}} |S_C(t)/\hat{S}_C(t)| = O_P(1)$.

We then apply Lemma 3.4 (ii) and Lemma 3.5 (ii) (using Assumption 0.7) to obtain that

for any $\beta > 1$,

$$\frac{1}{nh^d} \sup_{x\in\mathcal{X}} \left| \sum_{i=1}^{n} W_i K\left(\frac{X_i - x}{h}\right) (\tilde{r}_1(A_i) r_2(B_i) + r_1(A_i)\tilde{r}_2(B_i)) \right| = O_P(n^{\max(1/p''-1/(2\beta),0)}),$$

so that $\mathcal{T}_1 = O_P\left(\varepsilon_n n^{\max(1/p''-1/(2\beta),0)}\right)$.

This concludes the proof since we have $\limsup_{M\to\infty} \lim_{n\to\infty} \mathbb{P}(E_n(M)) = 1$ from Assumption 0.8.

■

3.6.6 Uniform rate of convergence of the stochastic term

In this section, we use a result from Einmahl & Mason (2005) to obtain uniform rates of convergence for our estimator. Lemma 3.4 below is a direct consequence of Proposition 1 in Einmahl & Mason (2005). Under a condition of moment of order $2\alpha > 2$, it provides uniform asymptotic bounds for sums of i.i.d. variables of the form of $\mathfrak{K}_n(x)$ (see below in equation (3.32)). The cases (i) and (ii) in Lemma 3.4 give two bounds for $\mathfrak{K}_n(x)$ according to the strength of the assumption we make on the speed of h. The Lemma 3.5 is a corollary of the Lemma 3.4 and gives weaker bounds when we only assume a moment of order $2\alpha \in [1/2; 1]$.

Lemma 3.4 *Let*

$$\mathfrak{K}_n(x) = \sum_{i=1}^{n} V_{i,n} K\left(\frac{X_i - x}{h}\right), \tag{3.32}$$

for $(V_{i,n})_{i=1,\dots,n}$ a sequence of i.i.d. positive random variables having the same distribution as a variable V_n.

Let $\|K_n\|_{\mathcal{X}} = \sup_{x\in\mathcal{X}} |\mathfrak{K}_n(x)|$. Assume that $\sup_{x\in\mathcal{X}} E[V_n^{2\alpha}|X = x] = O(\eta_n)$ for some $\alpha > 1$ and let $\tilde{\eta}_n := \sup_{x\in\mathcal{X}} E[V_n|X = x]$. We have the two following results:

(i) Assume that $\liminf_{n\to\infty} nh^{\frac{d}{1-1/\alpha}} [\log n]^{-1} > 0$, then

$$\|K_n\|_{\mathcal{X}} = O_P(nh^d\tilde{\eta}_n + n^{1/2}h^{d/2}[\log n]^{1/2}\eta_n^{1/(2\alpha)}).$$

(ii) Assume that $\liminf_{n\to\infty} nh^{2d} > 0$, then

$$\|K_n\|_{\mathcal{X}} = O_P(nh^d\eta_n^{1/(2\alpha)}).$$

Let us observe that the bound in (i) and (ii) is slightly different. Indeed, $\tilde{\eta}_n \leq \eta_n^{1/(2\alpha)}$ but can be significantly smaller in some cases. More precisely, in our use of Lemma 3.4 (where $2\alpha = p$, with $p > 2$), $\tilde{\eta}_n = O(\nu_n^{1-1/p})$ while $\eta_n^{1/p} \geq \nu_n^{1/p}$.

Proof.

• Proof of (i):

Let $(\varepsilon_1, ..., \varepsilon_n)$ denote i.i.d. Rademacher variables, i.e. $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 1/2$, that are assumed to be independent from the sample $(V_{i,n}, X_i)_{i=1,...,n}$.

Let $\gamma_n = (n\eta_n/\log(n))^{1/2\alpha}$ and define

$$\begin{aligned}
\mathfrak{K}_n^{\gamma}(x) &= \sum_{i=1}^{n} V_{i,n}\mathbf{1}_{V_{i,n}\leq\gamma_n} K\left(\frac{X_i - x}{h}\right), \\
\Delta\mathfrak{K}_n &= \sup_{x\in\mathcal{X}} \left|\mathfrak{K}_n^{\gamma}(x) - E[\mathfrak{K}_n^{\gamma}(x)] - \mathfrak{K}_n(x) + E[\mathfrak{K}_n(x)]\right|, \\
\mathcal{K}_n &= \sup_{x\in\mathcal{X}} \left|\mathfrak{K}_n^{\gamma}(x) - E[\mathfrak{K}_n^{\gamma}(x)]\right|, \\
\mathcal{K}_n^S &= \sup_{x\in\mathcal{X}} \left|\sum_{i=1}^{n} \varepsilon_i \left\{V_{i,n}\mathbf{1}_{V_{i,n}\leq\gamma_n} K\left(\frac{X_i - x}{h}\right)\right\}\right|,
\end{aligned}$$

such that we have

$$\|K_n\|_{\mathcal{X}} \leq \mathcal{K}_n + \Delta\mathfrak{K}_n + E[\mathfrak{K}_n(x)].$$

We first deal with the term $\Delta\mathfrak{K}_n$. Following the same idea as in the proof of Lemma 1 in Einmahl & Mason (2000), we have

$$\begin{aligned}
|\mathfrak{K}_n^{\gamma}(x) - E[\mathfrak{K}_n^{\gamma}(x)] - \mathfrak{K}_n(x) - E[\mathfrak{K}_n(x)]| &= \left|\sum_{i=1}^{n} V_{i,n}\mathbf{1}_{V_{i,n}>\gamma_n} K\left(\frac{X_i - x}{h}\right)\right. \\
&\quad \left. + nE\left[V_n\mathbf{1}_{V_n>\gamma_n} K\left(\frac{X - x}{h}\right)\right]\right|
\end{aligned}$$

with

$$\begin{aligned}
E\left[\sup_{x\in\mathcal{X}}\left|\sum_{i=1}^{n} V_{i,n}\mathbf{1}_{V_{i,n}>\gamma_n} K\left(\frac{X_i - x}{h}\right)\right|\right] &\leq n\|K\|_{\infty}\sup_{x\in\mathcal{X}} E[V_n\mathbf{1}_{V_n>\gamma_n}|X = x] \\
&\leq n\|K\|_{\infty}\sup_{x\in\mathcal{X}}\big(E[V_n^{2\alpha}|X = x]^{1/(2\alpha)} \\
&\qquad P(V_n > \gamma_n|X = x)^{1-1/(2\alpha)}\big) \\
&\leq n\|K\|_{\infty}\eta_n\gamma_n^{1-2\alpha} \\
&\leq \|K\|_{\infty}\eta_n^{1/(2\alpha)} n^{1/(2\alpha)}[\log(n)]^{1-1/(2\alpha)}
\end{aligned}$$

where we used Markov's inequality on line three. Using similar arguments, we can show that $nE\left[V_n \mathbf{1}_{V_n>\gamma_n} K\left(\frac{X-x}{h}\right)\right] = O(\eta_n^{1/(2\alpha)} h^d n^{1/(2\alpha)} [\log(n)]^{1-1/(2\alpha)})$.

Thanks to the assumptions, we have $n^{1/(2\alpha)}[\log(n)]^{1-1/(2\alpha)} = O(n^{1/2}h^{d/2}[\log(n)]^{1/2})$, so that

$$\Delta\mathfrak{K}_n = O(n^{1/2}h^{d/2}[\log(n)]^{1/2}\eta_n^{1/(2\alpha)}) \tag{3.33}$$

Then, we look for a bound to control the term $\mathcal{K}_n$. From Lemma 2.3.6 in A. W. Van der Vaart & Wellner (1996), we have

$$E[\mathcal{K}_n] \leq 2E[\mathcal{K}_n^S].$$

A bound for $E[\mathcal{K}_n^S]$ is obtained from Proposition 1 in Einmahl & Mason (2005). We apply this proposition to the random vector (V_n, X) and to the class of functions $\mathcal{G}_n = \{(v,u) \to v\mathbf{1}_{|v|\leq\gamma_n} K((u-x)/h), x \in \mathcal{X}\}$. In particular, Lemma 22 in Nolan & Pollard (1987) ensures that the proper bound holds for the covering number of the class $\mathcal{G}_n$. We then get

$$E[\mathcal{K}_n^S] = O\left(n^{1/2}h^{d/2}[\log n]^{1/2}\eta_n^{1/(2\alpha)}\right),$$

so that

$$\mathcal{K}_n = O\left(n^{1/2}h^{d/2}[\log n]^{1/2}\eta_n^{1/(2\alpha)}\right). \tag{3.34}$$

Finally the term $E[\mathfrak{K}_n(x)]$ is controlled thanks to the inequality

$$E[\mathfrak{K}_n(x)] = O(nh^d\tilde{\eta}_n), \tag{3.35}$$

The result is then the consequence of the inequalities (3.33), (3.34) and (3.35).

• Proof of (ii):

The principle is the same as in the proof of (i). Taking $\gamma_n = nh^d\eta_n^{1/(2\alpha)}$, we get that

$$\Delta\mathfrak{K}_n = O(nh^d\eta_n^{1/(2\alpha)}).$$

On the other hand the Corollary 4 from Einmahl & Mason (2005) is used (with $\beta = n^{1/2}h^d\eta_n^{1/2\alpha)}$ and $U = nh^d\eta_n^{1/(2\alpha)}$) instead of Proposition 1 to show that

$$\mathcal{K}_n = O\left(nh^d\eta_n^{1/(2\alpha)}\right).$$

Finally, notice that $\tilde{\eta}_n \leq \eta_n^{1/(2\alpha)}$, so that $E[\mathfrak{K}_n(x)] = O(nh^d\eta_n^{1/(2\alpha)})$. ■

Lemma 3.5 (Corollary of Lemma 3.4) *Let $\|K_n\|_{\mathcal{X}}$, η_n, $\tilde{\eta}_n$ and α as defined in Lemma 3.4, but now assume that $\alpha \in [1/2; 1]$. Also, let $\beta > 1$. We have the two following results:*

(i) Assume that $\liminf_{n\to\infty} nh^{\frac{d}{1-1/\beta}}[\log n]^{-1} > 0$, then

$$\|K_n\|_{\mathcal{X}} = O_P\left([nh^d\tilde{\eta}_n + n^{1/2}h^{d/2}[\log n]^{1/2}\eta_n^{1/(2\beta)}] \times n^{\frac{1-\alpha/\beta}{2\alpha}}\right).$$

(ii) Assume that $\liminf_{n\to\infty} nh^{2d} > 0$, then

$$\|K_n\|_{\mathcal{X}} = O_P\left([nh^d\eta_n^{1/(2\alpha)}] \times n^{\frac{1-\alpha/\beta}{2\alpha}}\right).$$

Proof. We have

$$|\mathfrak{K}_n(x)| \leq \left\{\sup_{1\leq i\leq n} V_{i,n}^{1-\alpha/\beta}\right\} \times \left|\sum_{i=1}^n V_{i,n}^{\alpha/\beta} K\left(\frac{X_i - x}{h}\right)\right|. \tag{3.36}$$

Let $\delta = 2\alpha/(1-\alpha/\beta)$. Since $E[V_{i,n}^{(1-\alpha/\beta)\delta}] < \infty$, $\sup_{1\leq i\leq n} V_{i,n}^{(1-\alpha/\beta)} = O_P(n^{1/\delta})$ (see e.g. the example following Lemma 2.2.1 in A. W. Van der Vaart & Wellner (1996)).

Thanks to the Lemma 3.4, the bound for the second term in (3.36) is $O_P(nh^d\tilde{\eta}_n + n^{1/2}h^{d/2}[\log n]^{1/2}\eta_n^{1/(2\beta)})$ for the case (i) and $O_P(nh^d\eta_n^{1/(2\beta)})$ for the case (ii). ■

3.6.7 Proof of Theorem 3.2

For the sake of simplicity, we assume in this section that $\theta \in \mathbb{R}$. The multidimensional case can be studied similarly, component by component, as we did in Section 3.6.1.

By Proposition 3.1, we already have that $\hat{\theta}_h(x) - \theta_h^*(x)$ tends uniformly to zero. To obtain the convergence rate, the key result consists in controlling the deviations of the process

$$\mathcal{Z}_h(x,\theta) = \frac{M_{n,h}(x,\theta) - M_{n,h}(x,\theta_h^*(x)) - M_f^{(c)}(x,\theta) + M_f^{(c)}(x,\theta_h^*(x))}{|\theta - \theta_h^*(x)|}.$$

Indeed, by definition of $\hat{\theta}_h(x)$ and $\theta_h^*(x)$,

$$M_{n,h}(x,\hat{\theta}_h(x)) - M_{n,h}(x,\theta_h^*(x)) \geq 0,$$

and

$$M_f^{(c)}(x,\theta_h^*(x)) - M_f^{(c)}(x,\hat{\theta}_h(x)) \geq 0.$$

Therefore,

$$0 \leq \frac{M_f^{(c)}(x,\theta_h^*(x)) - M_f^{(c)}(x,\hat{\theta}_h(x))}{|\hat{\theta}_h(x) - \theta_h^*(x)|} \leq \mathcal{Z}_h(x,\hat{\theta}_h(x)).$$

Moreover, by a second order Taylor expansion,

$$M_f^{(c)}(x,\hat{\theta}_h(x)) - M_f^{(c)}(x,\theta_h^*(x)) = \frac{(\hat{\theta}_h(x) - \theta_h^*(x))^2}{2} \nabla_\theta^2 M_f^{(c)}(x,\tilde{\theta}_h(x)),$$

where $\tilde{\theta}_h(x)$ belongs to the interval $[\hat{\theta}_h(x), \theta_h^*(x)]$. Due to Assumption 0.5 and to the consistency of $\hat{\theta}_h(x)$ shown in Proposition 3.1, $|\nabla_\theta^2 M_f^{(c)}(x,\tilde{\theta}_h(x))| \geq c_0' > 0$ for h small enough and n sufficiently large. The result of Theorem 3.2 then follows from Proposition 3.2 below.

Proposition 3.2 *Let $\beta > 1$. Then under Assumptions 0.6 to 0.8,*

$$\begin{aligned}\sup_{x,\theta} |\mathcal{Z}_h(x,\theta)| &= O_P(n^{-1/2}h^{-d/2}[\log n]^{1/2} + \nu_n^{1-1/p} + \varepsilon_n n^{\max(1/p''-1/(2\beta),0)} \\ &\qquad + n^{\max(1/p'-1/(2\beta),0)-1/2}).\end{aligned}$$

Proof of Proposition 3.2. With $M_{n,h}^*$ defined in (3.23), decompose

$$\mathcal{Z}_h(x,\theta) = \mathcal{Z}_h^*(x,\theta) + \mathcal{Z}_h^{(c)}(x,\theta),$$

where

$$\begin{aligned}\mathcal{Z}_h^*(x,\theta) &= \frac{M_{n,h}(x,\theta) - M_{n,h}(x,\theta_h^*(x)) - M_{n,h}^*(x,\theta) + M_{n,h}^*(x,\theta_h^*(x))}{|\theta - \theta_h^*(x)|}, \\ \mathcal{Z}_h^{(c)}(x,\theta) &= \frac{M_{n,h}^*(x,\theta) - M_{n,h}^*(x,\theta_h^*(x)) - M_f^{(c)}(x,\theta) + M_f^{(c)}(x,\theta_h^*(x))}{|\theta - \theta_h^*(x)|}.\end{aligned}$$

$\mathcal{Z}_h^*$ corresponds to the replacement of (A_i, B_i) by pseudo-observations $(\hat{A}_i, \hat{B}_i)$, while $\mathcal{Z}_h^{(c)}$ comes from the difference between the criterion when the margins are known and its expectation. These two terms are studied separately in Lemma 3.6 and Lemma 3.7. ■

Auxiliary Lemmas

Lemma 3.6 *Under Assumptions 0.7 and 0.8, and assuming that* $\liminf_{n\to\infty} nh^{2d} > 0$*, we have for any* $\beta > 1$*,*

$$\sup_{x,\theta} |\mathcal{Z}_h^*(x,\theta)| = O_P\left(\varepsilon_n n^{\max(1/p''-1/(2\beta),0)}\right).$$

Proof. From a first order Taylor expansion, we get

$$|\phi(a,b,\theta)-\phi(a,b,\theta_h^*(x))-\phi(\hat{a},\hat{b},\theta)+\phi(\hat{a},\hat{b},\theta_h^*(x))| \leq |\dot{\phi}(a,b,\tilde{\theta}_h(x))-\dot{\phi}(\hat{a},\hat{b},\tilde{\theta}_h(x))|\cdot|\theta-\theta_h^*(x)|,$$

for some $\tilde{\theta}_h(x)$ between θ and $\theta_h^*(x)$.

Next, from another Taylor expansion we have :

$$|\dot{\phi}(a,b,\tilde{\theta}_h(x)) - \dot{\phi}(\hat{a},\hat{b},\tilde{\theta}_h(x))| \leq |\partial_a\dot{\phi}(\tilde{a},\tilde{b},\tilde{\theta}_h(x))| \cdot |\hat{a}-a| + |\partial_b\dot{\phi}(\tilde{a},\tilde{b},\tilde{\theta}_h(x))| \cdot |\hat{b}-b|,$$

with, $\tilde{a} \in [a,\hat{a}]$ and $\tilde{b} \in [b,\hat{b}]$. Hence, we can use Assumption 0.7 to show that

$$|\mathcal{Z}_h^*(x,\theta)| \;\leq\; \frac{1}{nh^d}\sum_{i=1}^n W_{i,n} K\left(\frac{X_i-x}{h}\right)\left\{\tilde{r}_1(\tilde{A}_i)r_2(\tilde{B}_i)|\hat{A}_i-A_i| + r_1(\tilde{A}_i)\tilde{r}_2(\tilde{B}_i)|\hat{B}_i-B_i|\right\},$$

with $\tilde{A}_i \in [A_i,\hat{A}_i]$ (resp. $\tilde{B}_i \in [B_i,\hat{B}_i]$).

In order to obtain the desired result, we need to control terms of the same form as in the proof of Lemma 3.3. We then use similar arguments.

On the set $E_n(M) = \{\sup_{1\leq i\leq n} |\hat{A}_i - A_i| + |\hat{B}_i - B_i| \leq M\varepsilon_n\}$, which satisfies $\limsup_{M\to\infty} \lim_{n\to\infty} P(E_n(M)) = 1$ from Assumption 0.8, we have for some constants $C > 0$,

$$|\mathcal{Z}_h^*(x,\theta)| \leq CM\varepsilon_n \sup_{t\leq\mathfrak{S}_{(n)}}\left|\frac{S_C(t)}{\hat{S}_C(t)}\right| \cdot \frac{1}{nh^d}\sum_{i=1}^n W_i K\left(\frac{X_i-x}{h}\right)\big\{\tilde{r}_1(A_i)r_2(B_i) + r_1(A_i)\tilde{r}_2(B_i)\big\}\omega_i(\nu_n - M\epsilon_n),$$

with $\sup_{t\leq\mathfrak{S}_{(n)}} |S_C(t)/\hat{S}_C(t)| = O_P(1)$.

Next, let $\beta > 1$. From Lemma 3.4 (ii) and Lemma 3.5 (ii), we get

$$\frac{1}{nh^d}\sup_{x\in\mathcal{X}}\left|\sum_{i=1}^{n} W_i K\left(\frac{X_i - x}{h}\right)(\tilde{r}_1(A_i)r_2(B_i) + r_1(A_i)\tilde{r}_2(B_i))\right| = O_P(n^{\max(1/p''-1/(2\beta),0)}), \tag{3.37}$$

so that

$$\sup_{x\in\mathcal{X},\theta\in\Theta}|\mathcal{Z}_h^*(x,\theta)| = O_P\left(\varepsilon_n n^{\max(1/p''-1/(2\beta),0)}\right)$$

■

Lemma 3.7 *Let $\beta > 1$. Under Assumptions 0.6 and 0.8,*

$$\sup_{x,\theta}|\mathcal{Z}_h^{(c)}(x,\theta)| = O_P(n^{-1/2}h^{-d/2}[\log n]^{1/2} + \nu_n^{1-1/p} + n^{\max(1/p'-1/(2\beta),0)-1/2}).$$

Proof. Let

$$\phi_{\theta,\theta'}(y,z,x) = \frac{\log c_\theta(F_T(y|x),F_U(z|x)) - \log c_{\theta'}(F_T(y|x),F_U(z|x))}{(\theta-\theta')}.$$

Let $\mathcal{A}$ denote the class which contains all the functions $\phi_{\theta,\theta'}$. We have

$$\phi_{\theta,\theta'}(y,z,x) \le \Psi(y,z,x) := R(F_T(y|x),F_U(z|x)),$$

where we used Assumption 0.6.

Let

$$L_n(x,h,\theta) = \frac{1}{nh^d}\sum_{i=1}^{n} W_i K\left(\frac{X_i - x}{h}\right)\phi_{\theta,\theta_h^*(x)}(Y_i,Z_i,X_i),$$

and let $\beta > 1$. It follows from Lemma 3.1 and Lemma 3.2 that

$$\frac{M_{n,h}^*(x,\theta) - M_{n,h}^*(x,\theta_h^*(x))}{(\theta-\theta_h^*(x))} = L_n(x,h,\theta) + O_P\Big(n^{-1/2}h^{-d/2}[\log n]^{1/2} + \nu_n^{1-1/p} + n^{\max(1/p'-1/(2\beta),0)-1/2}\Big),$$

where the O_P-rate is uniform in x and θ.

On the other hand,

$$\sup_{x,\theta}\left|L_n(x,h,\theta) - \frac{M_f^{(c)}(x,\theta) - M_f^{(c)}(x,\theta_h^*(x))}{\theta-\theta_h^*(x)}\right| = O_P(n^{-1/2}h^{-d/2}[\log n]^{1/2}).$$

This result is obtained using Theorem 4 in Einmahl & Mason (2005). Let $\mathcal{A}^\delta = \{(y,z,c,x) \to \mathbf{1}_{y+z\leq c}\ a(y,z,x)S_C(y+z)^{-1} : a \in \mathcal{A}\}$, and $\Psi^\delta(T,U,C,X) = \mathbf{1}_{T+U\leq C}\Psi(T,U,X)S_C(T+U)^{-1}$. The conditions in Theorem 4 in Einmahl & Mason (2005) hold if we check that

$$N(\varepsilon, \mathcal{A}^\delta, \Psi^\delta) \leq \Delta\varepsilon^{-\alpha}, \tag{3.38}$$

for some $\alpha > 0$ and $\Delta > 0$, and if

$$E\left[\left\{\frac{\delta\Psi(Y,Z,X)}{S_C(Y+Z)}\right\}^p\right] < \infty, \tag{3.39}$$

for some $p > 2$. Condition (3.39) is easy to check, since

$$E\left[\left\{\frac{\delta\Psi(Y,Z,X)}{S_C(Y+Z)}\right\}^p\right] = E\left[\frac{\Psi(T,U,X)^p}{S_C(T+U)^{p-1}}\right]$$

which is finite from (3.10) in Assumption 0.6.

To check (3.38), observe that

$$\phi_{\theta_1,\theta_2}(y,z,x) - \phi_{\theta_3,\theta_4}(y,z,x) = \dot{\phi}(F_T(y|x), F_U(z|x), \tilde{\theta}) - \dot{\phi}(F_T(y|x), F_U(z|x), \bar{\theta}),$$

for $\tilde{\theta}$ between θ_1 and θ_2, and $\bar{\theta}$ between θ_3 and θ_4. Then we get, from a Taylor expansion,

$$\phi_{\theta_1,\theta_2}(y,z,x) - \phi_{\theta_3,\theta_4}(y,z,x) = \ddot{\phi}(F_T(y|x), F_U(z|x), \theta^*)(\tilde{\theta} - \bar{\theta}),$$

for θ^* between $\tilde{\theta}$ and $\bar{\theta}$. From Assumption 0.5, we deduce that the class $\mathcal{A}$ satisfies (3.38) thanks to Lemma 19.31 of A. Van der Vaart (1998). ■

Chapitre 4

The impact of churn on client value in health insurance, evaluation using a random forest under random censoring

In the insurance broker market, commissions received by brokers are closely related to so-called "customer value": the longer a policyholder keeps their contract, the more profit there is for the company and therefore the broker. Hence, predicting the time at which a potential policyholder will surrender their contract is essential in order to optimize a commercial process and define a prospect scoring. In this paper, we propose a weighted random forest model to address this problem. Our model is designed to compensate for the impact of random censoring. We investigate different types of assumptions on the censoring, studying both the cases where it is independent or not from the covariates. We compare our approach with other standard methods which apply in our setting, using simulated and real data analysis. We show that our approach is very competitive in terms of quadratic error in addressing the given problem.

4.1 Introduction

In insurance brokerage, an important problem is to evaluate what is known as "prospect value". Roughly speaking, this represents the rentability of a potential policyholder. More effort could then be dedicated to attract a prospect with probable high profitability. A first approach to evaluate this prospect value is to predict how long a given current customer will keep their contract. To investigate this, brokers now have access to large databases of potentially relevant information. In the present paper, our aim is to discuss an extension of the *random forest* algorithm which takes into account specific details of this framework. Among these, a crucial issue is to deal with right-censoring, which is a common problem when dealing with duration variables. Due to the temporal phenomenon that we study, a significant number of observations are incomplete and require particular care to counterbalance the corresponding lack of information. The method we propose consists of an appropriate weighting of the observations to compensate for censoring. We discuss different approaches and compare them using simulation and real data analysis.

The random forest algorithm was first proposed in Breiman (2001). The underlying aim is to estimate $E[\phi(T)|X]$ with the help of bootstrap data augmentation and aggregation of regression trees. In our case, T is a right-censored variable, X a vector of covariates, and ϕ a known gain function, namely the value of a customer remaining a time T in the portfolio. In the presence of censoring Hothorn, Bühlmann, et al. (2006) proposed a generalization extending the random forest algorithm. Using the *inverse probability of censoring weights* (IPCW) approach, described in Van der Laan & Robins (2003), they introduce weights to compensate for censoring. Additionally, Steingrimsson et al. (2016, 2018) combined doubly robust estimators with survival trees. The novelty of our approach is to compute IPCW in situations where censoring depends on the covariates, as it was suggested in Molinaro et al. (2004), and we will discuss its performance compared to other weighting strategies. The IPCW technique is a very general one: once the weights have been determined, any regression technique can be generalized to this framework – see for example Koul et al. (1981) for linear regression, Goldberg & Kosorok (2017) for support vector machines, and Lopez et al. (2016) for *classification and regression trees* (CART). The difficulty in defining appropriate weights is in determining which identifiability assumption is reasonable given the situation. The weights may be considerably different if censoring is allowed to depend on covariates X (see e.g., Lopez et al. (2013)), or not (see e.g., Stute (2003)), and, as we will show on real data, this has important consequences on the results of the procedure.

Another modification of random forests to the presence of censoring is the *random survival forest* (RSF) of Ishwaran et al. (2008) (see also Ishwaran & Kogalur (2007)). In this procedure, a log-rank test is used to split the observations at each computational step of the the regression trees. This type of procedure for growing a survival tree was previously proposed in Ciampi et al. (1986) and Segal (1988). It was also studied in LeBlanc & Crowley (1993) and in Hothorn, Hornik, & Zeileis (2006) in the context of conditional inference trees. Moreover, Zhu & Kosorok (2012) have studied the impact of recursively fitted RSF models on prediction quality. LeBlanc & Crowley (1992) introduced the *relative risk tree* (RRT) algorithm, where a proportional hazard likelihood is used as the criterion at the splitting step. Forests of relative risk trees were investigated in Hothorn et al. (2004) and Ishwaran et al. (2004). The extension to left-truncation of this approach was considered in Fu & Simonoff (2016). Additional splitting criteria examined in the literature include the exponential log-likelihood in Davis & Anderson (1989), and splits determined through analysis of residuals of the Cox model (Ahn & Loh (1994)). More recently, Zhu (2013) & Zhu et al. (2015) proposed the *reinforcement learning trees* (RLT) algorithm which consists in fitting an embedded model at each step of the tree construction to improve the selection of the splitting variable, while Li & Bradic (2019) explored censored quantile regression using random forests.

While the theoretical properties of tree and forest-based learning algorithms are not fully understood yet, the consistency of RSF was first investigated in Ishwaran & Kogalur (2010) under the assumption that X takes a finite number of values. These properties were further studied in Cui, Zhu, & Kosorok (2017) & Cui, Zhu, Zhou, & Kosorok (2017) which provided a theoretical framework to consider the consistency of survival forests, and established consistency under specific conditions that include random splitting rules and splitting rules with marginal signal checking. Cui, Zhu, Zhou, & Kosorok (2017) also underlined the problem of non optimal split selection for usual survival tree methods. The method we study in this article does not suffer from such problem since it requires less assumptions. Finally, a review of the literature on survival trees can be found in Bou-Hamad et al. (2011).

The rest of the paper is organized as follows. In Section 4.2, we present the specific details of the types of observations we examine here, and describe the relevant random forest algorithm we have adapted to censoring. A simulation study is performed in Section 4.3, while the real-world behavior of our procedure is illustrated in Section 4.4, where we present an application of this weighted random forest to model the churn behavior of

health insurance policy holders. Further results as well as explanations about the choice of the random forest parameters are provided in supplementary material.

4.2 Description of the method

4.2.1 The survival regression setting

To study the termination risk, we introduce the *lifetime* of a contract, denoted T. This duration is not directly observed due to the presence of right-censoring (which is a classical issue in survival analysis). Instead of observing T, we observe a pair of variables (Y, δ) defined as

$$Y = \min(T, C),$$
$$\delta = \mathbb{1}_{T \leq C}.$$

The need for the censoring variable C is due to the fact that some contracts may not have been terminated at the end of the observation period. Additionally, a vector of covariates $X \in \mathcal{X} \subset \mathbb{R}^p$ is available, in order to identify certain characteristics that may influence the termination time. Observations correspond to i.i.d. replicates $(Y_i, \delta_i, X_i)_{1 \leq i \leq n}$ of this set-up. For the sake of simplicity, we consider the case where T and C are continuous random variables. For any continuous random variable U, we will denote by $S_U(t) = P(U > t)$ its survival function.

Our aim is to estimate the function $f(x) = E[\phi(T)|X = x]$. In our application, ϕ is a pricing function that associates a unitary profit with a customer remaining a time T with the portfolio in question.

Since T is not directly observed because of censoring, classical estimators of f (random forests, as developed below, or any other regression estimator) are biased if no attempt is made to correct for this phenomenon. The procedure we propose is based on the inverse probability of censoring weighting (IPCW, see e.g., Van der Laan & Robins (2003)) which, through an appropriate weighting of the observations, aims to suppress this bias. This general method is described in Section 4.2.2 below.

4.2.2 Inverse probability of censoring weighting

Introduction

IPCW is a general principle which consists of correcting the bias introduced by censoring. Given a continuous random variable V (possibly multidimensional), the key task is to determine which weights should be put on an uncensored observation in order to retrieve the distribution of the variable of interest.

Proposition 4.1 (The IPCW principle) *Let* $\gamma = \begin{cases} 1 \text{ if } V \text{ is not censored} \\ 0 \text{ if } V \text{ is censored} \end{cases}$ *and* $V' = \gamma V$.

Let $p(v) = P(\gamma = 1 | V = v)$ *and assume that* $\forall v, p(v) > 0$.

Then for any function ψ,

$$E\left[W \cdot \psi(V')\right] = E\left[\psi(V)\right], \quad \text{with} \quad W = \frac{\gamma}{p(V')}.$$

This result states that given a random variable V that is not always observable, it is still possible to estimate its distribution based on the observations $(V'_i, \gamma_i)_{i=1,..,n}$. This is done by attributing weights $\frac{\gamma_i}{p(V'_i)}$ to the observations, with $p(v)$ the probability of V being non-censored, given $V = v$.

In our setting, we can apply the IPCW routine to the vector (T, X). This results in the equality:

$$E\left[W \cdot \psi(Y, X)\right] = E\left[\psi(T, X)\right], \tag{4.1}$$

where $W = \frac{\delta}{p(Y,X)}$ and $p(t, x) = P(\delta = 1 | T = t, X = x) = P(t \leq C | T = t, X = x)$.

In the survival setting, it is generally impossible to infer values for $p(t, x)$ since – as it is well known – the identifiability of the model requires assumptions on the dependence between T and C that cannot be statistically verified (see Section 4.1 in Lagakos (1979)). However, with assumptions on the dependence between T and C, it is possible to compute $p(t, x)$.

Let **H1** and **H2** denote the following hypotheses:

$$\begin{aligned}\textbf{H1}&: P(t \leq C|X = x, T = t) = S_C(t),\\ \textbf{H2}&: P(t \leq C|X = x, T = t) = S_C(t|X = x),\end{aligned}$$

with $S_C(\cdot|X = x)$ the conditional distribution function of C given $X = x$. Sufficient conditions for these hypotheses to be satisfied are, respectively, $(T, X) \perp\!\!\!\perp C$ (**H1**) and $T \perp\!\!\!\perp C$ conditionally on X (**H2**).

Clearly **H2** is more general than **H1**, but requires estimation of a conditional survival function, which is more tricky. Hence, if the strongest assumption **H1** is reasonable, it may be interesting to use it instead of **H2**, in order to facilitate computation of the appropriate weights. The obvious drawback is that, if the censoring is dependent on the covariates, **H1** does not hold.

In what follows, we consider these two assumptions as separate cases, but both are handled simultaneously. Depending on the hypothesis we make, we note $W = \delta/S_C(Y)$ (**H1**) or $W = \delta/S_C(Y|X)$ (**H2**), so that in both cases equation (4.1) holds. We also note for all $i \in \{1, .., n\}$, $W_i = \delta_i/S_C(Y_i)$ (or $W_i = \delta_i/S_C(Y_i|X = X_i)$) the exact IPCW weights. However, the function S_C (resp. $S_C(\cdot|X)$) is unknown and we have to estimate it in order to compute the weights W.

Computation of the weights

Computation of the IPCW weights requires us to estimate the survival function of the censoring variable. In many applications, the roles of the variables T and C are symmetric, and the censoring variable needs to be studied using survival models. In our work, we have used three strategies to estimate S_C (resp. $S_C(\cdot|X)$):

- S_C estimated with the Kaplan-Meier (KM) estimator (Kaplan & Meier (1958b)): $\hat{S}_C(t) = \prod_{Y_i \leq t} \left(1 - (1 - \delta_i)/\sum_{j=1}^{n} \mathbb{1}_{Y_j \geq Y_i}\right)$.
- $S_C(\cdot|X)$ estimated with the Cox model (Cox (1972)). This model assumes that the hazard rate has the form: $\lambda_C(t|X = x) = \lambda_0(t) \cdot e^{{}^t\beta \cdot x}$ with β a vector of coefficients, and for some random variable U, $\lambda_U = -S'_U/S_U$.
- $S_C(\cdot|X)$ estimated with the *random survival forest* (RSF) algorithm (Ishwaran et al. (2008)).

An alternative idea for estimating $S_C(\cdot|X)$ would be to use a kernel estimator such as the conditional Kaplan-Meier estimator of Beran (1981) and Dabrowska (1989). Nevertheless, this would rely on kernel smoothing whose behavior deteriorates when the dimension of X increases (i.e. $p > 3$) as noted in Lopez et al. (2013). This dimensional constraint is the reason for imposing a stronger assumption on the conditional distribution of the censoring as a feasible compromise.

Given estimators $\hat{S}_C$ (resp. $\hat{S}_C(\cdot|X)$) of S_C (resp. $S_C(\cdot|X)$), the weights $(\hat{W}_i)_{i=1,..n}$ are computed using the formula $\hat{W}_i = \delta_i/\hat{S}_C(Y_i)$ (resp. $\hat{W}_i = \delta_i/\hat{S}_C(Y_i|X_i)$). Therefore, each method used to estimate S_C leads to different IPCW weights, then to different estimators of the joint distribution of (T, X). In the following section, the weights $(\hat{W}_i)_{i=1,..n}$ refer to any collection of weights computed with one of the three methods.

4.2.3 A weighted random forest algorithm for the regression of right-censored data

Random forest is an ensemble learning algorithm which consists of an aggregation of elemental regression trees (base learners), see e.g., Breiman (2001) and Biau & Scornet (2016). A regression tree (see Breiman et al. (1984)) produces a partition of a dataset using successive binary splits based on values of the covariates. In each subset obtained (called a *leaf*), the observations correspond to a homogeneous group. The key behind group formation is the choice of the splitting criterion (e.g., least squares) which aims to reduce heterogeneity at each step.

The random forest algorithm is a combination of two methods: a *classification and regression tree* (CART) algorithm to compute each tree, and a bagging algorithm (bagging means *bootstrap aggregating*, see Breiman (1996)) to introduce randomness into partition formation by building bootstrap samples. Compared to a single regression tree approach, tree aggregation in random forests stabilizes results by making them less sensitive to changes in the database.

The rest of this section is devoted to adapting the random forest algorithm to the presence of right censoring. In Section 4.2.3, we propose a modification of the splitting criterion to work in our framework. Computation of predictions given a tree is shown in Section 4.2.3. An overall description of the algorithm is given in Section 4.2.3, and we discuss its parameters in Section 6.

Split selection

In the building of a binary tree, the main interest is in the way splits are determined. Here, we describe the split selection procedure used at each node of the tree, for the growing of one tree of the forest. This tree is built using a bootstrap sample of the initial data.

Let $\mathcal{D}_n$ be a list of indices in $\{1, \dots, n\}$ which represents a bootstrap sample of size n of the initial data, drawn uniformly with replacement. Given a set $B \subset \mathcal{X}$, let $n^{\mathcal{D}_n,w}(B) = \sum_{i\in\mathcal{D}_n} W_i \mathbb{1}_{X_i\in B}$ be the weighted number of observations of $\mathcal{D}_n$ that belong to B, and $\bar{\phi}_B^{\mathcal{D}_n,w} = 1/n^{\mathcal{D}_n,w}(B) \sum_{i\in\mathcal{D}_n} W_i \phi(Y_i) \mathbb{1}_{X_i\in B}$ the weighted mean of $\phi(Y_i)$ for the observations within B (we set $\bar{\phi}_B^{\mathcal{D}_n,w} = +\infty$ if $n^{\mathcal{D}_n,w}(B) = 0$).

Let $A \subset \mathcal{X}$ be a node of a tree, $j \in \{1, \dots, p\}$ a variable, and u real number. Define:

$$L^w(u, j, A, \mathcal{D}_n) = \frac{1}{n^{\mathcal{D}_n,w}(A)} \cdot \sum_{i\in\mathcal{D}_n} W_i \cdot \left(\phi(Y_i) - \bar{\phi}_{A_l}^{\mathcal{D}_n,w} \mathbb{1}_{X_i^{(j)}\le u} - \bar{\phi}_{A_r}^{\mathcal{D}_n,w} \mathbb{1}_{X_i^{(j)}>u}\right)^2 \cdot \mathbb{1}_{X_i\in A},$$

where $A_l = A \cap \{X^{(j)} \le u\}$ and $A_r = A \cap \{X^{(j)} > u\}$.

We also define $n^{\mathcal{D}_n,\hat{w}}(B)$, $\bar{\phi}_B^{\mathcal{D}_n,\hat{w}}$ and $L^{\hat{w}}(u, j, A, \mathcal{D}_n)$ as the same quantities with the weights W_i replaced by the weights $\hat{W}_i$.

The binary split chosen at node A is $A = A_l^* \cup A_r^*$, with $A_l^* = A \cap \{X^{(j^*)} \le u^*\}$, $A_r^* = A \cap \{X^{(j^*)} > u^*\}$, and (u^*, j^*) given by:

$$(u^*, j^*) = \underset{\substack{j\in\{1,..,p\} \\ u\in\left\{X_i^{(j)} \,/\, i\in\mathcal{D}_n \text{ s.t. } X_i\in A\right\}}}{\operatorname{argmin}} L^{\hat{w}}(u, j, A, \mathcal{D}_n).$$

For a heuristic understanding of the algorithm, consider a fixed set A. We abusively assume that A is independent from the data, which is not correct in practice since A is selected from the data. Under this assumption, we can study the asymptotics of $L^w(u, j, A, \mathcal{D}_n)$ since the observations $(Y_i, X_i)_{i\in\mathcal{D}_n}$ which belong to A are i.i.d. and follow the conditional distribution $\mathcal{L}((Y, X)|X \in A)$. Until the end of this section, we assume that A is independent from the data.

With A_z standing for either A_l or A_r, we have from the law of large numbers and equation (4.1),

$$\bar{\phi}_{A_z}^{\mathcal{D}_n,w} \underset{n\to+\infty}{\longrightarrow} E[W \cdot \phi(Y)|X \in A_z] = E[\phi(T)|X \in A_z],$$

and

$$\frac{1}{n^{\mathcal{D}_n,w}(A_z)} \cdot \sum_{\substack{i\in\mathcal{D}_n \\ X_i\in A_z}} W_i \cdot \left(\phi(Y_i) - \bar{\phi}_{A_z}^{\mathcal{D}_n,w}\right)^2$$

$$\underset{n\to+\infty}{\longrightarrow} E\left[W \cdot \left(\phi(Y) - E[W\cdot\phi(Y)|X\in A_z]\right)^2 \middle| X\in A_z\right]$$

$$= E\left[\left(\phi(T) - E[\phi(T)|X\in A_z]\right)^2 \middle| X\in A_z\right].$$

Then, by splitting the sum in $L^w(u,j,A,\mathcal{D}_n)$ between A_l and A_r, we get:

$$L^w(u,j,A,\mathcal{D}_n) \underset{n\to+\infty}{\longrightarrow} L^\infty(u,j,A),$$

with

$$L^\infty(u,j,A) = \frac{P(X\in A_l)}{P(X\in A)} \cdot E\left[(\phi(T) - E[\phi(T)|X\in A_l])^2 \middle| X\in A_l\right] + \frac{P(X\in A_r)}{P(X\in A)} \cdot E\left[(\phi(T) - E\left[\phi(T)|X\in A_r\right])^2 \middle| X\in A_r\right].$$

Therefore we observe that, at the cost of replacing $\hat{W}_i$ by W_i, the selected split is asymptotically the one which minimizes the sum of the within variances of the two child nodes, weighted by their relative importance.

It is also important to see that given $\mathcal{D}_n$ and the splitting criteria L, the growing of a tree is deterministic. Therefore we can define a function $\mathcal{T}(\mathcal{D},L)$ which outputs a tree given a sample and a splitting criterion.

Tree prediction

The growing of a binary tree $\mathcal{T}$ results in a partition $(\mathcal{X}_k)_{k=1,..,K}$ of the input space. In order to predict the mean value of the target variable, a natural estimator is $m^{\hat{w}}(x) = \sum_{k=1}^K \bar{\phi}_{\mathcal{X}_k}^{\mathcal{D}_n,\hat{w}} \mathbb{1}_{x\in\mathcal{X}_k}$. Under **H1**/**H2** we have for the denoised version m^w,

$$m^w(x) \underset{n\to+\infty}{\longrightarrow} \sum_{k=1}^K E\left[\phi(T)|X\in\mathcal{X}_k\right]\mathbb{1}_{x\in\mathcal{X}_k},$$

which shows that $m^{\hat{w}}$ approximates a piecewise constant estimator of f.

However, though the growing step is done using a least-square criterion – which is well-adapted to mean regression – the prediction made in the terminal leaf may not necessarily be a sample mean (e.g., it may be a quantile – see Meinshausen (2006)). We can formalize this by defining a function $\mathcal{M}(\mathcal{T}, \mathcal{D}, m)$ which outputs a predictor given a binary tree, a sample, and a type of terminal leaf-based estimator. In the following section, we propose a second terminal leaf estimator for the mean within a leaf.

Both for tree prediction and split selection, note that the method we propose is a generalization of the classical CART regression algorithm. Indeed, setting the weights W_i all equal to 1 results in Breiman's original algorithm. We note L and m the split criteria and terminal leaf estimator obtained when setting all weights W_i to 1.

Description of the algorithms

The tools we developed in previous sections can be used in different random forest algorithms. We have studied three survival weighted random forests (*swRF*) in this article: *swRF1*, *swRF2* and *swRF3*. These algorithms differ in the way weights $\hat{W}_i$ are introduced.

In *swRF1* the weights $\hat{W}_i$ are taken into account in the bootstrap part of the random forest. Based on the whole training set, we build $\hat{S}_C$ (or $\hat{S}_C(\cdot|X)$) and deduce $(\hat{W}_i)_{i=1,\dots,n}$. Then, all bootstrap samples of the forest are drawn with replacement and with sampling probabilities proportional to the weights $(\hat{W}_i)_{i=1,\dots,n}$. A CART tree is then grown on each bootstrap sample, using the classical algorithm (i.e. with L and m which denote the splitting criterion and the terminal leaf estimator with uniform weights).

In algorithms *swRF2* and *swRF3*, bootstrap samples are drawn uniformly with replacement, as it is done in the classical random forest algorithm, and the weights are computed on the bootstrap samples. This means that the weights of a given observation may vary from tree to tree. The only difference between *swRF2* and *swRF3* is the terminal leaf estimator used. In *swRF2*, $m^{\hat{w}}$ is used as described in Section 4.2.3, whereas in *swRF3* we use $m^{\hat{w}_{loc}}(x) = \sum_{k=1}^{K} \bar{\phi}_{\mathcal{X}_k}^{\mathcal{D}_n, \hat{w}_{loc}} \mathbb{1}_{x \in \mathcal{X}_k}$. In fact, $m^{\hat{w}_{loc}}$ is similar to $m^{\hat{w}}$ except that the weights used in each terminal leaf are computed using a Kaplan-Meier estimator inside the leaf.

Pseudocode of the three algorithms we study is shown in Algorithms 3 and 4 below.

Algorithme 3 : $swRF1$

Input : Data : $(Y_i, \delta_i, X_i)_{i=1,..,n}$, $M > 0$: number of trees

Output : Ensemble predictor $swRF1$

Compute weights $(\hat{W}_i)_{i=1,..,n}$: $\hat{W}_i = \delta_i / \hat{S}_C(Y_i)$ (or $\hat{W}_i = \delta_i / \hat{S}_C(Y_i|X_i)$)

for $j = 1, \ldots, M$ **do**

  Draw $\mathcal{D}_{n,j}$: sample n observations w.r.t. weights $(\hat{W}_i)_{i=1,..,n}$, with replacement

  Build $\mathcal{T}_j = \mathcal{T}(\mathcal{D}_{n,j}, L)$

  $\hat{m}_j = \mathcal{M}(\mathcal{T}_j, \mathcal{D}_{n,j}, m)$

end

return : $\hat{m}_{swRF1} = \frac{1}{M} \sum_{i=1}^{M} \hat{m}_j$

Algorithme 4 : $swRF2$ & $swRF3$

Input : Data : $(Y_i, \delta_i, X_i)_{i=1,..,n}$, $M > 0$: number of trees, $mode \in \{1, 2\}$: type of terminal leaf estimator

Output : Ensemble predictor $swRF2$ (or $swRF3$)

for $j = 1, \ldots, M$ **do**

  Draw $\mathcal{D}_{n,j}$: sample n observations uniformly with replacement

  Compute weights $(\hat{W}_i^{\mathcal{D}_{n,j}} : i \in \mathcal{D}_{n,j})$: $\hat{W}_i^{\mathcal{D}_{n,j}} = \delta_i / \hat{S}_C^{\mathcal{D}_{n,j}}(Y_i)$ $\left(\text{or } \hat{W}_i^{\mathcal{D}_{n,j}} = \delta_i / \hat{S}_C^{\mathcal{D}_{n,j}}(Y_i|X_i)\right)$

  Build $\mathcal{T}_j = \mathcal{T}(\mathcal{D}_{n,j}, L^{\hat{w}})$

  $\hat{m}_j = \mathcal{M}(\mathcal{T}_j, \mathcal{D}_{n,j}, m^{\hat{w}})$ if $mode = 1$ ($\hat{m}_j = \mathcal{M}(\mathcal{T}_j, \mathcal{D}_{n,j}, m^{\hat{w}_{loc}})$ if $mode = 2$)

end

return : $\hat{m}_{swRF(2,3)} = \frac{1}{M} \sum_{i=1}^{M} \hat{m}_j$

Remark 4.1 *For each of swRF1, swRF2 or swRF3, we can build three models which correspond to the three ways to estimate the weights $(W_i)_{i=1,\dots,n}$ developed in Section 4.2.2.*

A practical limitation of the weighted random forest algorithm is that for a large non-censored observation Y_i, the corresponding weight tends to be very big (Cui, Zhu, & Kosorok (2017)). This is even more so in the conditional case if X is multidimensional, since the estimated survival function $\hat{S}_C(\cdot|X = x)$ may decrease quickly for some values of x. Overly-large weights may give too much importance to a single observation, both in terms of split selection and terminal leaf estimation.

We propose two ways to deal with this problem. The first is to threshold the weights $\hat{W}_i$ so that the maximum ratio between two nonzero weights does not exceed some value. Let r_{max} be the maximum ratio allowed between two nonzero IPCW weights. The choice of the value for r_{max} corresponds to a typical bias-variance trade-off. Indeed, thresholding the weights introduces bias in the swRF procedure since large non-censored observations get smaller weights, and thus predicted values tend to be lower, but on the other hand this lowers the estimator's variance, which is high when some observations have very large weights.

The second strategy is to use swRF to model the time $T' = min(T, T_{max})$ and not the original time T. This is natural because we are working on the estimation of the expected value for T, and the problem of estimating the average of a right-censored time is generally hard to solve; indeed, there is high variance in the estimation of the tail of the distribution, which causes fluctuations in the estimated mean.

In the following, we will combine the two strategies to achieve good accuracy using swRF.

Parameters of the random forest algorithm

In our earlier presentation of random forests, we made no mention of the parameters which need to be specified to build them. Since there is a large number of parameters, and since parameters may differ according to the random forest implementation in question, we only mention the most common ones – those which have been shown to have the strongest impact on the resulting model.

The main parameters in a random forest are summarized in Tab. 4.1. The parameters *minleaf* and *maxdepth* impact tree size in the forest, and we discuss their influence on the performances of the algorithms in supplementary material. As for the parameter *mtry*, it is especially important in high dimensional settings.

Parameter	Description
$minleaf$	minimum value for the number of observations that should be present in a leaf; a split that would result in a leaf of less than $minleaf$ observations cannot be selected
$maxdepth$	length of the longest downward path from the root node to a leaf of the tree; a tree only consisting of the root node has depth 0.
$mtry$	number of candidate variables that are randomly sampled at each node when choosing the best split

Tab. 4.1: Parameters of the random forest algorithm.

4.2.4 Assessing the quality of a model's fit

In the applications which follow, our strategy to evaluate a fitted model relies on train-test approaches: let $\mathcal{D}_{tr}$ and $\mathcal{D}_{te}$ be train and test set of indices. However, due to the right censoring, the usual accuracy criterion used in regression settings cannot be computed on the test set.

In fact, since the problems of bias on the test and train samples are similar, we can use for model validation the same sample fitting method with IPCW that we use for model training. Let $(\hat{W}_i)_{i\in\mathcal{D}_{te}}$ be the estimated IPCW and $\hat{f}$ the predictor function of a fitted model. Then, the quantity $1/n_{te} \cdot \sum_{i\in\mathcal{D}_{te}} \hat{W}_i \cdot (\phi(Y_i) - \hat{f}(X_i))^2$ is an estimator of the mean squared error (MSE) of the model $\hat{f}$ (with $n_{te} = Card(\mathcal{D}_{te})$ the number of test observations). This is the method suggested in Gerds & Schumacher (2006) and the one we adopt in this article. One disadvantage is that again this approach raises the question of the choice of weights to consider for IPCW; in the same way as for training, we choose here to compute the validation error corresponding to three types of weight: KM, Cox and RSF. Therefore, we use not one but three criteria to compare model performance. Computation of the IPCW weights is performed separately on the training and test sets.

The problem of model validation in the presence of censoring has received considerable attention in the literature, and other performance measures exists. The C-index (Harrell et al. (1982)), which generalizes the Kendall tau for right-censored data, is very popular, so we also compute it when comparing models.

4.3 Simulated data example

4.3.1 Technical details

Data simulation

In our simulations, the covariates X are distributed as marginal uniform distributions on $[-1, 1]$ linked through a Gaussian copula with covariance given by an $AR(1)$ covariance matrix with coefficient ρ, i.e., a matrix K^ρ such that $K^\rho_{i,j} = \rho^{|i-j|}$. The coefficient ρ is random and uniformly distributed on $[0, 0.6]$.

We investigate three experimental settings which correspond to three different models for $\mathcal{L}(T|X)$, the distribution of T given X. In the following description of the three cases, λ_T and k_T are constants that will be specified later, and $Weibull(\lambda, k)$ refers to the distribution with density $g(u) = k/\lambda \cdot (u/\lambda)^{k-1} e^{-(u/\lambda)^k} \mathbb{1}_{u\geq 0}$:

- Case 1 (Weibull): $T|(X, \beta_T) \sim Weibull(\lambda_T e^{-{}^t\beta_T \cdot X}, k_T)$, with $\beta_T \sim N(\mu = 0, \sigma^2 = 0.04 \cdot I_p)$ and $\beta_T \perp\!\!\!\perp X$.

- Case 2 (Independent mixture of Weibulls): $T|(X, G, (\beta_{T,j})_{j=1,\dots,4}) \sim Weibull(\lambda_T e^{-{}^t\beta_{T,G} \cdot X}, k_T)$ with:

  - $G \sim Unif\{1, 2, 3, 4\}$ and $G \perp\!\!\!\perp X, (\beta_{T,j})_{j=1,\dots,4}$,
  - $(\beta_{T,j})_{j=1,\dots,4}$ i.i.d. with $\forall j \in [\![1, 4]\!]$, $\beta_{T,j} \sim N(\mu = 0, \sigma^2 = 0.09 \cdot I_p)$ and $(\beta_{T,j})_{j=1,\dots,4} \perp\!\!\!\perp X$.

- Case 3 (Covariate-dependent mixture of Weibulls, $p \geq 2$):

  $$T|(X, G, (\beta_{T,j})_{j=1,\dots,4}) \sim Weibull(\lambda_T e^{-{}^t\beta_{T,G} \cdot X}, k_T)$$

  with:

  - $G = \mathbb{1}_{X_1\geq 0, X_2\geq 0} + 2 \cdot \mathbb{1}_{X_1\geq 0, X_2<0} + 3 \cdot \mathbb{1}_{X_1<0, X_2\geq 0} + 4 \cdot \mathbb{1}_{X_1<0, X_2<0}$,
  - $(\beta_{T,j})_{j=1,..,4}$ i.i.d. with $\forall j \in [\![1, 4]\!]$, $\beta_{T,j} \sim N(\mu = 0, \sigma^2 = 0.04 \cdot I_p)$ and $(\beta_{T,j})_{j=1,\dots,4} \perp\!\!\!\perp X$.

The nature of $\mathcal{L}(C|X)$ is the same in every setting: $C|X \sim Weibull(\lambda_C e^{-{}^t\beta_C \cdot X}, k_C)$ with λ_C and β_C fitted to satisfy certain conditions (see Algorithm 5).

In each setting, we test the algorithm for two different functions ϕ: $\phi(t) = t$ and $\phi(t) = log(t+1)$. We also assess the sensitivity of the results to the censoring rate of the simulated data ($q = 0.1$, 0.3 or 0.5) and to the strength of the dependence between C and X (measured in terms of percentage of explained variance: $R2_C = 0.05$ or 0.1).

The process of data simulation is detailed in Algorithm 5.

Algorithme 5 : Simulated data generation

Input : n : number of simulated observations
p : dimension of X
q : proportion of censored observations
$R2_C$: proportion of explained variance for $C|X$
k_C : shape parameter of C
(λ_T, k_T) : scale and shape parameter for T
T_{max} : threshold for T

Output : Dataset of generated data

1 Draw ρ uniformly on $[0, 0.6]$
2 Simulate $(X_i)_{i\in 1,\dots,n}$ i.i.d. with $X \sim GaussianCopula(K^\rho)$ and marginals $Unif[-1, 1]$
3 Generate $(T_i)_{i=1,\dots,n}$ from case 1, 2 or 3
4 Draw $\beta_{C_0} \sim N(\mu = 0, \sigma^2 = 0.04 \cdot I_p)$, let $\beta_C = \eta \cdot \beta_{C_0}$, and calibrate η so that the empirical estimate of the explained variance $\widehat{R2}_C$ satisfies $\widehat{R2}_C \approx R2_C$
5 Calibrate λ_C so that the empirical proportion of censored observations $\hat{q}$ satisfies $\hat{q} \approx q$
6 Generate $(C_i)_{i=1,\dots,n}$ with $\forall i,\ C_i \sim Weibull(\lambda_C e^{-^t\beta_C \cdot X_i}, k_C)$
7 Check on the simulated $(C_i)_{i=1,..,n}$ that $\widehat{R2}_C \approx R2_C$ and $\hat{q} \approx q$
8 Build $\forall i = 1, \dots, n$, $Y_i = min(T_i, C_i)$, $\delta_i = \mathbb{1}_{T_i \leq C_i}$
9 Build $\forall i = 1, \dots, n$, $T'_i = min(T_i, T_{max})$, $Y'_i = min(T'_i, C_i)$, $\delta'_i = \mathbb{1}_{T'_i \leq C_i}$ (cf. Section 4.2.2. Remark: $\hat{q}$ at step 5 is derived from T' and not T)

return : Dataset with columns X, T, C, Y, δ, T', Y', δ'

We take $n = 2000$, $p = 6$, $\lambda_T = 100$, $k_T = 1$, and $k_C = 0.8$ as parameters for the data simulation. The parameter T_{max} is set to 1000.

Different models

In our experiments, we compare the performance of sixteen models. First, we have three *swRF1* models, associated with Kaplan-Meier (*swRF11*), Cox (*swRF12*) and random survival forest (*swRF13*) weights, respectively.

For *swRF2* and *swRF3*, many weight computations are necessary since weights are computed on the bootstrap samples. The use of RSF to compute the weights is thus

computationally intensive, so for these two models we only use KM (*swRF21* and *swRF31*) and Cox (*swRF22* and *swRF32*) weights.

We also test five other algorithms to compare with *swRF*. They all follow the same idea: first, fit a model to estimate $S_T(\cdot|X)$ (which gives an estimator $\hat{S}_T(\cdot|X)$), and second, integrate $\hat{S}_T(\cdot|X)$ to get an estimator of $E[\phi(T)|X = x]$: $\hat{f}(x) = -\int \phi \cdot d\hat{S}_T(\cdot|X = x)$. Five different models are used to estimate $S_T(\cdot|X)$ at the first step : the Cox model, the RSF algorithm, the forest of relative risk trees (RRT) algorithm of Ishwaran et al. (2004), and two versions of the reinforcement learning trees (RLT) algorithm described in the Section 4 of Zhu (2013) (one which uses reinforcement learning and one which does not). This gives five benchmarks for *swRF* that we call respectively *RSFr*, *Cr*, *RRTr*, *RLTr* (with reinforcement) and *nRLTr* (without reinforcement).

Since here we are working with simulated data, it is interesting to consider, as a baseline indicator, the performance of the random forest algorithm as if there was no censoring in the data. This random forest is trained on the non-censored data $(X_i, T_i)_{i=1,..n}$, and the associated model is denoted *RF*. Similarly, it is interesting to consider the exact IPCW weights W_i, as described in Section 4.2.2, for the use of *swRF*. The exact IPCW weights correspond to theoretical weights computed using the true conditional survival function of the censoring $S_C(\cdot|X)$ which is known in our simulated data application. This gives three additional comparison models called *swRF14*, *swRF24* and *swRF34*.

So that the different random forest models involved are comparable, all of the random forests in the simulations are set with the same values for the three parameters we evoked in Section 6. The parameter *mtry* is set to 6 so that $mtry = p$ and there is no randomness involved in split selection. We also take $maxdepth = 4$ and $minleaf = 50$, as justified in supplementary material.

The *swRF* models are trained on the threshold data (Y', δ', X) with $Y' = \min(T', C)$, setting $r_{max} = 50$ (see the definitions of T' and r_{max} in the Remark 4.1 of Section 4.2.3). *RF* is trained on the data (T', X), whereas *Cr*, *RSFr*, *RRTr*, *RLTr* and *nRLTr* are trained on the data (Y, δ, X). For the latter models, we then compute the prediction at the point x with $\hat{f}(x) = -\int \phi \cdot d\hat{S}_{T'}(\cdot|X = x)$, where $\hat{S}_{T'}(t|X) = \hat{S}_T(t|X) \cdot \mathbb{1}_{t \leq T_{max}}$.

4.3.2 Results and analysis

Performance of the models

To compare the models, we generated 100 simulated datasets of 2000 observations with Algorithm 5. For each iteration, we train all models on the same 1000 observations, and evaluate them on the 1000 remaining ones. Model accuracy is measured in terms of the mean squared error (MSE), which can can be computed in the simulated data context since we have access to all of the $\phi(T_i)$ values. The means of the MSE over the 100 i.i.d. replicates of the simulation process are given in Fig. 4.1. To keep the figure visually understandable, we chose to represent in the main text only the results for the most illuminating models. Nevertheless, we emphasize that the following analysis is consistent with the complete results, obtained for the sixteen models, that are given in supplementary material.

The results on simulated data help us to learn more about the various models. An initial observation is that the results are quite affected by the ϕ function being considered, and less by changes in the distribution of $\mathcal{L}(T|X)$. Of course, RF is generally the best model, except when the constrained form of the Cox model is well-suited to the problem, whereby Cr may outperform RF. Recall that the RF algorithm should only be seen as a benchmark that cannot be used in practice, since it relies on the complete (uncensored) observations. For $q = 0.1$, the performance of the $swRF$ models are very close to that of RF, which is natural since $swRF$ is equivalent to RF in the non-censored case (i.e., $q = 0$).

We can also see that $swRF$ models are more sensitive to an increase in censoring rate than the comparison methods which model directly $S_T(\cdot|X)$, namely: Cr, $RSFr$, $RRTr$ and $RLTr$. The $swRF$ models perform well for $q = 0.3$ and 0.1, but the MSE is much larger when $q = 0.5$. For $\phi(t) = t$, $RSFr$ and Cr perform very well overall, but for $\phi(t) = log(t + 1)$, the $swRF$ models are usually more accurate, especially when $q = 0.1$ and $q = 0.3$, or the data is simulated as a covariate-dependent mixture (Case 3).

We can also compare the $swRF$ models with each other. We first remark that the $swRF$ scores are organized in terms of the groups of $swRF$ ($swRF1$ and $swRF3$). While for $\phi(t) = t$ and $q = 0.5$ $swRF3$ models achieve high MSE due to particular iterations where $swRF3$ does not work well, the results for $\phi(t) = log(t + 1)$ are very different for $swRF1$ and $swRF3$. The difference is mostly due to the terminal leaf estimator used (this is confirmed by the results of $swRF2$ in supplementary material). Indeed, we can see that the results of $swRF3$ are closer to the results of $RSFr$, $RRTr$ and $RLTr$, and

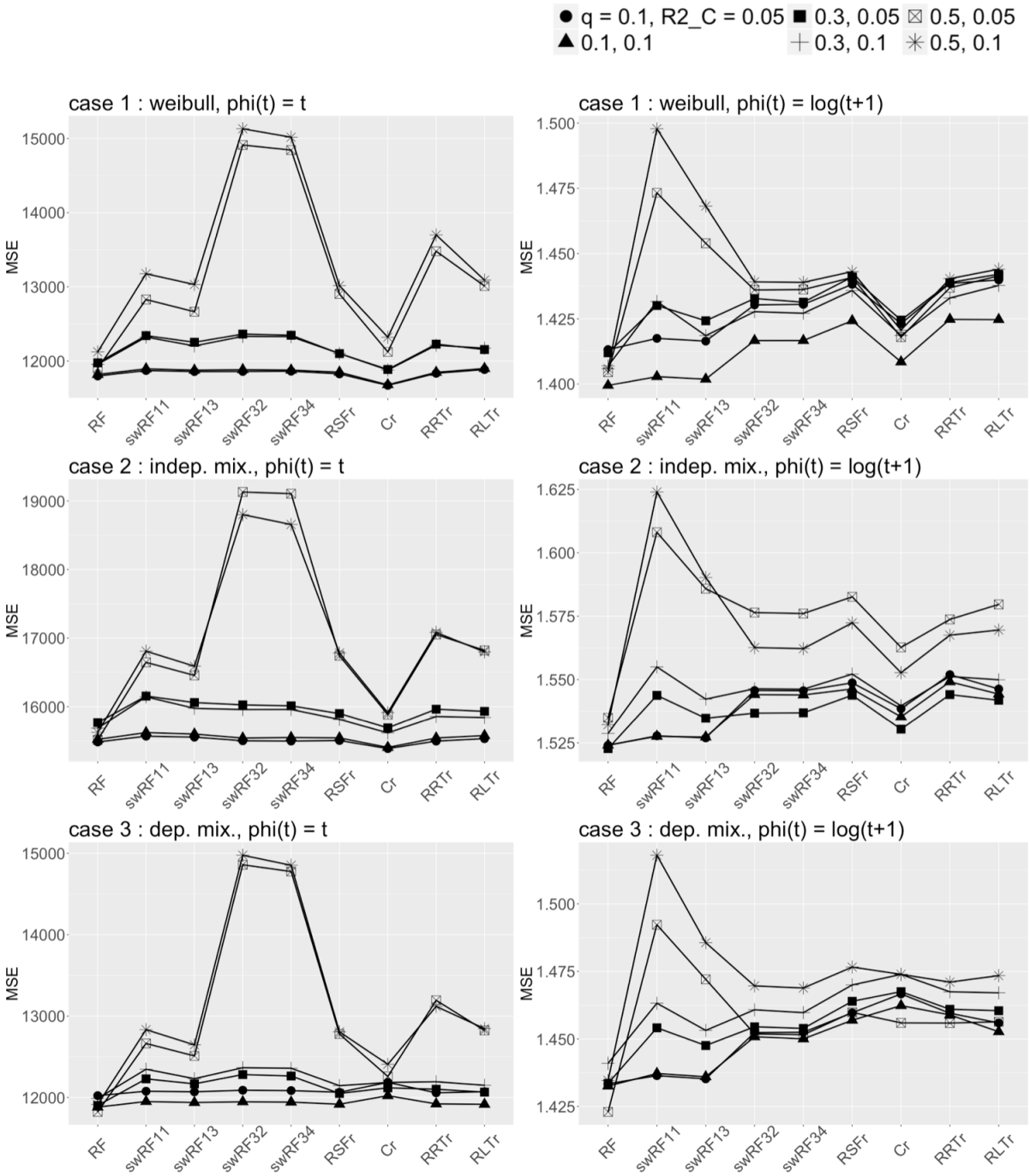

Fig. 4.1: Results on simulated data.

For each setting, the mean of the MSE values over 100 i.i.d. replicates of the simulation process is shown. The censoring rate q is equal to 0.1, 0.3, or 0.5, while the percentage of explained variance of C given X: $R2_C$, is set to 0.05 or 0.1.

the terminal leaf estimators used in these models are almost the same; $swRF3$ relies on a within leaf nonparametric estimation of the distribution of T using a KM estimator (i.e. $m^{\hat{w}_{loc}}$), whereas the other models rely on a Nelson-Allen estimator (Aalen (1978)) as explained in Ishwaran & Kogalur (2007). Also, the type of weights we consider has a second-order impact on the results, which becomes more significant with larger censoring rates. However, it is interesting to note that comparing the results of $swRF11$ and $swRF13$, we see that conditional weights computed with RSF tend to give better results than the Kaplan-Meier ones. By comparing the results of $swRF32$ and $swRF34$, we observe that Cox IPCW and exact IPCW lead to very close MSE. This shows that Cox IPCW (but also RSF IPCW as we will see in the next section) are good proxies for exact IPCW.

Correlations between weighted MSE (IPCW) and MSE

The results shown in Fig. 4.1 are estimated MSE which do not suffer from the effects of censoring. Since such estimators are not available for real data studies, in practice we rely – as explained in Section 4.2.4 – on weighted estimators (involving KM, Cox, RSF, or uniform weights) when comparing models. Therefore, it is of interest to compute, in the simulated data case, correlations between the non-censored MSE and weighted approximations of it. We added to the results the correlations with the exact (theoretical) weights which are available in the simulated data case. The average values of these correlations are displayed in Tab. 4.2. These results show that RSF and Cox weights better replicate the non-censored MSE than the KM ones, and therefore offer better comparison criteria for model selection. Let us note that RSF and Cox weights are even more accurate than exact weights. Such type of phenomena have already been observed in other censored regression models. For example, in the case of linear regression, Koul et al. (1981) noticed that the asymptotic variance of their slope estimator was smaller using KM weights rather than the true distribution function of the censoring (see their Remark 4.5 p. 1280). As a possible interpretation, as pointed by Koul et al. (1981), one may claims that the information contained in the censored observations would be completely lost if we were using exact weights, while estimation of the distribution of the censoring relies on the censored observations. On the other hand, the criterion based on uniform weighting of the non-censored observations has little correlation with the non-censored criterion, so we should not rely on it.

q	$R2_C$	RSF weights	Cox weights	exact weights	KM weights	unif. weights
0.10	0.05	0.91	0.91	0.90	0.90	0.70
0.10	0.10	0.92	0.92	0.91	0.91	0.70
0.30	0.05	0.78	0.77	0.72	0.74	0.20
0.30	0.10	0.76	0.75	0.70	0.66	0.18
0.50	0.05	0.72	0.71	0.66	0.65	0.14
0.50	0.10	0.61	0.64	0.59	0.49	0.09

Tab. 4.2: Correlations between weighted MSE (IPCW) and MSE.

Mean Spearman correlations between the non-censored and weighted (IPCW) MSE, as a function of q and $R2_C$. For each of the six cases, we have 600 calculations of the MSE and weighted MSE (2 choices for $\phi \times 3$ choices for $\mathcal{L}(T|X) \times 100$ iterations), and we calculate correlations between different fitting criteria.

4.4 Real data application

4.4.1 Modeling the churn behavior of policy holders

The problem of predictive modeling has gained interest in recent years in the insurance community, and optimization of underwriting is one relevant application (Frees et al. (2014)). For insurance contract brokers, forecasting customer value represents an opportunity to improve margins in a competitive environment (Cummins & Doherty (2006), Maas (2010)). Moreover, churn modeling is important for the estimation of customer value, as described in Verhoef & Donkers (2001). Here, we focus on applying the methodology developed in Section 4.2 to build a predictive model of the impact of churn on prospect value (as defined in Section 4.1) as predicted by insurance brokers.

A broker's approach to estimate the prospect values is given by the following multiplicative formula, where the hats indicate we are referring to estimated quantities:

$$\widehat{value} = \widehat{p}_{sub} \cdot \widehat{pr} \cdot \widehat{f}_{ew} \cdot \widehat{f}_{ch},$$

with $\widehat{p}_{sub}$ the probability of subscription, $\widehat{pr}$ the predicted premium, $\widehat{f}_{ew}$ the early withdrawal factor, and $\widehat{f}_{ch}$ the churn factor. We are only interested in modeling f_{ch} here. In this factorization, the churn factor only depends on the termination time of the contract via a function ϕ_{ch} as shown in Fig. 4.2.

The data that we study here is the customer base of an insurance broker from 1^{st} October 2009 to 31^{st} July 2016, and we focus in particular on complementary health

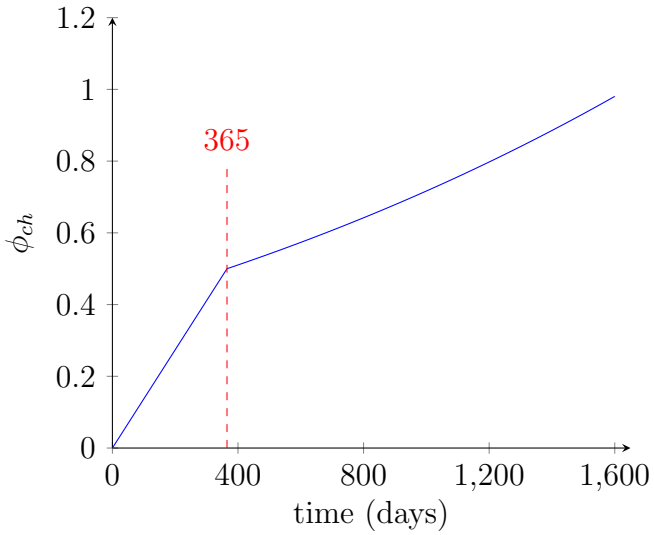

Fig. 4.2: The churn factor as a function of termination time.

Here, ϕ_{ch} is the broker's commission, expressed per unit of annual premium. It divides into two parts: a withholding part during the first year – which reaches 50% of the annual premium after one year – and a subsequent part, starting after one year, equal to 10% of annual premiums (taking into account annual resets of the premium).

insurance contracts. Data are available about the effective dates of the contracts, current states of contracts (active, terminated), and termination dates of contracts (if terminated). Before the underwriting, prospective information is given as 6 variables: age (8 levels starting from 18 years old), gender (2 levels: female, male), number of children insured (5 levels: 0, 1, 2, 3, ≥ 4), social security regime (7 levels: agricultural employee, employee, retired, retired self-employed, self-employed, student, unemployed), level of insurance (3 levels: low, medium, high), geographical zone (4 levels: Ile-de-France region, north, southern, other). Other client characteristics are available in the database but are only known after subscription occurs, hence it is impossible to use them to evaluate prospects.

We use this information to predict churn factors, which correspond to the variable $\phi(T)$. This variable is censored for a contract which is still active on the 31^{st} of July 2016. The censoring variable C is the age of the contract, i.e., the duration between the date of effect of the contract and the 31^{st} of July 2016. The 6 variables above constitute the covariate vector X.

4.4.2 Additional details about the experiments

The same models as in the simulated data application (see Section 9) are compared in this real data application, excluding RF and the $swRF$ models based on the exact weights ($swRF14$, $swRF24$, and $swRF34$), which cannot be used in the real data case. Also, T_{max} is set at 1465 days; this makes sense from the point of view of this application since we are then estimating the expected return of a prospect within the four years following subscription. All $swRF$, $RSFr$, $RRTr$ and $RLTr$ models are set with the parameters $mtry = 6$, $maxdepth = 5$, and $minleaf = 100$. For $swRF$ models, r_{max} is set to 50.

We test the various models with four different ϕ functions: $\phi = \phi_{ch}$ (Fig. 4.2), $\phi(t) = log(t+1)$, $\phi(t) = t$, and $\phi(t) = \mathbb{1}_{t>380}$. For the latter, T_{max} was set to 381 since, from 381 days on information on $\phi(T)$ is not censored.

4.4.3 Results and analysis

As in the simulated data case, we use a train-test approach (5000 observations each) for model testing, with 100 repetitions. Each training and testing set are drawn uniformly without replacement from the original dataset in such a way that the training and test sets do not overlap. The results in terms of MSE computed with KM, Cox and RSF weights, and on the C-index, are shown in Fig. 4.3. We give the results for the same models as in Section 4.3.2 except for RF and $swRF34$ which are not usable anymore with real data, and for $swRF22$ that we choose to incorporate into the figure. The results for the other models are given in supplementary material.

We first observe that the results for $\phi = \phi_{ch}$, $\phi(t) = log(t+1)$, and $\phi(t) = t$ look very similar, especially for $\phi = \phi_{ch}$ and $\phi(t) = log(t+1)$. Hence, the strong influence of the function ϕ that we noticed for the simulated data does not carry over to the real data. Moreover, there is a tendency in the results that $swRF$ models fitted using a certain type of weights perform very well with the MSE computed using the same type of weights. This is especially clear for the Cox and KM weights ($swRF11$ and $swRF22$ models), indicating that the type of weights considered has a real impact on the estimated distribution of (T, X).

We choose to consider the RSF weighted estimation of the MSE as the reference criteria for comparing the models. Indeed, we saw in Section 4.3.2 that Cox and RSF weights give approximations of the MSE that are the most correlated with the MSE computed on non-censored data. In addition, there is a risk of overfitting the type of weights we use,

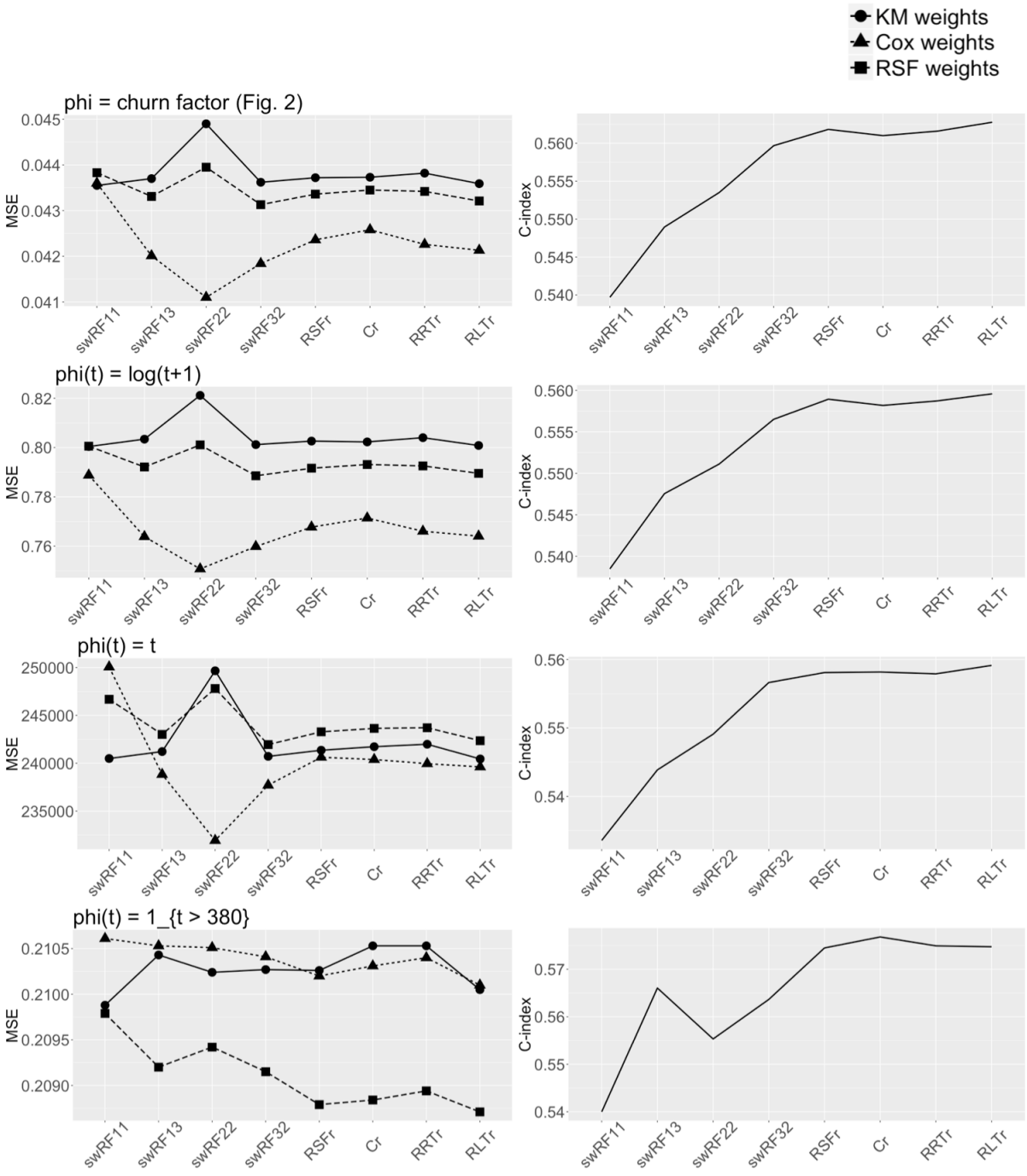

Fig. 4.3: Results on real data.

Left : the mean of the MSE values over 100 i.i.d. replicates of the simulation process, computed with KM, Cox or RSF weights. Right : the mean of the C-index over 100 i.i.d. replicates.

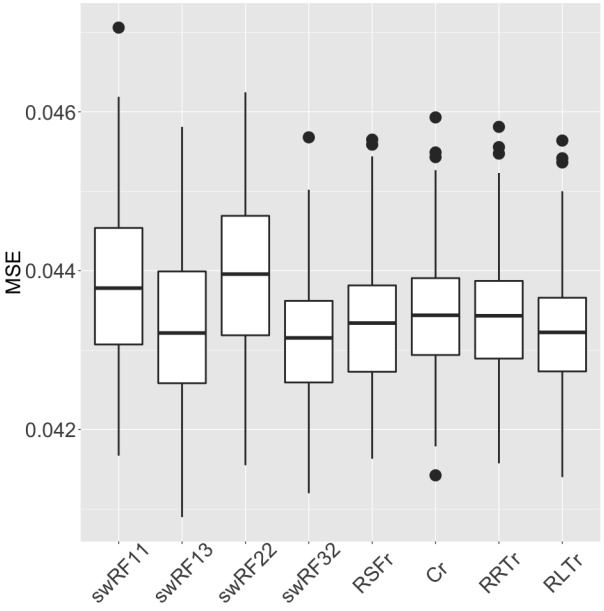

Fig. 4.4: Boxplot of the performances (MSE) of the models; $\phi = \phi_{ch}$, with RSF weights.

model	mean rank
$swRF11$	9.0
$swRF13$	4.8
$swRF22$	9.6
$swRF32$	2.8
$RSFr$	5.9
Cr	6.9
$RRTr$	6.9
$RLTr$	3.7

Tab. 4.3: Mean ranks of the models; $\phi = \phi_{ch}$, with RSF weights.

and since RSF weights are only involved in the model $swRF13$, they constitute a good choice. In Fig. 4.4 and Tab. 4.3, we show the performance of each model, calculated with RSF weights and for $\phi = \phi_{ch}$. Fig. 4.4 indicates the model $swRF32$ achieves on average the lowest error, followed by $RLTr$, an ordering confirmed by the table of rank statistics.

The C-index's statistic (Harrell et al. (1982)) corresponds to the proportion of ordered pairs in the test set which are well-ordered by the model $\hat{f}$, i.e. :

$$\frac{Card\left(\{(i,j) \in \mathcal{D}_{te}^2 / \hat{f}(X_i) > \hat{f}(X_j), Y_i > Y_j, \delta_j = 1\}\right)}{Card\left(\{(i,j) \in \mathcal{D}_{te}^2 / Y_i > Y_j, \delta_j = 1\}\right)}.$$

We can see in Fig. 4.3 that $RSFr$, Cr, $RRTr$ and $RLTr$ achieve the best C-index out of the set of models. This illustrates that it is important to consider a quadratic error criterion rather than a rank one when the goal is to estimate a mean value. It is the case in our situation since we are interested in getting the best prediction for the churn factor $f_{ch} = \phi_{ch}(T')$. The C-index values nevertheless show the benefit of using conditional weights within the $swRF$ algorithm in order to get models that rank the observations well.

4.5 Conclusion

In this paper, we have considered a class of weighted random forest algorithms, where the weight put on each observation is designed to compensate for censoring. A classical issue in censored regression models is the identifiability assumption that defines the dependence structure between the censoring and the variables involved in the model. Therefore, we have distinguished between two situations, namely the case where censoring is independent, and the case where this variable is allowed to depend on the duration variable T through the covariates. The latter case is the more general of the two, but leads to difficulties in computing appropriate weights, due to the strong impact of the dimension of the covariate vector. For this reason, we have proposed a compromise by modeling the conditional censoring distribution using a Cox model or a RSF model. We showed on simulated data that our method is competitive with other statistical methods in terms of accuracy. Also, this approach appeared to give the most accurate results in our application to modeling the commission received by an insurance broker, which is a non linear function of the time at which an insurance policyholder surrenders their contract. Moreover, it gives more accurate results than competing approaches such as random survival forest, relative risk forest, and reinforcement learning trees, in the setting we have considered. Finally, the weighting strategies we have proposed easily generalize to other regression-based approaches such as, to give one example, quantile regression.

As future work, it would be interesting to investigate the performances of the weighted random forest method in a setting where the dimension of the covariate vector is higher, which is an important application case (Zhu (2013), Ishwaran et al. (2010)). Indeed, in this situation the estimation of the conditional IPCW is hard and may require to adapt the algorithm. A suitable method might be to sample, for each tree of the forest, a subset of covariates to take into account in the computation of the conditional IPCW. Also, doubly robust survival trees studied in Steingrimsson et al. (2016) seems to be a great research direction. Promising results are already presented in Steingrimsson et al. (2018) using a relative risk tree to estimate $S_C(\cdot|X)$, and we might wonder if we could observe the same phenomenon in the results as we obtain in our work regarding the influence of the estimation of $S_C(\cdot|X)$: i.e. an improvement when using a conditional estimator instead of the KM estimator, and a benefit of using ensemble models such as RSF. Finally, theoretical work may be of interest as well; since the weighted random forest generalizes the (non-censored) random forest for regression, it may be possible to transpose consistency results obtained in the non-censored case (e.g. in Scornet et al.

(2015)) to the weighted case.

Software

All analyses were performed with the R package `sword` (`github.com/YohannLeFaou/sword`) which was developed for the purposes of this article. The code used for producing the results is available at the address:
`github.com/YohannLeFaou/impact-churn-health-insurance`.

Acknowledgements

We would like to thank the company Santiane, who provided the data that served for the experiments, and was always available to answer our questions. Also, we greatly thank the two referees for their valuable comments and their insightful suggestions.

4.6 Supplementary material

4.6.1 Choice of the parameters *minleaf* and *maxdepth*

The values taken by the parameters $minleaf$ and $maxdepth$ in the numerical applications of the article are optimized so that the different random forest algorithms achieve good performances. In this section, we first provide a sensitivity analysis performed on the simulated data presented in Section 3.1.1 which shows that the parameters $minleaf = 50$ and $maxdepth = 4$ are optimal for the majority of the random forest algorithms considered. We then discuss these optimal parameter values. Finally, we give the results of the sensitivity analysis performed on real data, along with some comments.

Study of the model's sensitivity to the parameters *minleaf* and *maxdepth* on simulated data

Setting and results : This study is based on the simulated datasets used in Section 3 which corresponds to a Weibull distribution (Case 1), a function $\phi(t) = \log(t+1)$, and a censoring rate $q \in \{0.1, 0.3, 0.5\}$. The same models as those represented on Fig. 1 are used, except the Cox model (Cr) which is not relevant in this analysis, and $swRF2$2 that we added to the compared models. Each model is evaluated under the settings

$maxdepth = 10$ & $minleaf \in \{10, 20, 50, 100, 200\}$, which are compared with the setting $maxdepth = 4$ & $minleaf = 50$. The means of the MSE over the 100 i.i.d. replicates of the simulation process are given in Fig. 4.5. For the sake of clarity, the models are divided in two subsets, each represented on the left side and right side of the figure.

For the cases $q = 0.1$ and $q = 0.3$, we can observe that every model, except $RRTr$ (when $q = 0.1$) and $swRF11$ (when $q = 0.3$), achieves its best performance with the setting $maxdepth \in \{4, 10\}$ & $minleaf = 50$. The results are more contrasted when the rate of censoring is higher ($q = 0.5$): while some models still achieve their best MSE with $maxdepth \in \{4, 10\}$ & $minleaf = 50$ ($swRF32$, $RSFr$, $RRRr$, $swRF34$), other models such as $swRF11$, $swRF13$ and $swRF22$ reach their best performances for $minleaf \in \{100, 200\}$. Thus, there is here a clear distinction between the models which use the Kaplan-Meier estimator in each terminal leaf and the models which employ in terminal leaves the IPCW used to grow the trees, these latter requiring more observations in each leaf to give good results.

Comments about the optimality of *minleaf*= 50 : The value $minleaf = 50$ is bigger than the $minleaf$ values usually reported in the literature. As an example, Zhu & Kosorok (2012) limit the number of observed failures in terminal nodes to six when measuring the performances of RSF and *recursively imputed survival trees* (RIST). Moreover, a common assertion found in the literature is that random forest algorithms perform well if the individual trees are grown to full size or nearly full size, which corresponds to small $minleaf$ values (e.g. $minleaf \leq 5$). For example, we refer to Sun (2010) or Biau & Scornet (2016). We would like to challenge this common belief in view of the results of our sensitivity analysis. Our results clearly demonstrate that, no matter which random forest model is used, a $minleaf$ of 10 is not big enough to reach the optimal range of MSE for our application. In fact, as supported by Segal (2004), the choice of the parameters which control the size of the trees involves a bias-variance trade-off. Scornet (2017) endorses this point of view, arguing that a high signal/noise ratio in the data, for a given classification of regression problem, causes large trees to perform well, while a lower signal/noise ratio leads to small trees performing better. The article analyzes the optimization of the parameters in the random forest algorithm and finds no theoretical reason to use the default values proposed by Breiman ($minsplit = 5$ for regression), concluding that optimizing the parameters which control the size of the terminal leafs and the size of the bootstrap samples improve the performance. From a practical point of view, many recent works insist on the importance of the tree size optimization for random forests : e.g. Boulesteix

et al. (2012), Huang & Boutros (2016), or Probst et al. (2018).

Sensitivity analysis on real data

The results of the sensitivity analysis performed on real data are given on Fig. 4.6. The chosen parameters $maxdepth = 5$ & $minleaf = 100$ are compared with the settings $maxdepth = 10$ & $minleaf \in \{50, 100, 200, 500\}$. The results justify our parameter choices and are coherent with the analysis made for simulated data (with $q = 0.5$). Indeed, our real data application is an example of a situation where the signal/noise ratio is low. The C-index values given in Fig. 3 are around 0.56, which is low compared to the results obtained with the various datasets (except the *transplant* dataset) in Ishwaran et al. (2008). The percentage of explained variance $R2$ is about 0.03, and thus quite low. Therefore, it is necessary to use small trees, with $maxdepth = 5$ & $minleaf = 100$ for a training set composed of 5000 observations, to achieve optimal performances with random forest algorithms.

4.6.2 Other results on simulated data

The Fig. 4.7 shows the complete results obtained for the sixteen models considered in the simulated data experiment of Section 3, completing the information presented in Fig. 1. Of course, the results presented on Fig. 4.7 are consistent with the analysis made in Section 3.2.1.

4.6.3 Other results on real data and further comments

The Fig. 4.8 & 4.9 complete the results obtained on real data presented in Section 4.3. It is worthwhile noting that the model $RLTr$ performs slightly better than the model $nRLTr$. It is not surprising to have such a small difference between the two models since the data is not high-dimensional, with only six covariates. The fact that $swRF32$ is the best model in our real data application suggests that conditional IPCW are effective at selecting the optimal splits in the early steps of the tree growing (with a split criteria being unbiased under the conditional independence assumption), but less reliable to estimate individual tree predictions in terminal leaves especially when the censoring rate is high and the leaves contain few non-censored observations.

Even if the explained variance of our model is only about three percent, it is still very useful to optimize a model for churn prediction, because at the aggregated level of

an insurance portfolio made of 200 000 policies, small improvements in the prediction of individual risks result in large impacts on the cash flows of the company. In fact, this situation of low signal/noise ratio is common in the domain of survival analysis. Sometimes, even when a model manages to rank the relative risks of the observations with a good accuracy, it can not predict the target duration with a low uncertainty. For instance in reliability analysis, it is usually hard to predict precisely the failure time of a component. This is illustrated in the work of Hong et al. (2009), who found in their study that the prediction intervals for the remaining lifetimes of power transformers are very large.

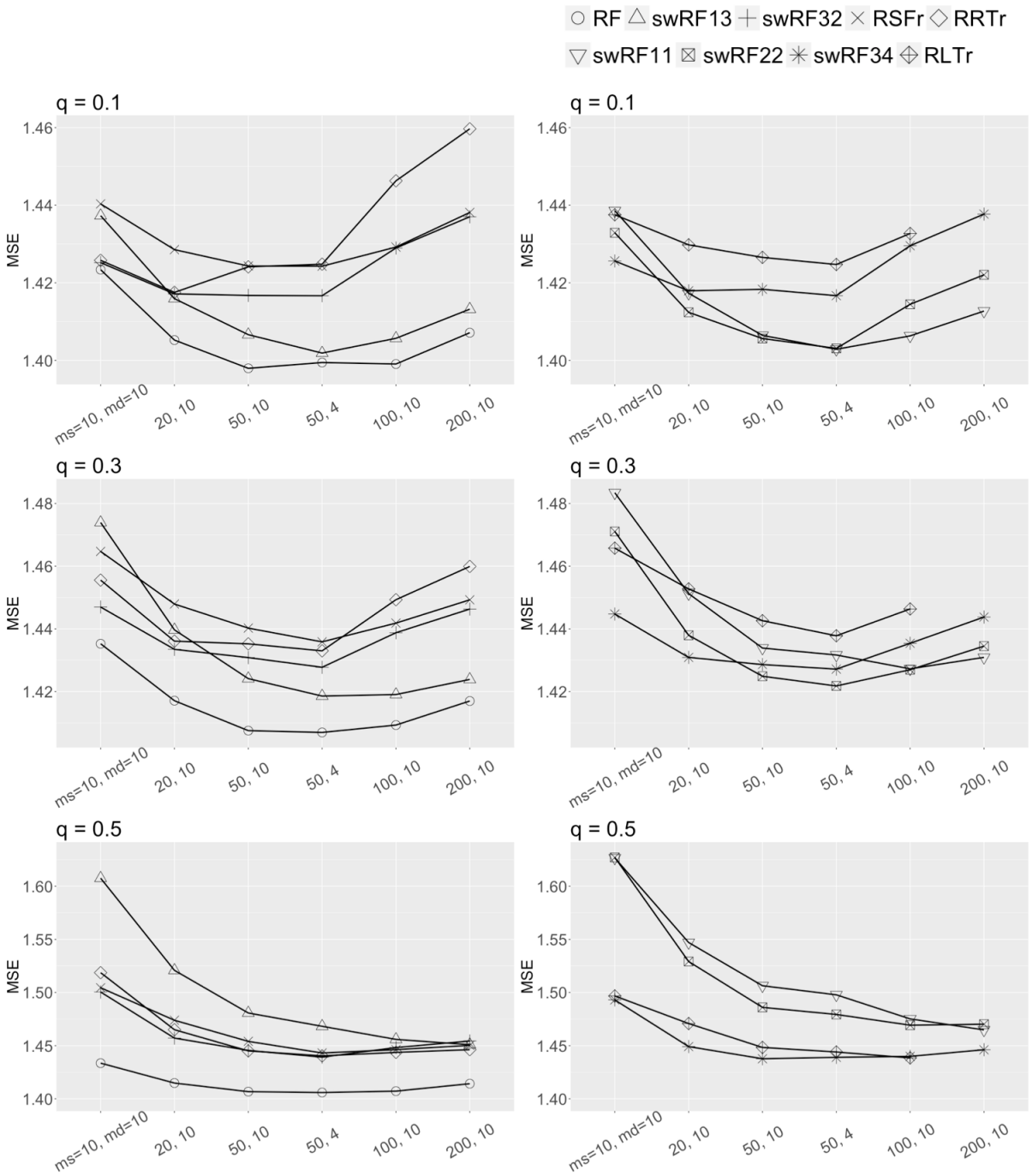

Fig. 4.5: Results of the sensitivity analysis on simulated data.

The mean of the MSE over the 100 i.i.d. replicates of the simulation process, for each random forest model and each pair of parameters *maxdepth* (*md*) & *minleaf* (*ms*). For *RLTr*, *embed.ntrees* is set to 10. For each random forest, $ntree = 100$ and $mtry = p = 6$.

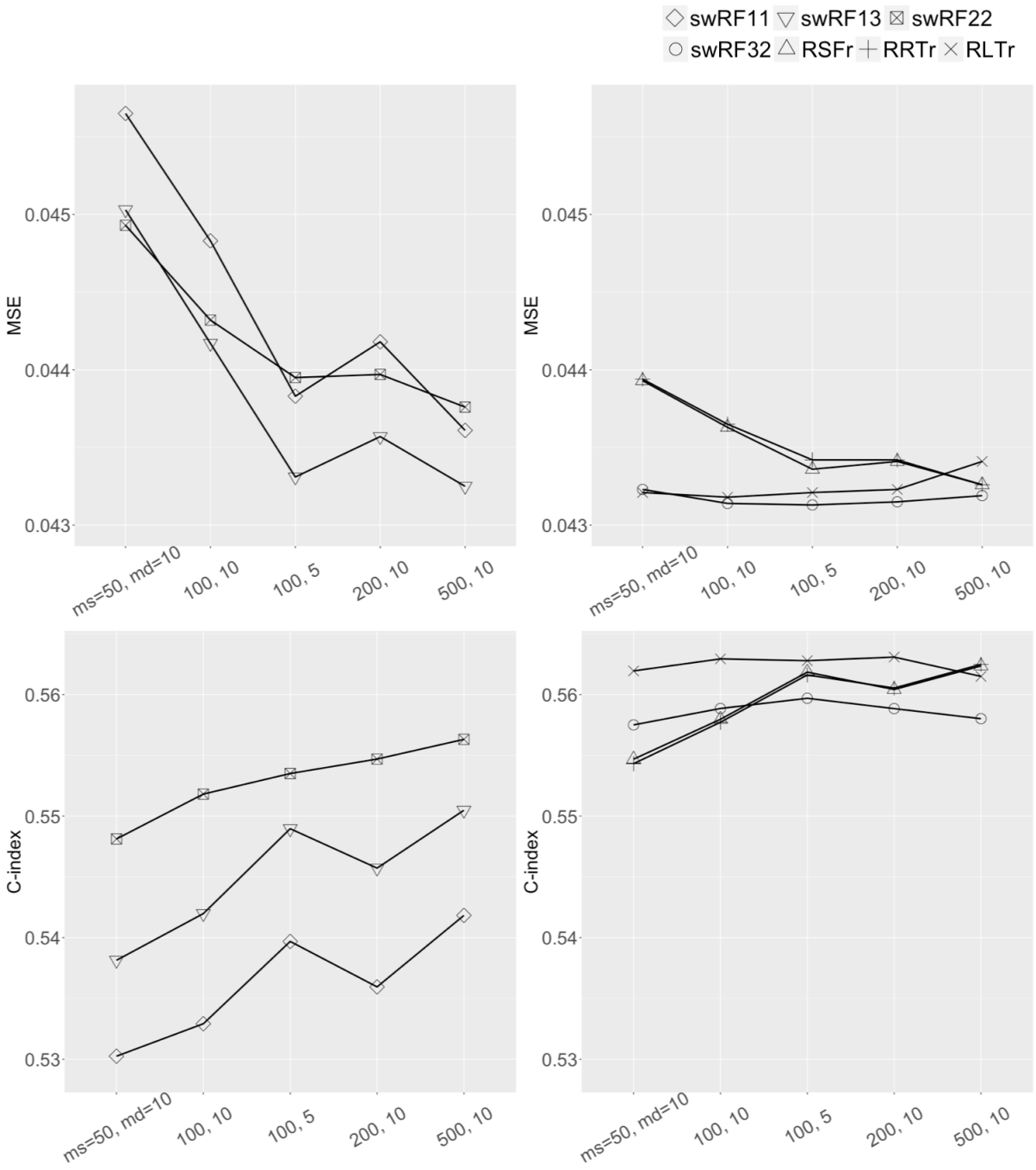

Fig. 4.6: Results of the sensitivity study on real data.

The mean of the MSE (estimated with RSF weights) and C-index over the 100 i.i.d. replicates of the simulation process, for each random forest model and each pair of parameters $maxdepth$ (md) & $minleaf$ (ms). For $RLTr$, $embed.ntrees$ is set to 10. For each random forest, $ntree = 100$ and $mtry = p = 6$.

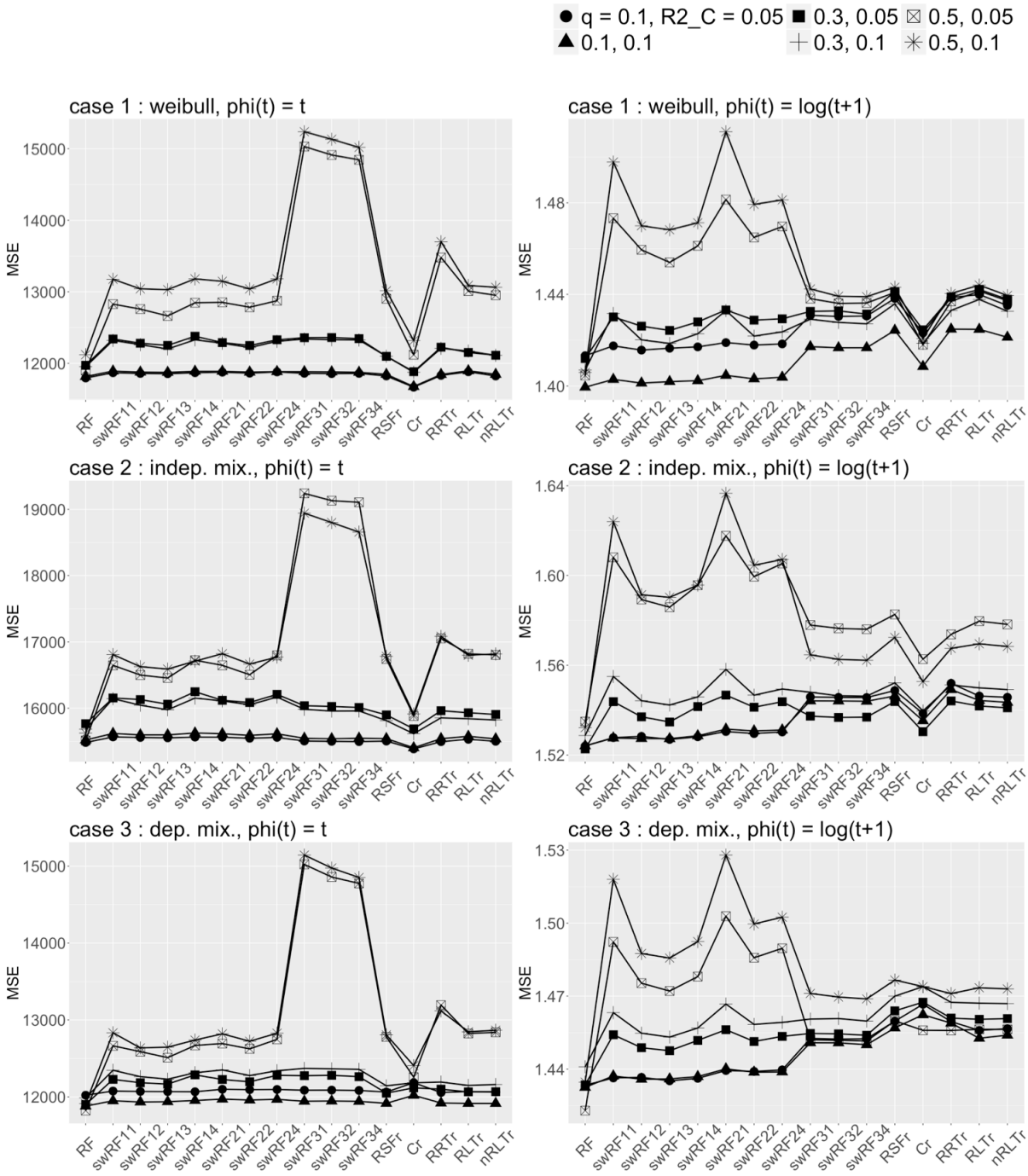

Fig. 4.7: Results on simulated data - all models

For each setting, the mean of the MSE values over 100 i.i.d. replicates of the simulation process is shown. The censoring rate q is equal to 0.1, 0.3, or 0.5, while the percentage of explained variance of C given X: $R2_C$, is set to 0.05 or 0.1.

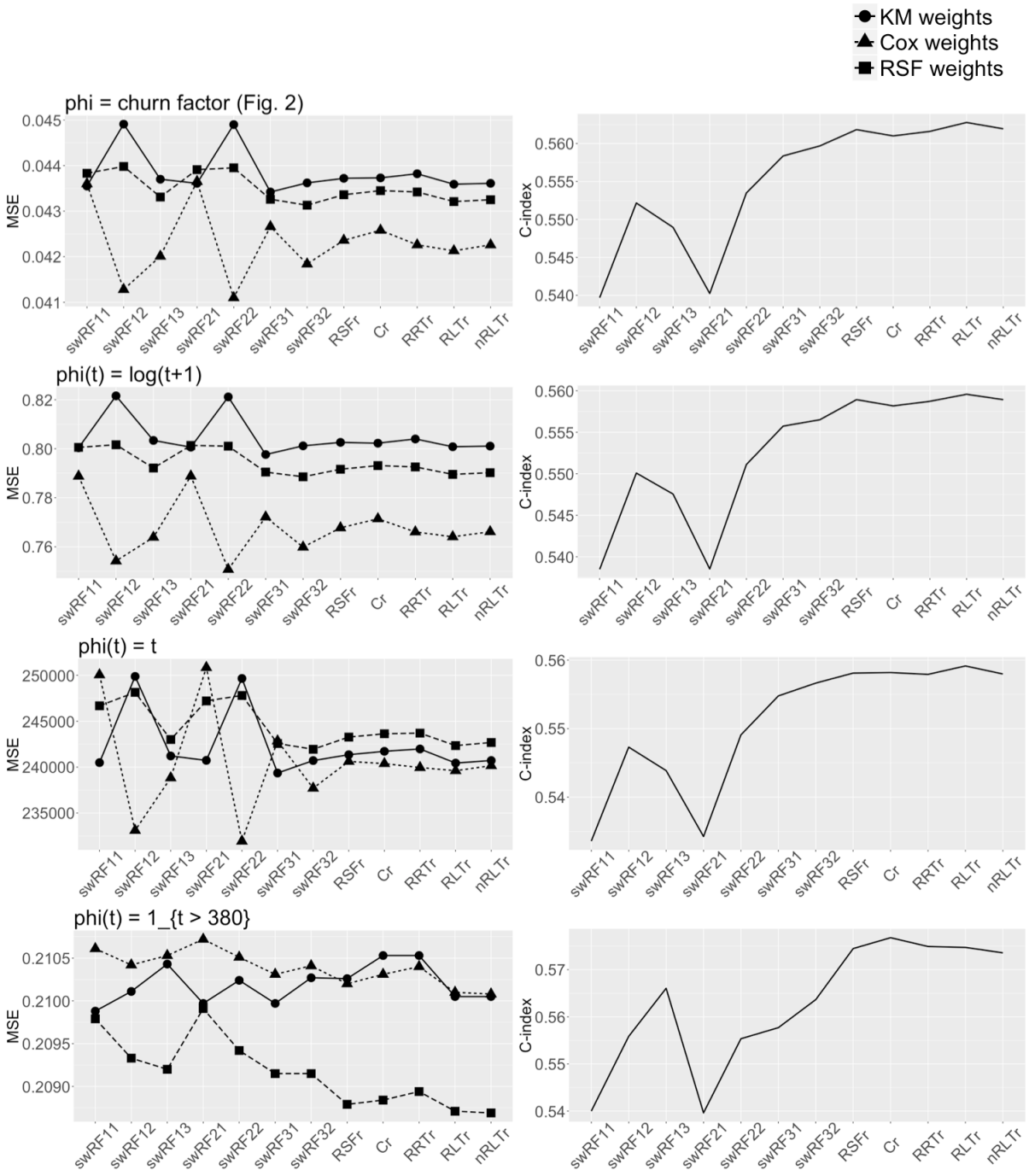

Fig. 4.8: Results on real data - all models

Left : the mean of the MSE values over 100 i.i.d. replicates of the simulation process, computed with KM, Cox and RSF weights. Right : the mean of the C-index over 100 i.i.d. replicates.

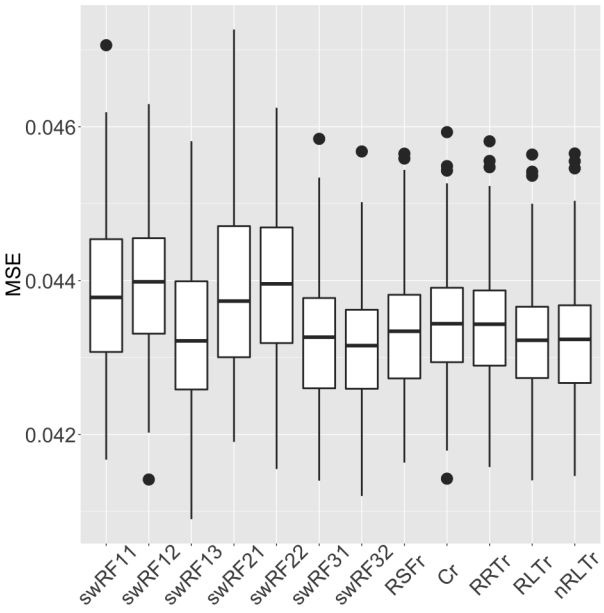

Fig. 4.9: Boxplot of the performances (MSE) of the models; $\phi = \phi_{ch}$, with RSF weights.

model	mean rank
$swRF11$	9.0
$swRF12$	9.5
$swRF13$	4.8
$swRF21$	9.6
$swRF22$	9.6
$swRF31$	4.7
$swRF32$	2.8
$RSFr$	5.9
Cr	6.9
$RRTr$	6.9
$RLTr$	3.7
$nRLTr$	4.4

Tab. 4.4: Mean ranks of the models; $\phi = \phi_{ch}$, with RSF weights.

Conclusion et perspectives

Le but de cette thèse était de mettre à profit les données recueillies par le courtier pour étudier la résiliation des contrats d'assurance complémentaire santé et approfondir les connaissances méthodologiques liées à l'analyse et à la prédiction des variables de durée. L'étude statistique des durées nécessite en effet un traitement spécifique du fait du phénomène de censure qui rend l'information disponible incomplète. Pour atteindre cet objectif la première étape, qui est l'objet du Chapitre 2, a été de comprendre les tenants et les aboutissants de l'activité du courtier, d'explorer les données disponibles et de définir le périmètre d'étude adéquat. Par la suite, nous avons travaillé sur deux problématiques visant à améliorer la compréhension des résiliations de contrats d'assurance chez les clients du courtier, et à optimiser le système de scoring des prospects utilisé par le courtier. Au Chapitre 3, nous avons proposé une méthode non paramétrique pour estimer la dépendance conditionnelle entre deux durée successives, et nous avons appliqué cette méthode pour étudier la dépendance entre la durée avant effet et la durée de résiliation d'un contrat. Au Chapitre 4, nous avons utilisé l'algorithme de forêt aléatoire conjointement au concept de poids IPCW pour optimiser la prédiction d'une quantité $E[\phi(T)|X = x]$ lorsque T est une variable de durée censurée à droite. Cette méthode a été appliquée sur les données du courtier pour améliorer l'estimation de la composante de résiliation dans le calcul du score d'un prospect.

Notre travail ouvre de nombreuses perspectives pour des recherches futures. En effet, sur le plan académique, les résultats théoriques de notre contribution du Chapitre 3 ont mis en évidence que la méthode d'estimation de copule conditionnelle développée souffre de la malédiction de la dimension lorsque le vecteur des variables explicatives X est multidimensionnel. Différentes méthodes, telles que les modèles single-index peuvent être envisagées pour améliorer les résultats dans une telle situation. Aussi, si l'on souhaitait étudier la dépendance entre plus que deux durées, il serait intéressant d'utiliser une approche par *vine copula*, qui permettrait de réduire le problème de l'estimation de la dépendance multivariée à plusieurs estimations bivariées. Le problème de la grande dimen-

sion est également une piste d'investigation intéressante pour ce qui concerne notre travail du Chapitre 4 sur la forêt aléatoire. Nous avons suggéré dans la conclusion du Chapitre 4 une méthode pour y remédier. Le développement de nouveaux algorithmes pour exploiter les données censurées à droite est un domaine actif de recherche, et nous pensons que des approches telles que le gradient boosting ou les réseaux de neurones pourraient apporter des améliorations. Sur le plan professionnel des métiers de l'assurance, nous pensons que de nombreux sujets restent à explorer en matière de modélisation des variables de durée, que ce soit pour étudier la durée de résiliation des contrats ou bien d'autres durées. Comme nous l'avons vu, la modélisation des durées intervient dans de nombreuses situations en assurance. Nous retenons deux axes d'investigation. Premièrement, l'utilisation de modèles dynamiques, qui puissent s'adapter à des modifications de l'environnement. Nous pensons par exemple à des outils de visualisation en temps réel, ou à des systèmes d'optimisation dynamique comme on en trouve aujourd'hui dans certaines entreprises de l'internet (publicité en ligne par exemple). De tels systèmes pourraient, dans le cas du courtier, permettre de s'adapter aux offres de la concurrence en temps réel. Deuxièmement, nous pensons au recueil et à l'exploitation de nouvelles données dans le but d'améliorer la connaissance client. L'acquisition de nouvelles données, structurées ou non structurées, est en effet rendue possible aujourd'hui du fait de la numérisation de l'économie. Toutes les pistes énumérées ici, suggèrent de nombreux travaux à venir, auxquels nous espérons contribuer.

References

Aalen, O. (1976). Nonparametric inference in connection with multiple decrement models. *Scandinavian Journal of Statistics*, 15–27.

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 701–726.

Abegaz, F., Gijbels, I., & Veraverbeke, N. (2012). Semiparametric estimation of conditional copulas. *Journal of Multivariate Analysis*, *110*, 43–73.

ACPR. (2018). Principes du conseil en assurance. *Publications de l'ACPR*.

Ahn, H., & Loh, W.-Y. (1994). Tree-structured proportional hazards regression modeling. *Biometrics*, 471–485.

Akritas, M. G., & Keilegom, I. V. (2003). Estimation of bivariate and marginal distributions with censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(2), 457–471.

Akritas, M. G., et al. (2000). The central limit theorem under censoring. *Bernoulli*, *6*(6), 1109–1120.

Altshuler, B. (1970). Theory for the measurement of competing risks in animal experiments. *Mathematical Biosciences*, *6*, 1–11.

Aubin, I., & Rolland, A. (2010). Lignes directrices de la construction des lois de maintien en incapacité et en invalidité. *Note méthodologique de l'Institut des Actuaires*.

Bargès, M., Cossette, H., & Marceau, E. (2009). Tvar-based capital allocation with copulas. *Insurance: Mathematics and Economics*, *45*(3), 348–361.

Bartram, S. M., Taylor, S. J., & Wang, Y.-H. (2007). The euro and european financial market dependence. *Journal of Banking & Finance*, *31*(5), 1461–1481.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, *S53*.

Benoumechiara, N., Michel, B., Saint-Pierre, P., & Bousquet, N. (2018). Detecting and modeling worst-case dependence structures between random inputs of computational reliability models. *arXiv preprint arXiv:1804.10527*.

Beran, R. (1981). *Nonparametric regression with randomly censored survival data* (Tech. Rep.). Univ. California, Berkeley.

Bernoulli, D. (1766). Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculation pour la prévenir. *Histoire de l'académie royale des sciences, année 1760 (disponible sur : http://gallica.bnf.fr)*.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197–227.

Biessy, G. (2017). Continuous-time semi-markov inference of biometric laws associated with a long-term care insurance portfolio. *ASTIN Bulletin: The Journal of the IAA*, *47*(2), 527–561.

Boudreault, M., Cossette, H., & Marceau, É. (2014). Risk models with dependence between claim occurrences and severities for atlantic hurricanes. *Insurance: Mathematics and Economics*, *54*, 123–132.

Boudreault, M., Cossette, H., & Marceau, E. (2017). On a joint frequency and severity loss model applied to earthquake risk.

Bou-Hamad, I., Larocque, D., & Ben-Ameur, H. (2011). A review of survival trees. *Statistics Surveys*, *5*, 44–71.

Boulesteix, A.-L., Janitza, S., Kruppa, J., & König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(6), 493–507.

Boumezoued, A., El Karoui, N., & Loisel, S. (2017). Measuring mortality heterogeneity with multi-state models and interval-censored data. *Insurance: Mathematics and Economics*, *72*, 67–82.

Bravais, A. (1844). *Analyse mathématique sur les probabilités des erreurs de situation d'un point*. Impr. Royale.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

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

Caragata Nasvadi, G., & Wister, A. (2009). Do restricted driver's licenses lower crash risk among older drivers? a survival analysis of insurance data from british columbia. *The Gerontologist*, *49*(4), 474–484.

Carriere, J. F. (2000). Bivariate survival models for coupled lives. *Scandinavian Actuarial Journal*, *2000*(1), 17–32.

Cebrian, A. C., Denuit, M., Lambert, P., et al. (2003). Analysis of bivariate tail dependence using extreme value copulas: An application to the soa medical large claims database. *Belgian Actuarial Journal*, *3*(1), 33–41.

Charpentier, A. (2013). Copules et risques multiples. *Chapter 6 in the book "Statistique du risque"*.

Charpentier, A., Fermanian, J.-D., & Scaillet, O. (2007). The estimation of copulas: Theory and practice.

Chauvigny, M., Devineau, L., Loisel, S., & Maume-Deschamps, V. (2011). Fast remote but not extreme quantiles with multiple factors: applications to solvency ii and enterprise risk management. *European Actuarial Journal*, *1*(1), 131–157.

Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, *99*(6), 323–329.

Cho, H. J., & Hong, S.-M. (2008). Median regression tree for analysis of censored survival data. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *38*(3), 715–726.

Ciampi, A., Thiffault, J., Nakache, J.-P., & Asselain, B. (1986). Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates. *Computational statistics & data analysis*, *4*(3), 185–204.

Clayton, D. G. (1978). 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.

Cossette, H., Côté, M.-P., Mailhot, M., & Marceau, E. (2014). A note on the computation of sharp numerical bounds for the distribution of the sum, product or ratio of dependent risks. *Journal of Multivariate Analysis*, *130*, 1–20.

Cossette, H., Marceau, E., & Mtalai, I. (2018). Collective risk models with dependence. *Available at SSRN: https://ssrn.com/abstract=3104912*.

Cossette, H., Marceau, E., Nguyen, Q. H., & Robert, C. Y. (2018). Tail approximations for sums of dependent regularly varying random variables under archimedean copula models. *Methodology and Computing in Applied Probability*, 1–30.

Cousin, A., & Di Bernardino, E. (2013). On multivariate extensions of value-at-risk. *Journal of multivariate analysis*, *119*, 32–46.

Cousin, A., & Di Bernardino, E. (2014). On multivariate extensions of conditional-tail-expectation. *Insurance: Mathematics and Economics*, *55*, 272–282.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society*, *34*, 187–220.

Cuberos, A., Masiello, E., & Maume-Deschamps, V. (2019). Copulas checker-type approximations: Application to quantiles estimation of sums of dependent random variables. *Communications in Statistics-Theory and Methods*, 1–19.

Cui, Y., Zhu, R., & Kosorok, M. (2017). Tree based weighted learning for estimating individualized treatment rules with censored data. *Electronic journal of statistics*, *11*(2), 3927–3953.

Cui, Y., Zhu, R., Zhou, M., & Kosorok, M. (2017). Consistency of survival tree and forest models: splitting bias and correction. *arXiv preprint arXiv:1707.09631v4*.

Cummins, J. D. (1973). *Development of life insurance surrender values in the united states*. SS Huebner Foundation for Insurance Education, Wharton School, University of ....

Cummins, J. D., & Doherty, N. A. (2006). The economics of insurance intermediaries. *Journal of Risk and Insurance*, *73*(3), 359–396.

Dabrowska, D. M. (1989, 09). Uniform consistency of the kernel conditional kaplan-meier estimate. *Ann. Statist.*, *17*(3), 1157–1167. Retrieved from `https://doi.org/10.1214/aos/1176347261`

Dabrowska, D. M., et al. (1988). Kaplan-meier estimate on the plane. *The Annals of Statistics*, *16*(4), 1475–1489.

Davis, R. B., & Anderson, J. R. (1989). Exponential survival trees. *Statistics in Medicine*, *8*(8), 947–961.

Deheuvels, P. (1979). La fonction de dependence empirique et ses proprietes, un test non parametrique d'independance. *Bulletin de la classe des sciences, Academie Royale de Belgique, 5e serie*, *65*, 274–292.

Denuit, M., Dhaene, J., Goovaerts, M., & Kaas, R. (2006). *Actuarial theory for dependent risks: measures, orders and models*. John Wiley & Sons.

Denuit, M., & Legrand, C. (2018). Risk classification in life and health insurance: extension to continuous covariates. *European Actuarial Journal*, 1–11.

Derumigny, A., & Fermanian, J.-D. (2017). About tests of the simplifying assumption for conditional copulas. *Dependence modeling*(5), 154 - 197.

Devineau, L., & Loisel, S. (2009). Risk aggregation in solvency ii: How to converge the approaches of the internal models and those of the standard formula? *Bulletin Français d'Actuariat*, *9*(18), 107–145.

Dias, A., Embrechts, P., et al. (2004). Dynamic copula models for multivariate high-frequency data in finance. *Manuscript, ETH Zurich*, *81*.

Di Bernardino, E., Maume-Deschamps, V., & Prieur, C. (2013). Estimating a bivariate tail: a copula based approach. *Journal of Multivariate Analysis*, *119*, 81–100.

Doukhan, P., Fermanian, J.-D., Lang, G., et al. (2005). *Copulas of a vector-valued stationary weakly dependent process*. INSEE.

Drees. (2017). Les dépenses de santé depuis 1950. *Études et résultats de la Drees*.

Drees. (2018). Les dépenses de santé en 2017 - résultats des comptes de la santé. *Panoramas de la Drees - santé*.

Dreyfus, M. (2011). L'histoire de la mutualité: quatre grands défis. *Les Tribunes de la sante*(2), 49–54.

Dutang, C. (2012). The customer, the insurer and the market. *Bulletin Français d'Actuariat*.

Einmahl, U., & Mason, D. M. (2000, Jan 01). An empirical process approach to the uniform consistency of kernel-type function estimators. *Journal of Theoretical Probability*, *13*(1), 1–37.

Einmahl, U., & Mason, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.*, *33*(3), 1380–1403.

Embrechts, P., De Haan, L., Huang, X., et al. (2000). Modelling multivariate extremes. *Extremes and integrated risk management*, 59–67.

Euler, L. (n.d.). Recherches générales sur la mortalité et la multiplication du genre humain. *disponible sur http://gallica.bnf.fr*.

Fantazzini, D., & Figini, S. (2009). Random survival forests models for sme credit risk measurement. *Methodology and computing in applied probability*, *11*(1), 29–45.

Favre, A.-C., El Adlouni, S., Perreault, L., Thiémonge, N., & Bobée, B. (2004). Multivariate hydrological frequency analysis using copulas. *Water resources research*, *40*(1).

Ferger, D., Manteiga, W. G., Schmidt, T., & Wang, J.-L. (2017). *From statistics to mathematical finance*. Springer.

Fermanian, J.-D., & Lopez, O. (2018). Single-index copulas. *Journal of Multivariate Analysis*, *165*, 27–55.

Fermanian, J.-D., Radulovic, D., Wegkamp, M., et al. (2004). Weak convergence of empirical copula processes. *Bernoulli*, *10*(5), 847–860.

Fermanian, J.-D., & Scaillet, O. (2003). Nonparametric estimation of copulas for time series. *Journal of Risk*, 25–54.

Fermanian, J.-D., & Scaillet, O. (2005). *Some statistical pitfalls in copula modelling for financical applications, e. klein (ed.), capital formation, gouvernance and banking.* Nova Science Publishing, New York.

Fleming, T. R., & Harrington, D. P. (2011). *Counting processes and survival analysis* (Vol. 169). John Wiley & Sons.

Frank, M. J. (1979). On the simultaneous associativity off (x, y) andx+y- f (x, y). *Aequationes mathematicae*, *19*(1), 194–226.

Frees, E. W., Derrig, R. A., & Meyers, G. (2014). *Predictive modeling applications in actuarial science* (Vol. 1). Cambridge University Press.

Frees, E. W., & Valdez, E. A. (1998). Understanding relationships using copulas. *North American actuarial journal*, *2*(1), 1–25.

Fu, W., & Simonoff, J. S. (2016). Survival trees for left-truncated and right-censored data, with application to time-varying covariate data. *Biostatistics*, *18*(2), 352–369.

Geerdens, C., Acar, E. F., & Janssen, P. (2018). Conditional copula models for right-censored clustered event time data. *Biostatistics*, *19*(2), 247-262.

Genest, C., Gendron, M., & Bourdeau-Brien, M. (2013). The advent of copulas in finance. In *Copulae and multivariate probability distributions in finance* (pp. 13–22). Routledge.

Genest, C., Ghoudi, K., & Rivest, L.-P. (1995a). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, *82*(3), 543–552.

Genest, C., Ghoudi, K., & Rivest, L.-P. (1995b). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, *82*(3), 543–552.

Genest, C., & MacKay, J. (1986a). Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données. *Canadian Journal of Statistics*, *14*(2), 145–159.

Genest, C., & MacKay, J. (1986b). The joy of copulas: Bivariate distributions with uniform marginals. *The American Statistician*, *40*(4), 280–283.

Genest, C., Nešlehová, J., & Ben Ghorbal, N. (2011). Estimators based on kendall's tau in multivariate copula models. *Australian & New Zealand Journal of Statistics*, *53*(2), 157–177.

Gerds, T. A., & Schumacher, M. (2006). Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, *48*(6), 1029–1040.

Gibaud, B. (1986). *De la mutualité à la sécurité sociale: conflits et convergences*. Editions de l'Atelier.

Gibaud, B. (2008). Mutualité/sécurité sociale (1945-1950): la convergence conflictuelle. *Vie sociale*(4), 39–52.

Gijbels, I., & Herrmann, K. (2014). On the distribution of sums of random variables with copula-induced dependence. *Insurance: Mathematics and Economics*, *59*, 27–44.

Gijbels, I., Omelka, M., & Veraverbeke, N. (2017). Nonparametric testing for no covariate effects in conditional copulas. *Statistics*, *51*(3), 475–509.

Gijbels, I., Veraverbeke, N., & Omelka, M. (2011). Conditional copulas, association measures and their applications. *Computational Statistics & Data Analysis*, *55*(5), 1919 - 1932.

Gill, R. (1983). Large sample behaviour of the product-limit estimator on the whole line. *Ann. Statist.*, *11*(1), 49–58.

Goldberg, Y., & Kosorok, M. R. (2017). Support vector regression for right censored data. *Electronic Journal of Statistics*, *11*(1), 532–569.

Goodwin, B. K., & Hungerford, A. (2014). Copula-based models of systemic risk in us agriculture: implications for crop insurance and reinsurance contracts. *American Journal of Agricultural Economics*, *97*(3), 879–896.

Graunt, J. (1662). Bills of mortality. *Natural and political observations*.

Gribkova, S., & Lopez, O. (2015). Non-parametric copula estimation under bivariate censoring. *Scandinavian Journal of Statistics*, *42*(4), 925–946. (10.1111/sjos.12144)

Gribkova, S., Lopez, O., & Saint-Pierre, P. (2013). A simplified model for studying bivariate mortality under right-censoring. *Journal of Multivariate Analysis*, *115*, 181–192.

Guibert, Q., Lopez, O., & Piette, P. (2017). Forecasting mortality rate improvements with a high-dimensional var.

Guibert, Q., & Planchet, F. (2017). Utilisation des estimateurs de kaplan-meier par génération et de hoem pour la construction de tables de mortalité prospectives. *disponible sur le HAL*.

Guibert, Q., & Planchet, F. (2018). Non-parametric inference of transition probabilities based on aalen-johansen integral estimators for acyclic multi-state models: application to ltc insurance. *Insurance: Mathematics and Economics*.

Guillén, M., Nielsen, J. P., Scheike, T. H., & Pérez-Marín, A. M. (2012). Time-varying effects in the analysis of customer loyalty: A case study in insurance. *Expert systems with Applications*, *39*(3), 3551–3558.

Gumbel, E. J. (1960). Bivariate exponential distributions. *Journal of the American Statistical Association*, *55*(292), 698–707.

Günther, C.-C., Tvete, I. F., Aas, K., Sandnes, G. I., & Borgan, Ø. (2014). Modelling and predicting customer churn from an insurance company. *Scandinavian Actuarial Journal*, *2014*(1), 58–71.

Haberman, S., & Pitacco, E. (1998). *Actuarial models for disability insurance*. CRC Press.

Hainaut, D. (2018). A neural-network analyzer for mortality forecast. *ASTIN Bulletin: The Journal of the IAA*, *48*(2), 481–508.

Hainaut, D., & Robert, C. Y. (2014). Credit risk valuation with rating transitions and partial information. *International Journal of Theoretical and Applied Finance*, *17*(07), 1450046.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, *247*(18), 2543–2546.

Henckaerts, R., Antonio, K., Clijsters, M., & Verbelen, R. (2018). A data driven binning strategy for the construction of insurance tariff classes. *Scandinavian Actuarial Journal*, *2018*(8), 681–705.

Henebry, K. L. (1997). A test of the temporal stability of proportional hazards models for predicting bank failure. *Journal of Financial and Strategic Decisions*, *10*(3), 1–11.

Hong, Y., Meeker, W. Q., McCalley, J. D., et al. (2009). Prediction of remaining life of power transformers based on left truncated and right censored lifetime data. *The Annals of Applied Statistics*, *3*(2), 857–879.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., & Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics*, *7*(3), 355–373.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, *15*(3), 651–674.

Hothorn, T., Lausen, B., Benner, A., & Radespiel-Tröger, M. (2004). Bagging survival trees. *Statistics in medicine*, *23*(1), 77–91.

Hougaard, P. (2012). *Analysis of multivariate survival data*. Springer Science & Business Media.

Huang, B. F., & Boutros, P. C. (2016). The parameter sensitivity of random forests. *BMC bioinformatics*, *17*(1), 331.

Institut des actuaires, I. (2006). Lignes directrices mortalité. *Note méthodologique de l'Institut des Actuaires*.

Ishwaran, H., Blackstone, E. H., Pothier, C. E., & Lauer, M. S. (2004). Relative risk forests for exercise heart rate recovery as a predictor of mortality. *Journal of the American Statistical Association*, *99*(467), 591–600.

Ishwaran, H., Gerds, T. A., Kogalur, U. B., Moore, R. D., Gange, S. J., & Lau, B. M. (2014). Random survival forests for competing risks. *Biostatistics*, *15*(4), 757–773.

Ishwaran, H., & Kogalur, U. B. (2007). Random survival forests for r. *R news*, *7*.

Ishwaran, H., & Kogalur, U. B. (2010). Consistency of random survival forests. *Statistics & probability letters*, *80*(13-14), 1056–1064.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The annals of applied statistics*, 841–860.

Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J., & Lauer, M. S. (2010). High-dimensional variable selection for survival data. *Journal of the American Statistical Association*, *105*(489), 205–217.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. Chapman and Hall/CRC.

Joe, H., Smith, R. L., & Weissman, I. (1992). Bivariate threshold methods for extremes. *Journal of the Royal Statistical Society: Series B (Methodological)*, *54*(1), 171–183.

Jondeau, E., & Rockinger, M. (2006). The copula-garch model of conditional dependencies: An international stock market application. *Journal of international money and finance*, *25*(5), 827–853.

Kaplan, E. L., & Meier, P. (1958a). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, *53*, 457–481.

Kaplan, E. L., & Meier, P. (1958b). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, *53*(282), 457–481.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1/2), 81–93.

Kim, C. (2005). Modeling surrender and lapse rates with economic variables. *North American Actuarial Journal*, *9*(4), 56–70.

Klein, J. P., & Moeschberger, M. L. (2006). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Koul, H., Susarla, V. v., & Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. *The Annals of Statistics*, 1276–1288.

Labit Hardy, H. (2012). *Modélisation de la durée de vie des contrats d'assurance santé* (Unpublished master's thesis). Institut de Science Financière et d'Assurances.

Lagakos, S. (1979). General right censoring and its impact on the analysis of survival data. *Biometrics*, 139–156.

Lakhal, L., Rivest, L.-P., & Beaudoin, D. (2009). Ipcw estimator for kendall's tau under bivariate censoring. *The International Journal of Biostatistics*, *5*(1).

Lakhal-Chaieb, M. (2010). Copula inference under censoring. *Biometrika*, *97*(2), 505–512.

LeBlanc, M., & Crowley, J. (1992). Relative risk trees for censored survival data. *Biometrics*, 411–425.

LeBlanc, M., & Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association*, *88*(422), 457–467.

Lescourret, L., & Robert, C. Y. (2006). Extreme dependence of multivariate catastrophic losses. *Scandinavian Actuarial Journal*, *2006*(4), 203–225.

Li, A. H., & Bradic, J. (2019). Censored quantile regression forests. *arXiv preprint arXiv:1902.03327*.

Logan, B. R., Klein, J. P., & Zhang, M.-J. (2008). Comparing treatments in the presence of crossing survival curves: an application to bone marrow transplantation. *Biometrics*, *64*(3), 733–740.

Lopez, O. (2012). A generalization of the kaplan–meier estimator for analyzing bivariate mortality under right-censoring and left-truncation with applications in model-checking for survival copula models. *Insurance: Mathematics and Economics*, *51*(3), 505–516.

Lopez, O. (2018). A censored copula model for micro-level claim reserving.

Lopez, O., Milhaud, X., & Thérond, P.-E. (2016). Tree-based censored regression with applications in insurance. *Electronic journal of statistics*, *10*(2), 2685–2716.

Lopez, O., Patilea, V., & Van Keilegom, I. (2013). Single index regression models in the presence of censoring depending on the covariates. *Bernoulli*, *19*(3), 721–747.

Lopez, O., & Saint-Pierre, P. (2012). Bivariate censored regression relying on a new estimator of the joint distribution function. *Journal of Statistical Planning and Inference*, *142*(8), 2440–2453.

Maas, P. (2010). How insurance brokers create value—a functional approach. *Risk Management and Insurance Review*, *13*(1), 1–20.

Mallet-Bricout, B. (2005). Loi" chatel": de nouvelles avancées dans la protection du consommateur. *droit et patrimoine*, 38.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, *50*, 163–170.

Marshall, A. W., & Olkin, I. (1967). A generalized bivariate exponential distribution. *Journal of Applied Probability*, *4*(2), 291–302.

Maume-Deschamps, V., Rullière, D., & Usseglio-Carleve, A. (2017). Quantile predictions for elliptical random fields. *Journal of Multivariate Analysis*, *159*, 1–17.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*(Jun), 983–999.

Meira-Machado, L., Sestelo, M., & Gonçalves, A. (2016). Nonparametric estimation of the survival function for ordered multivariate failure time data: A comparative study. *Biometrical Journal*, *58*(3), 623–634.

Milhaud, X. (2013). Exogenous and endogenous risk factors management to predict surrender behaviours. *ASTIN Bulletin: The Journal of the IAA*, *43*(3), 373–398.

Milhaud, X., & Dutang, C. (2018). Lapse tables for lapse risk management in insurance: a competing risk approach. *European Actuarial Journal*, *8*(1), 97–126.

Milhaud, X., Loisel, S., & Maume-Deschamps, V. (2011). Surrender triggers in life insurance: what main features affect the surrender behavior in a classical economic context? *Bulletin Français d'Actuariat*, *11*(22), 5–48.

Molinaro, A. M., Dudoit, S., & Van der Laan, M. J. (2004). Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis*, *90*(1), 154–177.

Molinaro, A. M., Olshen, A., & Strawderman, R. (2014). Tree derived survival risk groups in differentiating risk for glioma patients. *Neuro-oncology*, *16*(Suppl 3), iii2.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, *9*(1), 141–142.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology*, *1*(1), 27–52.

Nolan, D., & Pollard, D. (1987). U-processes: rates of convergence. *The Annals of Statistics*, 780–799.

Oakes, D. (2008). On consistency of kendall's tau under censoring. *Biometrika*, *95*(4), 997–1001.

Outreville, J. F. (1990). Whole-life insurance lapse rates and the emergency fund hypothesis. *Insurance: Mathematics and Economics*, *9*(4), 249–255.

Paris, V., Devaux, M., & Wei, L. (2010). Health systems institutional characteristics.

Patton, A. J. (2004). On the out-of-sample importance of skewness and asymmetric dependence for asset allocation. *Journal of Financial Econometrics*, *2*(1), 130–168.

Patton, A. J. (2006a). Estimation of multivariate models for time series of possibly different lengths. *Journal of applied econometrics*, *21*(2), 147–173.

Patton, A. J. (2006b). Modelling asymmetric exchange rate dependence. *International economic review*, *47*(2), 527–556.

Peto, R., & Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, 185–207.

Planchet, F. (2005). Tables de mortalité d'expérience pour des portefeuilles de rentiers. *Note méthodologique de l'Institut des Actuaires*.

Planchet, F., & Winter, P. (2010). L'utilisation des splines bidimensionnels pour l'estimation de lois de maintien en arr\ˆ et de travail. *arXiv preprint arXiv:1001.1907*.

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2018). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, e1301.

Renshaw, A., & Haberman, S. (1986). Statistical analysis of life assurance lapses. *Journal of the Institute of Actuaries*, *113*(3), 459–497.

Satten, G. A., & Datta, S. (2001). The kaplan–meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician*, *55*(3), 207–210.

Schmid, M., Wright, M. N., & Ziegler, A. (2016). On the use of harrell's c for clinical risk prediction via random survival forests. *Expert Systems with Applications*, *63*, 450–459.

Scornet, E. (2017). Tuning parameters in random forests. *ESAIM: Proceedings and Surveys*, *60*, 144–162.

Scornet, E., Biau, G., & Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, *43*(4), 1716–1741.

Segal, M. R. (1988). Regression trees for censored data. *Biometrics*, 35–47.

Segal, M. R. (2004). *Machine learning benchmarks and random forest regression* (Tech. Rep.). UCSF: Center for Bioinformatics and Molecular Biostatistics.

Shih, J. H. (1998). Modeling multivariate discrete failure time data. *Biometrics*, 1115–1128.

Shih, J. H., & Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 1384–1399.

Shorack, G. R., & Wellner, J. A. (2009). *Empirical processes with applications to statistics* (Vol. 59). Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. (Reprint of the 1986 original [MR0838963])

Silvapulle, P., Kim, G., Silvapulle, M. J., et al. (2004). Robustness of a semiparametric estimator of a copula. In *Econometric society 2004 australasian meetings.*

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, *8*, 229–231.

Spearman, C. (1904). " general intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–292.

Steingrimsson, J. A., Diao, L., Molinaro, A. M., & Strawderman, R. L. (2016). Doubly robust survival trees. *Statistics in medicine*, *35*(20), 3595–3612.

Steingrimsson, J. A., Diao, L., & Strawderman, R. L. (2018). Censoring unbiased regression trees and ensembles. *Journal of the American Statistical Association*, 1–14.

Stepanova, M., & Thomas, L. (2002). Survival analysis methods for personal loan data. *Operations Research*, *50*(2), 277–289.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The annals of Statistics*, 1348–1360.

Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis*, *45*(1), 89–103.

Stute, W. (1995). The central limit theorem under random censorship. *Ann. Statist.*, *23*(2), 422–439.

Stute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scandinavian journal of statistics*, 461–471.

Stute, W. (1999). Nonlinear censored regression. *Statistica Sinica*, 1089–1102.

Stute, W. (2003). Kaplan–meier integrals. In *Advances in survival analysis* (Vol. 23, p. 87 - 104). Elsevier. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0169716103230054`

Stute, W., & Wang, J.-L. (1993). The strong law under random censorship. *The Annals of Statistics*, 1591–1607.

Sun, Y. V. (2010). Multigenic modeling of complex disease by random forests. In *Advances in genetics* (Vol. 72, pp. 73–99). Elsevier.

Sécu. (2018). Les chiffres clés de la sécurité sociale 2017. *Édition de la Sécurité sociale*.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, *72*(1), 20–22.

Tsukahara, H. (2005). Semiparametric estimation in copula models. *Canadian Journal of Statistics*, *33*(3), 357–375.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., & Wei, L. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, *30*(10), 1105–1117.

Van den Poel, D., & Lariviere, B. (2004a). Customer attrition analysis for financial services using proportional hazard models. *European journal of operational research*, *157*(1), 196–217.

Van den Poel, D., & Lariviere, B. (2004b). Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services. *Expert Systems with Applications*, *27*(2), 277–285.

Van Der Laan, M. J., et al. (1996). Efficient estimation in the bivariate censoring model and repairing npmle. *The Annals of Statistics*, *24*(2), 596–627.

Van der Laan, M. J., & Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media.

Van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge University Press.

Van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence*. Springer.

Van Keilegom, I., & Akritas, M. G. (1999). Transfer of tail information in censored regression models. *Annals of Statistics*, 1745–1784.

Van Keilegom, I., Akritas, M. G., & Veraverbeke, N. (2001). Estimation of the conditional distribution in regression with censored data: a comparative study. *Computational Statistics & Data Analysis*, *35*(4), 487–500.

Van Keilegom, I., & Veraverbeke, N. (2001). Hazard rate estimation in nonparametric regression with censored data. *Annals of the Institute of Statistical Mathematics*, *53*(4), 730–745.

Veraverbeke, N., Gijbels, I., & Omelka, M. (2011). Estimation of a conditional copula and association measures. *Scandinavian Journal of Statistics*, *38*(4), 766-780.

Verbelen, R., Antonio, K., & Claeskens, G. (2016). Multivariate mixtures of erlangs for density estimation under censoring. *Lifetime data analysis*, *22*(3), 429–455.

Verbelen, R., Gong, L., Antonio, K., Badescu, A., & Lin, S. (2015). Fitting mixtures of erlangs to censored and truncated data using the em algorithm. *ASTIN Bulletin: The Journal of the IAA*, *45*(3), 729–758.

Verhoef, P. C., & Donkers, B. (2001). Predicting customer potential value an application in the insurance industry. *Decision support systems*, *32*(2), 189–199.

Wang, W., & Wells, M. T. (1998). Nonparametric estimation of successive duration times under dependent censoring. *Biometrika*, *85*(3), 561–572.

Wang, W., & Wells, M. T. (2000). Estimation of kendall's tau under censoring. *Statistica Sinica*, 1199–1215.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 359–372.

Wüthrich, M. V. (2018). Machine learning in individual claims reserving. *Scandinavian Actuarial Journal*, *2018*(6), 465–480.

Xue, X., & Brookmeyer, R. (1996). Bivariate frailty model for the analysis of multivariate survival time. *Lifetime Data Analysis*, *2*(3), 277–289.

Youn, H., & Shemyakin, A. (1999). Statistical aspects of joint life insurance pricing. *1999 Proceedings of the Business and Statistics Section of the American Statistical Association*, *34138*.

Youn, H., & Shemyakin, A. (2001). Pricing practices for joint last survivor insurance. *Actuarial Research Clearing House*, *1*(2), 3.

Zhu, R. (2013). *Tree-based methods for survival analysis and high-dimensional data* (Unpublished doctoral dissertation). The University of North Carolina at Chapel Hill.

Zhu, R., & Kosorok, M. R. (2012). Recursively imputed survival trees. *Journal of the American Statistical Association*, *107*(497), 331–340.

Zhu, R., Zeng, D., & Kosorok, M. R. (2015). Reinforcement learning trees. *Journal of the American Statistical Association*, *110*(512), 1770–1784.

Résumé

Dans cette thèse, nous nous intéressons aux modèles de durée dans le contexte de la modélisation des durées de résiliation de contrats d'assurance santé. Identifié dès le 17ème siècle et les études de Graunt (1662) sur la mortalité, le biais induit par la censure des données de durée observées dans ce contexte doit être corrigé par les modèles statistiques utilisés. À travers la problématique de la mesure de la dépendance entre deux durées successives, et la problématique de la prédiction de la durée de résiliation d'un contrat d'assurance, nous étudions les propriétés théoriques et pratiques de différents estimateurs basés sur une méthode de pondération des observations (méthode dite IPCW) visant à corriger ce biais. L'application de ces méthodes à l'estimation de la valeur client en assurance est également détaillée.

Mots-clés : modèles de durée, données censurées, assurance, copule, forêt aléatoire

Abstract

In this thesis, we study duration models in the context of the analysis of contract termination time in health insurance. Identified from the 17th century and the original work of Graunt (1662) on mortality, the bias induced by the censoring of duration data observed in this context must be corrected by the statistical models used. Through the problem of the measure of dependence between successives durations, and the problem of the prediction of contract termination time in insurance, we study the theoretical and practical properties of different estimators that rely on a proper weighting of the observations (the so called IPCW method) designed to compensate this bias. The application of these methods to customer value estimation is also carefully discussed.

Keywords : survival analysis, censored data, insurance, copula, random forest