UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ANALYSE DE LA DÉPENDANCE INTRA-TRAJECTOIRE DANS LA MODÉLISATION INDIVIDUELLE DES RÉSERVES EN ASSURANCE NON-VIE

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN MATHÉMATIQUES

PAR

MARIE MICHAELIDES

JUIN 2024

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ANALYSIS OF INTER-RISKS DEPENDENCE IN MICRO-LEVEL CLAIMS RESERVING FOR NON-LIFE INSURANCE

THESIS

PRESENTED

AS PARTIAL REQUIREMENT

OF THE DOCTORATE IN MATHEMATICS

BY

MARIE MICHAELIDES

JUNE 2024

REMERCIEMENTS

Je tiens tout d'abord à remercier mes co-directeurs de thèse, Mathieu Pigeon et Hélène Cossette. Mathieu, merci d'avoir toujours été si disponible pendant ma thèse et même avant, lors de notre première rencontre à Louvain-la-Neuve. Savoir qu'à tout moment je pouvais te contacter en cas de question m'a été d'une grande aide dès mon arrivée à Montréal. Merci aussi pour ton soutien et ton engagement qui ont mené à l'aboutissement de cette thèse et m'ont permis de vivre de nombreuses expériences enrichissantes ces trois dernières années. Hélène, merci pour ton soutien, ta grande implication dès le début de ma thèse et merci de m'avoir accueillie à de nombreuses reprises à Québec. Merci pour nos échanges toujours stimulants qui m'ont très souvent inspirée et motivée à me dépasser. Vous m'avez tous les deux énormément appris et c'est grâce à vous que je désire maintenant poursuivre la recherche après mon doctorat.

Je remercie également la Chaire Co-operators en analyse des risques actuariels pour le soutien financier et pour les données qui m'ont permis de mener à bien cette thèse. Merci à Jean-Philippe Boucher pour son accueil et ses pokemon, et merci aussi aux autres étudiants de la chaire pour les chouettes activités passées ensemble. Je remercie également les étudiants d'Hélène Cossette et Étienne Marceau à l'université Laval pour leur accueil et les bons moments partagés ensemble lors de mes séjours à Québec ou lors de conférences et séminaires.

Mes remerciements vont aussi aux membres du jury, Jean-Philippe, Katrien Antonio et Yang Lu, qui ont accepté d'évaluer ma thèse.

Je tiens aussi à remercier Hélène Guérin, Arthur Charpentier et leurs différents étudiants qui se sont succédés dans le bureau en face du mien. Merci de m'avoir toujours inclue dans divers évènements et pour tous les bons moments partagés.

Merci à ma famille, particulièrement à mon père Philippe qui a toujours cru en moi, ma soeur Margaux pour sa patience et son expertise en français et mon frère Alexandre pour m'avoir toujours motivée et fait rire. Être loin de vous n'est pas facile tous les jours mais c'est grâce à votre soutien que j'ai pu mener à bien cette thèse.

Je remercie également mes amis, de Belgique, du Québec and from everywhere else. Dank ook aan mijn

voormalige collega's en professoren die mij hebben geholpen om dit punt te bereiken.

Sist men inte minst vill jag tacka min pojkvän, Jack. Tack för ditt stora stöd och för att du tror på mig när jag inte gör det. Jag hade inte kunnat göra det utan dig och jag älskar dig (mer).

En mémoire de ma grand-mère, Gilberte Vanderkel, pour toujours ma source d'inspiration.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉ

Cette thèse explore l'impact de la modélisation de la dépendance dans les modèles de micro-réserves en assurance non-vie. Compte tenu de la quantité croissante de données dont disposent les assureurs, nous développons des modèles qui reflètent mieux l'évolution des sinistres tout au long de leur durée de vie dans les portefeuilles des compagnies d'assurance. Par ailleurs, avec l'essor des ressources technologiques, ainsi qu'avec l'expansion de nombreuses compagnies à travers différents pays et différentes activités, leurs portefeuilles de risque se sont plus que jamais diversifiés. Pour modéliser les structures de plus en plus complexes des portefeuilles des assureurs, nous explorons dans cette thèse divers modèles de dépendance qui peuvent être incorporés dans des modèles de provisionnement. Plus spécifiquement, nous nous concentrons tout au long de nos travaux sur la modélisation de la dépendance entre plusieurs couvertures d'assurance fournies au sein des mêmes polices de portefeuilles. Dans un premier chapitre, nous construisons un modèle de provisionnement granulaire complet, en utilisant une régression logistique multinomiale pour les couvertures dépendantes. Ce modèle nous permet de prédire l'évolution complète d'une réclamation, de sa déclaration à son règlement. Dans le deuxième chapitre, nous utilisons des copules archimédiennes pour capturer cette dépendance en cas de divers types de censure. Au lieu d'ajuster un modèle de copule prédéfini, nous présentons une méthode par laquelle les assureurs peuvent bénéficier de la grande quantité de données à leur disposition, via un estimateur non paramétrique pour la fonction génératrice des copules de cette famille. Nous démontrons comment effectuer des simulations directement à partir de l'estimation de ce générateur, menant à des prédictions plus proches des données originales que si nous nous étions limités à un modèle de copule existant. Dans le troisième chapitre, nous étendons le modèle du deuxième chapitre en introduisant un estimateur paramétrique pour les générateurs de copules archimédiennes. Nous démontrons comment cette nouvelle approche permet d'incorporer des variables explicatives dans le modèle, et d'évaluer leur impact à la fois sur la force mais aussi sur la forme de la dépendance.

ABSTRACT

This thesis explores the impact of dependence modelling in micro-level claims reserving for non-life insurance. Considering the increasing quantities of data available to insurers, we investigate models that better reflect the development of claims throughout their lifetimes in the portfolios. In addition, with the rise of technological resources, as well as with the expansion of a lot of insurance companies across countries and businesses, insurance portfolios have become more diversified than ever before. To capture these intricate portfolio structures, we explore in this thesis various dependence models that can be incorporated in loss reserving frameworks. More specifically, we focus throughout our work on modelling the dependence between multiple insurance coverages provided within policies of insurance portfolios. In a first chapter, we build a full reserving model, using a multinomial logit regression for dependent coverages. This model allows us to predict the full development of a claim from declaration to settlement. In the second chapter, we use Archimedean copulas to capture this dependence in the presence of highly flexible censoring scenarios. Instead of fitting a predefined copula model, we show how insurers can benefit from their large datasets by presenting a non-parametric estimator for the generator function of copulas from this particular family. We demonstrate how simulations can then be performed directly from this estimated generator, leading to predictions that are closer to the original data than if we had constrained ourselves to a defined copula model. In the third chapter, we extend the model from the second chapter by introducing a parametric estimator for Archimedean copula generators. We show how this new approach allows to incorporate covariates and assess their impact on both the strength and the shape of dependence.

INTRODUCTION

Motivation

Cette thèse se situe à la croisée des chemins entre le provisionnement en assurance et la modélisation de la dépendance. Elle se penche sur le calcul des micro-réserves, dans un but de modéliser au plus près le développement de chaque sinistre individuel, tout au long de leur temps de vie dans le portefeuille des assureurs. À cette fin, nous portons dans cette thèse un intérêt particulier à la modélisation de la dépendance au sein de ces portefeuilles, afin de quantifier cette dernière et d'évaluer l'impact qu'elle peut avoir sur le montant de la réserve finale.

Cet intérêt pour les modèles de calcul de réserves individuelles et pour l'inclusion de la dépendance au sein de ceux-ci provient de la complexification et de la diversification accrue depuis plusieurs années des portefeuilles d'assurance, s'accompagnant d'une augmentation du nombre et du niveau d'exigence des régulations du marché de l'assurance à travers le monde. Ces tendances forcent les compagnies à devoir sans cesse repenser et améliorer leurs modèles.

Le monde de l'assurance a en effet bien évolué depuis ses prémices. Des premiers contrats formels d'assurance maritime au Moyen-Âge aux polices d'assurance automobile offrant à elles seules un grande nombre de couvertures variées, l'assurance non-vie a connu, particulièrement depuis l'émergence des technologies et d'internet, une forte croissance et une grande dynamisation. Dans l'ère des mégadonnées, les actuaires assument souvent de nouveaux rôles, tels que scientifiques de données, et font face à de nombreux risques auxquels leurs entreprises sont confrontées. Ces risques, en plus de créer de nouveaux produits ou branches d'assurance, complexifient également la structure des portefeuilles, créant parfois de la dépendance entre risques qui étaient jusqu'alors considérés comme non-corrélés.

Face à cette ramification des profils et portefeuilles des assureurs, les réglementations visant le monde de l'assurance ont fortement évolué. La modélisation et la quantification du risque, sous forme de besoins en capital, est au coeur du métier d'actuaire. Lors de chaque exercice comptable, ceux-ci doivent évaluer leurs réserves : le montant prédit des réclamations. Ces réserves permettront ensuite, en les combinant avec d'autres éléments du bilan, d'estimer les besoins en capital de la compagnie.

La spécificité de l'assurance est que le cycle de production y est inversé : les actuaires doivent estimer

les primes d'assurance, autrement dit le prix associé aux contrats offerts par la compagnie, bien avant de connaître le coût de ces contrats. L'estimation des réserves est donc un élément crucial pour la santé financière et la solvabilité de la compagnie, étroitement surveillées par les régulateurs des différents pays.

Dans l'Union Européenne, le cadre réglementaire Solvabilité II, introduit par l'Autorité Européenne des Assurances et des Pensions Professionnelles (AEAPP) pour les compagnies d'assurance et de réassurance, impose de prendre en compte la dépendance entre les différents modules de risques qui composent le besoin en capital des compagnies, via une matrice de corrélation. Il s'agit de la formule standard, appliquée par une majorité des assureurs européens. Comme discuté par (M. et Stahl, 2021), cette formule présente de nombreux désavantages et a souvent été critiquée pour son aspect trop réducteur qui ne permet pas de refléter les différents profils de risques de façon appropriée. Des alternatives existent : la Directive Solvabilité II permet aux compagnies d'implémenter leurs propres modèles internes, complets ou partiels, pour le calcul de capital. Ces modèles sont cependant soumis à des contrôles rigoureux et à un long et ardu processus d'approbation par les régulateurs. De plus, le coût financier et administratif lié à l'implémentation d'un tel modèle peut être dissuasif pour beaucoup de compagnies qui préfèreront dès lors conserver la formule standard de l'AEAPP.

Au Canada, la dépendance entre les différents modules de risques qui composent le capital requis des compagnies est inclue via le crédit de diversification, imposé par le Bureau du Surintendant des Institutions Financières (BSIF). Dans certains cas, les compagnies optent pour un modèle interne qui reflétera plus adéquatement leur profil de risques. Comme en Europe, le processus d'approbation, de nombreux contrôles et un coût significatif peuvent cependant contribuer à persuader les assureurs d'utiliser le modèle standard.

Au vu de ces nombreuses régulations et des contraintes, souvent lourdes, imposées aux compagnies sur les modèles de calcul de risques et de capital qu'elles utilisent, le monde de l'assurance a réagi assez tardivement à l'arrivée massive de données ainsi qu'aux avancées technologiques et numériques. Malgré les grandes quantités de données désormais à leur disposition, les compagnies d'assurance, ainsi que de réassurance, tendent à conserver des modèles avec lesquels elles ne peuvent pas tirer profit de ces ressources. Entre autres, elles utilisent encore très souvent des modèles de provisionnement qui ne permettent pas de modéliser dans le détail le développement des sinistres, ni de bénéficier de la grande quantité d'information disponible sur ces sinistres ou sur les assurés. Plus encore, malgré la grande diversité de leurs portefeuilles de risques et les nombreuses réglementations auxquelles elles sont soumises, les modèles utilisés par la

plupart des compagnies ne prennent pas ou peu en compte la dépendance entre ces risques. Pendant très longtemps, les compagnies d'assurance, à l'égard de divers autres acteurs des marchés financiers, se sont basées sur des hypothèses d'indépendance entre risques et sur la théorie des grands nombres.

Cette thèse s'inscrit précisément dans cette problématique. Le but est double : premièrement, nous explorons des modèles de provisionnement individuels, permettant une modélisation plus précise du développement des sinistres tout au long de leur temps de vie. L'objectif est de tirer parti des données collectées par les assureurs sur les réclamations et les détenteurs de polices afin d'estimer les réserves individuelles en modélisant le développement complet des sinistres. Ceci contraste avec l'approche usuelle de l'assurance par laquelle seuls les montants finaux des réserves sont estimés, en ne tenant pas ou peu compte des divers événements qui ponctuent la vie d'un sinistre, de son occurrence à sa clôture, et qui peuvent impacter la réserve finale. Deuxièmement, au vu de la diversification des portefeuilles d'assurance et de l'augmentation et de la complexification des réglementations mises en place, nous incluons la dépendance dans nos modèles de réserves individuelles.

Afin d'appliquer et de valider au mieux les différentes méthodologies qui seront étudiées au cours de cette thèse, nous utiliserons une base de données d'assurance automobile d'un assureur canadien. Dans ces données, chaque police d'assurance offre quatre couvertures distinctes aux assurés : une couverture pour indemnités d'accident, pour blessures corporelles, pour dommages au véhicule et pour perte d'usage du véhicule. Des analyses empiriques réalisées sur ces données nous permettent de voir très rapidement que de la dépendance existe entre ces quatre couvertures d'assurance. En effet, il est aisé d'imaginer qu'un accident qui implique des dommages au véhicule de l'assuré puisse aussi souvent impliquer une perte d'usage du véhicule endommagé et la nécessité d'obtenir un véhicule de remplacement. Un tel sinistre implique alors pour l'assureur les couvertures pour dommages au véhicule et pour perte d'usage de ce dernier. Dans le même idée, un accident suffisamment grave pour causer des blessures à l'assuré pourrait également avoir causé des blessures à un tiers. Dans ce cas, les couvertures pour indemnités d'accident et blessures corporelles, voire également dommages au véhicule et perte d'usage de ce dernier pourraient toutes être impliquées.

Cette thèse se structure en trois chapitres. Dans le Chapitre 1, nous explorons un modèle de provisionnement individuel prenant en compte la dépendance entre différentes couvertures d'assurance offertes sous une même police. Nous introduisons le concept de *schémas activation* d'une couverture comme étant le

premier moment auquel l'assureur se rend compte qu'un sinistre donné impacte la couverture en question. Pour capturer la dépendance entre ces couvertures d'assurance, nous utilisons une distribution multivariée au niveau des schémas d'activation : la distribution multinomiale logistique. Dans ce chapitre, nous modélisons le développement d'un sinistre au niveau des différentes lignes d'affaire qu'il impacte. Nous commençons par déterminer quel(les) couverture(s) est (sont) impactée(s). Nous modélisons ensuite la survenance de paiements et leurs sévérités, pour obtenir une estimation de la réserve totale du portefeuille. Ce chapitre s'inscrit dans l'effort poursuivi dans la littérature actuarielle de proposer des modèles de provisionnement qui permettent aux assureurs de prédire non seulement le coût total de leur portefeuille, mais également le développement des sinistres tout au long de leur vie, afin d'avoir une meilleure connaissance et un meilleur contrôle sur leur portefeuille.

Dans le Chapitre 2, nous travaillons en temps continu et définissons le concept de *délai d'activation*, correspondant au temps écoulé entre le moment de déclaration d'un sinistre et le moment auquel l'assureur enregistre pour la première fois que celui-ci a impacté une des couvertures offertes par la police d'assurance. Nous utilisons des copules archimédiennes pour modéliser la dépendance entre les délais d'activation de plusieurs couvertures. Travaillant avec des données censurées, nous présentons un estimateur non-paramétrique de la fonction génératrice des copules appartenant à cette famille, et nous l'utilisons ensuite pour sélectionner la copule la plus appropriée aux données par une approche graphique. Nous présentons un algorithme de simulations pour directement utiliser la fonction génératrice dans des modèles de prédiction. Ceci nous permet de passer outre tout modèle paramétrique et d'obtenir des prédictions plus précises.

Dans le Chapitre 3, nous étendons les techniques utilisées dans le Chapitre 2 pour y inclure des variables explicatives, permettant ainsi de tirer profit des données collectées par les assureurs. Nous proposons une alternative paramétrique à l'estimateur de la fonction génératrice des copules archimédiennes. Ce nouvel estimateur permet d'analyser l'impact qu'une variable explicative peut avoir non seulement sur la force de la dépendance via le tau de Kendall, mais aussi sur sa forme. Nous illustrons notre méthodologie à l'aide de deux applications : l'une sur les données d'une étude sur la rétinopathie diabétique, l'autre sur les données d'un portefeuille d'assurance automobile canadien. Ce chapitre s'inscrit dans une tendance récente de la littérature sur les copules visant à inclure l'effet de variables explicatives dans la modélisation de la dépendance.

Dans le reste de cette introduction, nous présentons une brève revue des modèles de provisionnement et de dépendance, afin de remettre dans leurs contextes académiques les différents outils qui seront utilisés dans cette thèse.

Modèles de provisionnement

Les modèles classiques de provisionnement sont les modèles dits *agrégés*, encore très utilisés aujourd'hui par les compagnies d'assurance. Ils tirent leur nom du fait qu'ils agrègent les montants des réclamations par période de survenance et de développement (traditionnellement par année comptable), et les prédictions sont ensuite réalisées sur cette même base. Contrairement aux modèles individuels, ils ne permettent donc pas d'estimer le montant des réserves par contrat du portefeuille. Le premier de ces modèles agrégés est le modèle *Chain Ladder*. Ce modèle déterministe repose sur l'hypothèse que l'évolution des coûts des sinistres reste constante d'une période à l'autre, c'est-à-dire que les montants des sinistres ne devraient pas être significativement différents au cours d'une période donnée par rapport à ce qui a été observé précédemment. Ce modèle n'a pas de fondement statistique. Plusieurs variantes ont été proposées pour l'améliorer, notamment le modèle *Bornhuetter-Ferguson* par (Bornhuetter et Ferguson, 1972) qui permet d'atténuer l'impact des plus anciennes années de survenance dans la modélisation, ou la méthode *Cape Cod*, qui estime des poids pour les différentes périodes de survenance à partir de l'exposition et des périodes de développement. Il faut cependant attendre le début des années 1990 et le modèle *Chain Ladder* de Mack (Mack, 1993), pour que le calcul de réserves passe pour la première fois du monde déterministe vers le monde stochastique, et que les estimations reposent sur des hypothèses statistiques.

Les bases de ce qui rendra possible par la suite la modélisation individuelle des réserves furent cependant posées dès la fin des années 1980, entre autres par (Norberg, 1986). Grâce, également, à l'introduction par (Nelder et Wedderburn, 1972) des modèles linéaires généralisés (*Generalized Linear Models*, ou GLM), la transition du provisionnement des modèles agrégés, dits *en triangles*, vers des modèles plus élaborés, offrant une prédiction de la réserve au niveau des sinistres individuels, allait bientôt s'opérer. Les modèles linéaires généralisés permettent de modéliser une variable réponse, Y, à partir d'une ou de plusieurs variable(s) explicative(s), $\boldsymbol{X}$. Contrairement aux modèles linéaires classiques, les GLM n'imposent ni à la variable réponse de suivre la distribution normale, ni la linéarité du lien entre Y et les variables explicatives. La variable réponse peut provenir de toute distribution appartenant à la famille exponentielle linéaire, c'est-

à-dire à la famille de distributions dont les fonctions de densité s'expriment sous la forme

$$f_Y(y) = c(y, \phi)\exp\Big(\frac{y\theta - a(\theta)}{\phi}\Big),$$

où θ est le paramètre *canonique* de la distribution, ϕ le paramètre de *dispersion*, $c(.)$ est une fonction appelée *mesure de base* et $a(.)$ est la fonction de *log-partition*. Cette famille regroupe de nombreuses distributions discrètes et continues fréquemment utilisées en actuariat, notamment les distributions binomiale, Poisson, normale, exponentielle ou encore gamma.

Sous un modèle linéaire généralisé, pour un échantillon d'observations de taille n, l'espérance de la variable réponse Y est alors modélisée par

$$\mathrm{E}[Y_i] = g^{-1}(\boldsymbol{X}_i^{'}\boldsymbol{\beta}),$$

pour chaque observation $i = 1, ..., n$. Le vecteur $\boldsymbol{X}_i = (X_{i,1}, ..., X_{i,p})$ représente les p variables explicatives associées à l'observation i, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$ sont les coefficients du modèle, et $g(.)$ est la fonction de lien qui définit la relation entre l'espérance de la variable réponse et les variables explicatives.

Les GLM sont aujourd'hui très utilisés dans la modélisation des réserves, tant dans l'industrie que dans la littérature actuarielle. Ils souffrent cependant du désavantage de ne pas correctement capturer les effets d'une variable explicative continue. Puisque le lien entre $g(\mathrm{E}[Y_i])$ et $\boldsymbol{X}_i^{'}\boldsymbol{\beta}$ est linéaire, ce type de modèle est moins approprié lorsqu'il s'agit d'y inclure une variable continue qui pourrait impacter la variable réponse de façon non-linéaire. (Hastie et Tibshirani, 1990) proposent une extension des GLM pour palier à ce désavantage : les modèles additifs généralisés (*Generalized Additive Models*, ou GAM). Ces modèles permettent d'inclure des variables continues transformées par des fonctions de lissage dans le prédicteur, et s'expriment comme suit :

$$\mathrm{E}[Y_i] = g^{-1}(\beta_0 + f_1(X_1) + f_2(X_2) + ... + f_p(X_p)),$$

où $f(.)$ sont les fonctions de lissage. La flexibilité accrue des GAM s'accompagne cependant d'une perte d'interprétabilité des modèles, et les assureurs optent souvent pour des GLM dans lesquels ils ne conservent que des variables catégorielles, ou dans lesquels ils introduisent des variables continues préalablement regroupées en diverses catégories. Certains auteurs, comme (Henckaerts *et al.*, 2018), proposent des solutions visant notamment à développer des stratégies de regroupement de données continues en différentes classes pouvant être utilisées dans les GLM.

Enfin, (Stasinopoulos *et al.*, 2017) proposent plus récemment une nouvelle généralisation des GLM et des GAM au travers des modèles additifs généralisés pour l'emplacement, l'échelle et la forme (*Generalized Additive Models for the Location, Scale and Shape*, ou GAMLSS). Ceux-ci permettent de considérer des distributions en-dehors de la famille exponentielle linéaire et de modéliser non seulement l'espérance de la variable réponse mais également sa variance et ses paramètres de forme tels que les coefficients d'aplatissement et d'asymétrie, avec ou sans fonctions de lissage pour les variables continues. Ces modèles présentent donc un degré de flexibilité accru et très utile en modélisation actuarielle où les variables réponses, que ce soit la fréquence ou la sévérité des sinistres, présentent souvent des caractéristiques que les GLM ou GAM seuls ne sont pas complètement en mesure de capturer. Dans cette thèse, nous utiliserons principalement les GAMLSS pour construire nos modèles de provisionnement, tirant ainsi parti de leur haut niveau de flexibilité pour modéliser, entre autres, les sévérités d'un portefeuille canadien complexe d'assurance automobile.

Malgré le développement de ces différents modèles dès la fin des années 1990, il faudra attendre le début des années 2010 et une plus grande disponibilité des ressources informatiques et computationnelles, pour que ceux-ci prennent vraiment leur essor. Le modèle proposé par (Antonio et Plat, 2014) est aujourd'hui considéré comme le premier modèle complet pour le calcul de réserves individuelles. Ces modèles fleurissent depuis dans la littérature actuarielle, avec des variantes très différentes, du paramétrique au non-paramétrique, en temps discret comme continu. L'avènement de l'apprentissage par machine a également apporté son lot de nouvelles techniques pour le provisionnement en assurance, utilisant des arbres de régression (Wüthrich, 2018), des algorithmes de boosting (Duval et Pigeon, 2019) ou même des réseaux de neurones (Delong *et al.*, 2021). Le premier objectif de cette thèse s'inscrit dans la continuité de cet effort de développer des modèles de provisionnement flexibles et précis, qui tirent profit des données disponibles aux assureurs et permettent de prédire l'évolution des sinistres tout au long de leur développement.

Modélisation de la dépendance

Le deuxième objectif de cette thèse est, au vu de la diversification des portefeuilles d'assurance ainsi que des réglementations mises en place à travers le monde, d'inclure la dépendance entre risques dans les modèles de provisionnement. La modélisation de la dépendance est un élément clef de la quantification des risques dans un portefeuille, et plusieurs auteurs dans la littérature actuarielle, mais aussi statistique, se sont penchés sur ce sujet. La dépendance dans un portefeuille d'assurance peut prendre de nombreuses formes. Lors de la modélisation du développement d'une réclamation, il peut s'agir de dépendance tempo-

relle, par exemple entre le délai de développement et le montant final payé pour le sinistre, comme dans (Lopez, 2019). Il peut également être question de la dépendance entre le moment de survenance d'un certain événement dans la vie du sinistre et son délai de développement, tel qu'illustré par (Zhou et Zhao, 2010). La dépendance s'immisce également entre risques concurrents d'un portefeuille. La diversification des produits d'assurance non-vie mène les actuaires à jongler avec de plus en plus de lignes d'affaires dont ils doivent souvent combiner les réserves en un seul montant à rapporter au régulateur. Au-delà du simple calcul de réserves, et comme déjà mentionné précédemment, les assureurs doivent également inclure dans leurs calculs de besoins en capital la dépendance entre leurs différents modules de risques.

Le type de dépendance sur lequel se penche cette thèse s'inscrit dans cette lignée, avec une attention particulière à la dépendance entre différentes couvertures d'assurance offertes sous une même police. Ce type de problématique a déjà été abordé pour des contextes de tarification dans la littérature actuarielle, entre autres par (Frees et Valdez, 2008) qui proposent un modèle prenant en compte la dépendance entre différents types de réclamations pour le calcul de primes.

Depuis la découverte du concept de corrélation, attribuée à Francis Galton en 1888, les méthodes de quantification et modélisation de la dépendance ont lRgement évolué. Pour mesurer le niveau de corrélation entre deux variables, les mesures d'association constituent encore aujourd'hui la méthode la plus simple et directe. Le coefficient de corrélation de Pearson permet, par exemple, de capturer le degré de dépendance linéaire entre des variables aléatoires. D'autres mesures, basées sur les rangs des observations plutôt que sur les observations des variables elles-mêmes, permettent de mesurer la corrélation entre celles-ci sans que des transformations non-linéaires des variables ne viennent perturber les résultats. Il s'agit notamment du rho de Spearman (Spearman, 1904) et du tau de Kendall, introduit par (Kendall, 1938), que nous utiliserons à travers cette thèse. Ce dernier se base sur le concept de concordance.

Definition 0.1 (Concordance) *Les paires d'observations* (X_1, Y_1) *et* (X_2, Y_2) *sont dites concordantes (resp. discordantes) si, lorsque* $X_1 > X_2$*, alors* $Y_1 > Y_2$ *(resp.* $Y_1 < Y_2$*).*

La probabilité de concordance (resp. discordance) est donc $\mathsf{P}[(X_1 - X_2)(Y_1 - Y_2) > 0]$ (resp. $\mathsf{P}[(X_1 - X_2)(Y_1 - Y_2) < 0]$, et le tau de Kendall est défini comme

$$\tau = \mathsf{P}[(X_1 - X_2)(Y_1 - Y_2) > 0] - \mathsf{P}[(X_1 - X_2)(Y_1 - Y_2) < 0].$$

Les mesures d'association, notamment la corrélation de Pearson, permettent de construire des distributions multivariées. Celles-ci consistent en une combinaison linéaire pondérées des variables corrélées, dont la dépendance est capturée par une matrice de corrélation qui doit être définie positive. Pour certaines distributions très utilisées en actuariat, telles que la normale, la distribution multivariée portera le même nom que les distributions marginales, avec pour seule différence la matrice de corrélation. Par exemple, soit $\boldsymbol{X} = (X_1, X_2)$ un vecteur de distribution normale bivariée. Les distributions marginales des composantes X_1 et X_2 sont également des distributions normales, telles que $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ et $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Les paramètres de la distribution bivariée sont alors donnés par

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \text{et} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \text{cov}(X_1, X_2) \\ \text{cov}(X_1, X_2) & 1 \end{pmatrix}.$$

Pour d'autres types de distributions dont les marginales n'appartiendraient pas à la même famille que celle de la distribution multivariée, il peut être moins souhaitable de travailler avec cette dernière. Bien qu'elles s'avèrent utiles dans certains cas, il peut devenir difficile de construire des structures de dépendance plus complexes avec des distributions multivariées. Le nombre de paramètres à estimer avec ces méthodes peut aussi devenir large assez rapidement, rendant l'estimation et l'interprétation de modèles plus ardues.

Une méthode alternative pour modéliser la dépendance s'est cependant développée depuis le milieu du XXe siècle et a fortement gagné en popularité dans de nombreux domaines depuis ; il s'agit des copules. Une copule est une fonction qui permet de combiner plusieurs distributions univariées en une seule distribution jointe. Abe Sklar a posé les bases théoriques de l'application des copules, notamment par le théorème éponyme (Sklar, 1959) :

Theorem 0.2 (Théorème de Sklar ((Nelsen, 2006))) *Soit $H(.)$ une fonction de distribution bivariée avec $F(.)$ et $G(.)$ comme distributions marginales. Il existe alors une copule $C(.)$ telle que pour toutes valeurs x et y dans $\mathbb{R}$,*

$$H(x, y) = C(F(x), G(y)). \tag{1}$$

Si F et G sont continues, alors la copule C est unique. Sinon, C est définie de façon unique sur Ran$(F) \times$ Ran(G). De façon équivalente, si $C(.)$ est une copule et si F et G sont des fonctions de répartition, alors la fonction $H(.)$ définie par l'Équation (2) est une fonction de répartition jointe avec F et G comme distributions marginales.

Le grand avantage des copules est qu'elles peuvent s'appliquer peu importe les distributions marginales des variables aléatoires. Elles permettent d'établir une relation de dépendance qui est distincte des distributions marginales. De plus, elles sont un outil très flexible qui permet de construire de nombreuses structures de dépendance, même les plus complexes. Les copules Gaussiennes sont par exemple utilisées pour modéliser des structures de dépendance symétriques.

Plusieurs familles de copules ont été décrites dans la littérature. Une des plus populaires et utilisées en assurance mais aussi dans de nombreux autres domaines tels que la finance, la médecine et même la météorologie, est la famille des copules archimédiennes amplement décrite par (Nelsen, 2006). Les copules membres de cette famille s'expriment à partir d'une fonction génératrice, dénotée $\psi(.) : [0, \infty] \rightarrow [0, 1]$, avec $\psi(0) = 1$ et $\lim_{\nu \rightarrow \infty} \psi(\nu) = 0$:

$$C(F_1(x_1), ..., F_d(x_d)) = \psi\Big\{\psi^{-1}[F_1(x_1)] + ... + \psi^{-1}[F_d(x_d)]\Big\}.$$

La popularité de cette famille découle principalement de leur simplicité et de leur flexibilité. Grâce à leur expression via la fonction génératrice, elles sont en effet faciles à construire et utiliser, et possèdent moult propriétés intéressantes. De plus, la dépendance y est établie par peu de paramètres par rapport à d'autres familles, certaines n'en possédant d'ailleurs qu'un seul. Il existe de très nombreuses copules archimédiennes, permettant de construire des structures de dépendance différentes. Dans le cadre de cette thèse, nous explorons quatre copules en particulier : les copules de Clayton (Clayton, 1978), Frank (Frank, 1979), Gumbel (Gumbel, 1960) et Joe (Joe, 1993). Ces quatre copules sont très populaires dans la litérature actuarielle mais aussi financière. Elles font parties des copules archimédiennes ne dépendant que d'un seul paramètre qui gouverne la force de dépendance. Elles permettent toutes également de modéliser des structures de dépendance très variées, comme illustré à la Figure 0.2. Ces graphiques représentent les nuages de points d'observations bivariées simulées à partir des quatre copules pour un tau de Kendall égal à 0.8. On y observe que la copule de Clayton permet de modéliser la dépendance des petites valeurs de la distribution, la copule de Frank modélise une dépendance plus uniforme sur l'ensemble des observations, la copule de Gumbel est utile pour établir la dépendance des valeurs extrêmes (entre les petites et les grandes valeurs) et la copule de Joe représente mieux la dépendance dans les grandes valeurs dans la queue de la distribution.

Ces quatre copules sont donc des choix appropriés pour modéliser la dépendance dans un jeu de données. Elles permettent, entre elles, d'analyser quatre formes de dépendance distinctes. Elles sont aussi très faciles à implémenter car elles ne dépendent que d'un seul paramètre de dépendance et des expressions simples

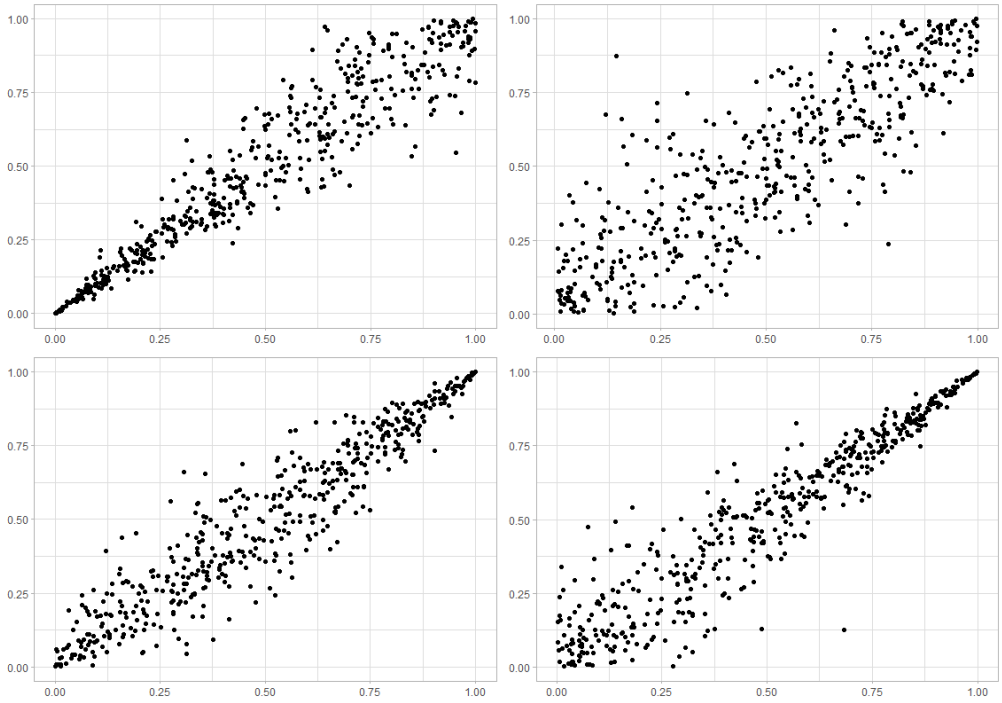

Figure 0.1 – Nuages de points des copules de Clayton (en haut à gauche), Frank (en haut à droite), Gumbel (en bas à gauche) et Joe (en bas à droite) pour un tau de Kendall égal à 0.8.

existent pour leurs fonctions génératrices, tel qu'illustré dans l'Annexe B.1.

Motivation

This thesis lies at the intersection of loss reserving and dependence modeling. It focuses on the calculation of micro-reserves with the goal to model the development of each individual claim as closely as possible throughout their lifetime in the insurers' portfolios. To this end, particular attention is paid to modeling dependence within these portfolios to quantify it and assess its impact on the final reserve amount.

The interest in micro-level loss reserving and in the inclusion of dependence therein stems from the increasing complexity and diversification of insurance portfolios over the past years, coupled with rising regulatory requirements in insurance markets worldwide. These trends compel companies to continually rethink and improve their models.

The insurance world has indeed evolved significantly over time. From the early formal maritime insurance contracts in the Middle Ages to automobile insurance policies offering a wide range of coverages, the non-life insurance sector has experienced substantial growth and dynamism, particularly with the emergence of technology and the internet. In the era of big data, actuaries often assume new roles, such as data scientists, and face numerous risks that their companies encounter. These risks not only lead to the creation of new insurance products or lines of business but also complicate the structure of their portfolios, sometimes creating dependence between risks that were previously considered uncorrelated.

Faced with the diversification of insurers' profiles and portfolios, regulations governing the insurance world have evolved significantly. Risk modeling and quantification, in the form of capital requirements, lie at the heart of actuarial work. Actuaries must estimate their reserves, the predicted claim amounts, in each reporting period. These reserves, combined with other items on the balance sheet, will then be used to estimate the company's capital requirements.

One peculiarity of insurance is that the production cycle is reversed : actuaries must estimate insurance premiums, i.e., the price associated with the contracts offered by the company, long before knowing the cost of these contracts. Reserve estimation is thus a crucial element for the financial health and solvency of companies, closely monitored by the regulators of each country.

In the European Union, the Solvency II regulatory framework, introduced by the European Insurance and Occupational Pensions Authority (EIOPA) for insurance and reinsurance companies, requires consideration of dependence between the different risk modules composing the capital requirement of companies, via a correlation matrix. This is the standard formula applied by a majority of European insurers. As discussed by (M. et Stahl, 2021), this formula has many disadvantages and has often been criticized for its overly simplistic nature, which fails to adequately reflect the different risk profiles. Alternatives exist : the Solvency II Directive allows companies to implement their own internal models, either full or partial, for capital calculation. However, these models are subject to rigorous controls and a lengthy and challenging approval process by regulators. Moreover, the financial and administrative costs associated with implementing such models can be prohibitive for many companies, leading them to stick to the EIOPA standard formula.

In Canada, dependence between the different risk modules composing the required capital of insurance companies is included via the diversification credit, imposed by the Office of the Superintendent of Finan-

cial Institutions (OSFI). In some cases, companies opt for an internal model that better reflects their risk profile. As in Europe, the approval process, numerous controls, and significant costs can, however, discourage insurers from using a more personalized model.

Given these numerous regulations and heavy constraints imposed on companies regarding the risk and capital calculation models they use, the insurance world has responded relatively late to the massive influx of data as well as technological and digital advancements. Despite the large amounts of data now available to them, insurance and reinsurance companies often still use models that do not allow them to leverage these resources. Among other things, they still frequently use reserving models that do not allow for detailed modeling of claims development or benefit from the wealth of information available on these claims or policyholders. Furthermore, despite the great diversity of their risk portfolios and the numerous regulations to which they are subject, the models used by most companies do not take into account or barely take into account the dependence between these risks. For a long time, insurance companies, like various other players in financial markets, have relied on independence assumptions between risks and on the theory of large numbers.

In this thesis, we address this issue. The goal is twofold : first, we explore individual reserving models, allowing for a more precise modeling of claim development throughout their lifetimes in insurers' portfolios. We seek to leverage the data collected by insurers on claims and policyholders to estimate individual reserves by modeling the complete development of claims. This contrasts with the usual approach in insurance where only final reserve amounts are estimated, with little or no consideration of the various events that punctuate the life of a claim, from occurrence to settlement, and that can impact the final reserve. Second, given the diversification of insurance portfolios and the increase and complexity of regulations, we seek to include dependence in our micro-level reserving models.

In order to best apply and validate the various methodologies that will be investigated in this thesis, we use a Canadian automobile insurance dataset. In these data, each insurance policy offers four distinct coverages to policyholders : Accident Benefits, Bodily Injury, Vehicle Damage and Loss of Use. Empirical analyses carried out on these data quickly show that there is a dependency between these four insurance coverages. Indeed, it's easy to imagine that an accident involving damage to the insured's vehicle can also often involve loss of use of the damaged vehicle and the need to obtain a replacement vehicle. For the insurer, a claim of this kind implies both the Vehicle Damage and Loss of Use coverages. Similarly, an accident serious enough to

cause injury to the insured may also have caused injury to a third party. In this case, the Accident Benefits and Bodily Injury coverages, and even the Vehicle Damage and Loss of Use coverages, could all be involved.

This thesis is structured into three chapters. In Chapter 1, we explore an individual reserving model taking into account the dependence between different insurance coverages offered under the same policy. We introduce the concept of *activation patterns* of the coverages as the first moment when the insurer realizes that a given claim impacts the coverage in question. To capture the dependence between these insurance coverages, we use a multivariate distribution at the activation pattern level : the logit multinomial distribution. In this chapter, we model the development of a claim at the level of the different lines of business it impacts. We begin by determining which coverage(s) is (are) activated. We then model the occurrence of payments and their severities to obtain an estimate of the total portfolio reserve. This chapter contributes to the ongoing effort in actuarial literature to propose reserving models that allow insurers to predict not only the total cost of their portfolio but also the development of claims throughout their lifetime, to have a better understanding and control over these portfolio.

In Chapter 2, we work in continuous time and define the concept of *activation delay*, corresponding to the time elapsed between the declaration of a claim and the moment at which the insurer first records that it has impacted one of the coverages offered within the insurance policy. We use Archimedean copulas to model dependence between the activation delays of multiple coverages. Working with censored data, we present a non-parametric estimator of the generator function of copulas belonging to this family. We then use it to select the most appropriate copula for the data using a graphical approach. We present a simulation algorithm to directly use the generator function in prediction models. This allows us to bypass any parametric model and obtain more accurate predictions.

In Chapter 3, we extend the techniques used in Chapter2 to include explanatory variables, thus leveraging the data collected by insurers. We propose a parametric alternative to the estimator of the generator function of Archimedean copulas. This new estimator allows us to analyze the impact that a covariate can have not only on the strength of dependence via Kendall's tau but also on its shape. We illustrate our methodology using two applications : one on data from a diabetic retinopathy study, the other on data from a Canadian automobile insurance portfolio. This chapter is part of a recent trend in copula literature aiming to include the effect of explanatory variables in dependence modeling.

In the rest of this introduction, we provide a brief review of reserving and dependence models to place in their academic contexts the various tools that will be used in this thesis.

Loss reserving

Classical reserving models are the so-called *aggregated* models, still widely used today by insurance companies. They derive their name from the fact that they aggregate claim amounts by occurrence and development period (traditionally by accounting year), and predictions are then made on this basis. Unlike individual models, they do not allow to estimate reserve amounts per claim in the portfolio. The first of these aggregated models is the *Chain Ladder* model. This deterministic model is based on the assumption that the evolution of claim costs remains constant from one period to another, i.e. claims amounts are not expected to be significantly different in any given period compared to what has been previously observed. This model has no statistical foundation. Several variants have been proposed to improve it, including the *Bornhuetter-Ferguson* model by (Bornhuetter et Ferguson, 1972), which mitigates the impact of older occurrence years in modeling, or the *Cape Cod* method, which estimates weights for different occurrence periods based on exposure and development periods. However, it was not until the early 1990s and Mack's *Chain Ladder* model (Mack, 1993) that the calculation of reserves shifted from the deterministic world to the stochastic world for the first time, and estimates relied on statistical assumptions.

The foundations of what would later enable the individual modeling of reserves were laid as early as the late 1980s, among others by (Norberg, 1986). Thanks also to the introduction by (Nelder et Wedderburn, 1972) of Generalized Linear Models (GLMs), the transition from aggregate, so-called *triangle* reserving models to more elaborate models, offering prediction of reserves at the level of individual claims, was soon to occur. Generalized Linear Models allow for modeling a response variable, Y, based on one or more explanatory variables, $\boldsymbol{X}$. Unlike classical linear models, GLMs do not impose the response variable to follow a normal distribution, nor the linearity of the relationship between Y and the explanatory variables. The response variable can come from any distribution belonging to the exponential family, i.e., the family of distributions whose density functions are expressed in the form

$$f_Y(y) = c(y, \phi)\exp\Big(\frac{y\theta - a(\theta)}{\phi}\Big),$$

where θ is the *canonical* parameter of the distribution, ϕ is the *dispersion* parameter, $c(.)$ is a function called the *base measure*, and $a(.)$ is the *log-partition* function. This family includes many discrete and continuous distributions frequently used in actuarial science, including binomial, Poisson, normal, exponential, or gamma distributions. Under a Generalized Linear Model, for a sample of observations of size n, the ex-

pectation of the response variable Y is then modeled by

$$\mathrm{E}[Y_i] = g^{-1}(\boldsymbol{X}_i'\boldsymbol{\beta}),$$

for each observation $i = 1, ..., n$. The vector $\boldsymbol{X}_i = (X_{i,1}, ..., X_{i,p})$ represents the p explanatory variables associated with observation i, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$ are the model coefficients, and $g(.)$ is the link function that defines the relationship between the expectation of the response variable and the explanatory variables.

GLMs are widely used in reserving models, both in industry and in the actuarial literature. However, they suffer from the disadvantage of not adequately capturing the effects of a continuous explanatory variable. Since the link between $g(\mathrm{E}[Y_i])$ and $\boldsymbol{X}_i'\boldsymbol{\beta}$ is linear, this type of model is less suitable when including a continuous variable that could impact the response variable nonlinearly. (Hastie et Tibshirani, 1990) proposed an extension of GLMs to overcome this disadvantage : Generalized Additive Models (GAMs). These models allow for including continuous variables transformed by smoothing functions in the predictor, and are expressed as follows :

$$\mathrm{E}[Y_i] = g^{-1}(\beta_0 + f_1(X_1) + f_2(X_2) + ... + f_p(X_p)),$$

where $f(.)$ are smoothing functions. The increased flexibility of GAMs, however, comes with a loss of model interpretability, and insurers often opt for GLMs in which they retain only categorical variables or in which they introduce continuous variables pre-grouped into various categories. Some authors, such as (Henckaerts *et al.*, 2018), propose solutions aimed at developing data grouping strategies for continuous data into different classes that can be used in GLMs. Finally, (Stasinopoulos *et al.*, 2017) recently proposed a new generalization of GLMs and GAMs through Generalized Additive Models for Location, Scale, and Shape (GAMLSS). These models allow to consider distributions outside the exponential family and to model not only the expectation of the response variable but also its variance and shape parameters such as kurtosis and skewness coefficients, with or without smoothing functions for continuous variables. These models therefore offer a higher level of flexibility, which is very useful in actuarial modeling where response variables, whether it be claim frequency or severity, often exhibit characteristics that GLMs or GAMs alone are not completely able to capture. In this thesis, we will primarily use GAMLSS to build our reserving models, thus taking advantage of their high level of flexibility to estimate, among others, the severities of a complex Canadian automobile insurance portfolio.

Despite the development of these different models since the late 1990s, it was not until the early 2010s,

when computational resources and computing power became more readily available, that they truly gained momentum. The model proposed by (Antonio et Plat, 2014) is now considered the first comprehensive model for individual reserving. These models have since proliferated in the actuarial literature, with very different variants, from parametric to non-parametric, in both discrete and continuous time. The advent of machine learning has also brought new techniques for insurance reserving, using regression trees (Wüthrich, 2018), boosting algorithms (Duval et Pigeon, 2019), or even neural networks (Delong *et al.*, 2021). The first objective of this thesis is in line with this ongoing effort to develop flexible and accurate reserving models that leverage the data available to insurers and allow for predicting claims development throughout their lifetime.

Dependence modeling

The second objective of this thesis is, given the diversification of insurance portfolios and the market regulations worldwide, to include inter-risks dependence in reserving models. Modeling dependence is a key element in risk quantification in a portfolio, and several authors in the actuarial literature, as well as statistics, have explored this topic. Dependence in an insurance portfolio can take many forms. When modeling the development of a claim, it can involve temporal dependence, for example, between the development lag and the final amount paid for the claim, as in (Lopez, 2019). It can also be a matter of dependence between the timing of a certain event in the life of the claim and its development lag, as illustrated by (Zhou et Zhao, 2010). Dependence also arises between concurrent risks in a portfolio. The diversification of non-life insurance products leads actuaries to juggle with more and more lines of business, often needing to combine reserves into a single amount to report to regulators. Beyond simply calculating reserves, and as mentioned previously, insurers must also include in their capital requirements calculations the dependence between their different risk modules.

The type of dependence addressed in this thesis falls into this line, with particular attention given to dependence between different insurance coverages offered under a single policy. This type of issue has already been addressed for pricing contexts in the actuarial literature, among others by (Frees et Valdez, 2008), who propose a model that takes into account the dependence between different types of claims for premium calculation.

Since the discovery of the concept of correlation, attributed to Francis Galton in 1888, methods for quantifying and modeling dependence have greatly evolved. To measure the level of correlation between two

variables, association measures are still the simplest and most direct method today. The Pearson correlation coefficient, for example, captures the degree of linear dependence between random variables. Other measures, based on the ranks of observations rather than on the observations of the variables themselves, allow to measure the correlation between them, without nonlinear transformations of the variables affecting the results. These include Spearman's rho (Spearman, 1904) and Kendall's tau, introduced by (Kendall, 1938), which we will use throughout this thesis. The latter is based on the concept of concordance.

Definition 0.3 (Concordance) *The pairs of observations* (X_1, Y_1) *and* (X_2, Y_2) *are said to be concordant (resp. discordant) if, when* $X_1 > X_2$*, then* $Y_1 > Y_2$ *(resp.* $Y_1 < Y_2$*).*

The probability of concordance (resp. discordance) is therefore $\mathsf{P}[(X_1 - X_2)(Y_1 - Y_2) > 0]$ (resp. $\mathsf{P}[(X_1 - X_2)(Y_1 - Y_2) < 0]$), and Kendall's tau is defined as

$$\tau = \mathsf{P}[(X_1 - X_2)(Y_1 - Y_2) > 0] - \mathsf{P}[(X_1 - X_2)(Y_1 - Y_2) < 0].$$

Association measures, especially Pearson's correlation, allow to build multivariate distributions. These consist of a weighted linear combination of correlated random variables, with dependence captured by a correlation matrix that must be positive definite. For some distributions widely used in actuarial science, such as the normal distribution, the multivariate distribution will bear the same name as the marginal distributions, with the only difference being the correlation matrix. For example, let $\boldsymbol{X} = (X_1, X_2)$ be a bivariate normal distribution vector. The marginal distributions of components X_1 and X_2 are also normal distributions, such as $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$. The parameters of the bivariate distribution are then given by

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \mathsf{cov}(X_1, X_2) \\ \mathsf{cov}(X_1, X_2) & 1 \end{pmatrix}.$$

For other types of distributions whose marginals do not belong to the same family as the multivariate distribution, it may be less desirable to work with the latter. Although they are useful in some cases, it can become difficult to construct more complex dependency structures with multivariate distributions. The number of parameters to estimate with these methods can also quickly become large, making the estimation and interpretation of models more difficult.

An alternative method for modeling dependence, however, has developed since the mid-20th century and has gained popularity in many fields since then; these are copulas. A copula is a function that combines

several univariate distributions into a single joint distribution. Abe Sklar laid the theoretical foundations for the application of copulas, notably through the eponymous theorem (Sklar, 1959) :

Theorem 0.4 (Sklar's Theorem ((Nelsen, 2006))) *Let $H(.)$ be a bivariate distribution function with $F(.)$ and $G(.)$ as marginal distributions. Then there exists a copula $C(.)$ such that for all values x and y in $\mathbb{R}$,*

$$H(x, y) = C(F(x), G(y)). \tag{2}$$

If F and G are continuous, then the copula C is unique. Otherwise, C is uniquely defined on Ran$(F) \times$ Ran(G). Equivalently, if $C(.)$ is a copula and if F and G are cumulative distribution functions, then the function $H(.)$ defined by Equation (2) is a joint distribution function with F and G as marginal distributions.

The great advantage of copulas is that they can be applied regardless of the marginal distributions of the random variables. They establish a dependence relationship that is distinct from the marginal distributions. Moreover, they are a very flexible tool that allows to capture many dependency structures, even the most complex ones. Gaussian copulas, for example, are used to model symmetric dependence structures.

Several families of copulas have been described in the literature. One of the most popular and used in insurance but also in many other fields such as finance, medicine, and even meteorology, is the family of Archimedean copulas extensively described by (Nelsen, 2006). Copulas belonging to this family are expressed using a generator function, denoted $\psi(.) : [0, \infty] \to [0, 1]$, with $\psi(0) = 1$ and $\lim_{\nu \to \infty} \psi(\nu) = 0$:

$$C(F_1(x_1), ..., F_d(x_d)) = \psi\Big\{\psi^{-1}[F_1(x_1)] + ... + \psi^{-1}[F_d(x_d)]\Big\}.$$

The popularity of this family stems mainly from its simplicity and flexibility. Thanks to their expression via the generator function, they are easy to build and use, and possess many interesting properties. Moreover, dependence is capture by few parameters compared to other families, some of them having only one. There are many Archimedean copulas, allowing for the construction of very different dependence structures. In this thesis, we explore four copulas in particular : the Clayton (Clayton, 1978), Frank (Frank, 1979), Gumbel (Gumbel, 1960), and Joe (Joe, 1993) copulas. These four copulas are very popular in actuarial literature but also in finance. They are part of the Archimedean copulas depending on a single parameter that governs the strength of dependence. Between the four of them, they allow to capture very different dependence structures, as illustrated in Figure 0.2. These graphs show scatter plots of bivariate observations simulated from the four copulas for a Kendall's tau equal to 0.8. We observe that the Clayton copula models the

dependence of small values in the distribution, the Frank copula models a more uniform dependence over all observations, the Gumbel copula is useful for establishing the dependence of extreme values (between small and large values), and the Joe copula better represents the dependence in the large values in the tail of the distribution.

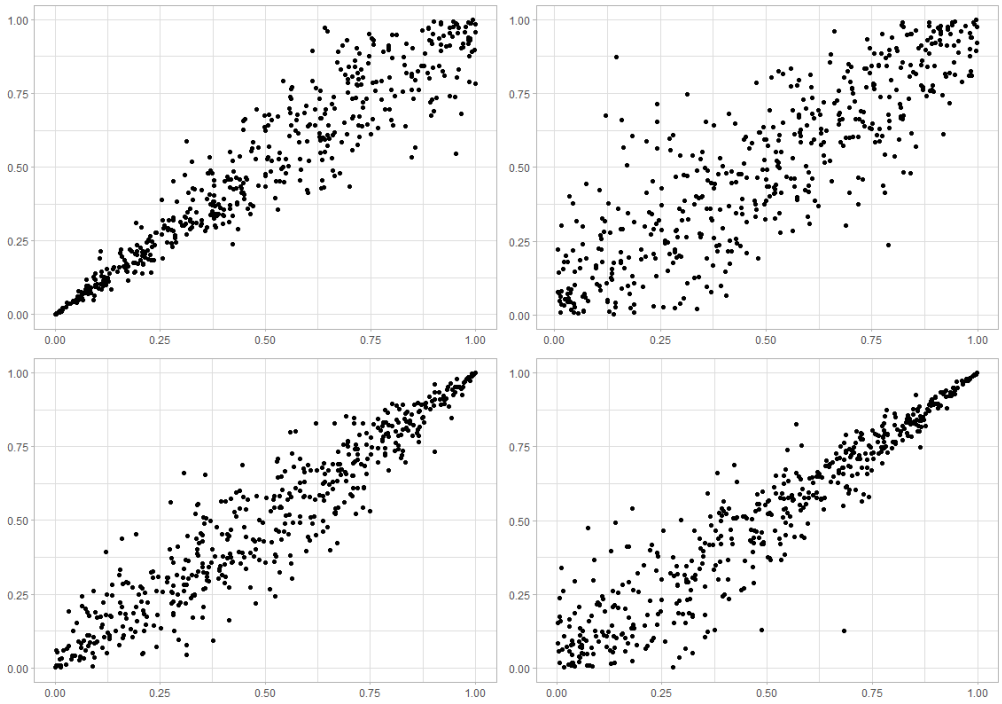

Figure 0.2 – Scatter plots of Clayton (top left), Frank (top right), Gumbel (bottom left), and Joe (bottom right) copulas for Kendall's tau equal to 0.8.

These four copulas are therefore appropriate choices for modeling dependence in a dataset. They allow for analyzing four distinct forms of dependence among them. They are also very easy to implement as they depend only on a single dependence parameter, and simple expressions exist for their generator functions, as illustrated in Appendix B.1.

CHAPITRE 1

INDIVIDUAL CLAIMS RESERVING USING ACTIVATION PATTERNS

Résumé

Une réclamation impacte souvent non pas une, mais plusieurs couvertures d'assurance fournies au sein d'un même contrat. Pour tenir compte de cet aspect multivarié des réclamations, nous proposons un nouveau modèle de calcul de micro-réserves construit autour de l'activation des différentes couvertures d'assurance. En utilisant le cadre de la régression logistique multinomiale, nous modélisons l'activation des différentes couvertures d'assurance pour chaque réclamation et leur évolution au cours des périodes suivantes, c'est-à-dire l'activation d'autres couvertures au fil des périodes ultérieures et tous les paiements possibles qui pourraient en découler. Ainsi, le modèle nous permet de compléter le développement individuel des réclamations ouvertes dans le portefeuille. En utilisant une base de données récente d'assurance automobile provenant d'une grande compagnie d'assurance canadienne, nous démontrons que cette approche génère des prédictions précises pour les montants des réserves totales, ainsi que pour les montants par couverture d'assurance. Cette analyse permet à l'assureur d'obtenir de nouvelles perspectives sur la dynamique de ses réserves.

Abstract

A claim often impacts not one but multiple insurance coverages provided in the contract. To account for this multivariate feature, we propose a new individual claim reserving model built around the activation of the different coverages to predict the reserve amounts. Using the framework of multinomial logistic regression, we model the activation of the different insurance coverages for each claim and their development in the following periods, i.e., the activation of other coverages in the later periods and all the possible payments that might result from them. As such, the model allows us to complete the individual development of the open claims in the portfolio. Using a recent automobile dataset from a major Canadian insurance company, we demonstrate that this approach generates accurate predictions of the total reserves and the reserves per insurance coverage. This analysis allows the insurer to get new insights into the dynamics of his claims reserves.

1.1 Introduction

Claims reserving is known to be one of the most crucial tasks performed by actuaries in insurance companies all over the world. Insurers must accurately predict future liabilities related to open claims. It allows them to answer the reporting standards they are subject to and to preserve sufficient capital with which they can set aside adequate reserves to fulfil their obligations to the policyholders and avoid financial ruin.

In actuarial practice, claims data is commonly aggregated on an occurrence year and development year basis in run-off triangles. These then help actuaries to evaluate the reserves for the portfolio as a whole. One ubiquitous method that uses such triangles is the Chain Ladder model introduced by (Mack, 1993) and further discussed in (Mack, 1994), (Mack, 1999) or (Mack et Venter, 2000).

Over the years, several authors have challenged the robustness of this model. In particular, the recent increase in the quantity and availability of data has contributed to questioning the use of such aggregate methods. The early work of (Buhlmann *et al.*, 1980), (Hachemeister, 1980), and (Norberg, 1986) attempted to benefit from these larger quantities of data. However, a few more years were required to obtain the necessary computing resources to move from the classical run-off triangles to the so-called micro-level claims reserving models. (Antonio et Plat, 2014) were the first to truly incorporate information related to the policyholder or even the claim itself into their model by building on the work mentioned above as well as on prior work performed by (Norberg, 1993), (Norberg, 1999) and (Haastrup et Arjas, 1996). They demonstrate that using the detailed information available to the insurer at the claim level allows them to obtain more accurate predictions for the reserves. Following their lead, (Pigeon *et al.*, 2013) also propose a fully parametric discrete-time model for the payments and then extend their work in (Pigeon *et al.*, 2014) to include incurred losses as well. Many authors have since then contributed to this area of research which is still very active today, as can be seen, for example, from the recent contribution of (Crèvecoeur *et al.*, 2022) who focus on the treatment of RBNS claims.

Other authors have also opened the way to non-parametric approaches to claims reserving. (Wüthrich, 2018) was the first to introduce the use of (Breiman *et al.*, 1984)'s Classification and Regression Tree (CART) algorithm in a micro-level reserving model. Building on this work, (Lopez, 2019) and (Lopez *et al.*, 2016) apply the CART algorithm with censored data using, respectively, survival analysis and copulas to account for the possible dependence between the development time of the claim and its ultimate amount. More recently, (Delong et Wüthrich, 2020) and (Delong *et al.*, 2021) propose to use neural networks to model the

joint development of individual payments and claims incurred, thus benefiting from individual claims data in a collective reserving framework.

In addition to the constant increase in the quantity of data, the growing diversification of the products offered by insurers contributes to further complexifying the work of actuaries. Evolving in a very competitive world and taking advantage of the rise of new technologies, insurers must constantly remain aware of changes and evolutions in their clients' needs. To answer them, they often diversify their offer, multiplying coverages provided within a policy. Actuaries must refine their models to keep up with this diversity in their portfolios. To do so, they must consider the correlation between these different risks. Although some authors have already contributed to the claims reserving literature with models that include diverse forms of dependence, the specific inter-coverage dependence that we highlight in this paper has yet to be modelled in such a context, to the best of our knowledge. Among others, (Zhou et Zhao, 2010) and (Lopez, 2019) used copulas to model the dependence between the event times and delay in the development of a claim or the development time and the final amount of the claim. (Pešta et Okhrin, 2014) use time series and copulas to consider the dependence between payment amounts made at different stages in the development of a claim.

In the pricing literature, we find examples of joint modelling for correlated risks within a portfolio. In particular, (Frees et Valdez, 2008), and (Frees *et al.*, 2009) used copulas to model the dependence between different claim types in automobile insurance. They begin by identifying the coverage(s) impacted by a claim, then those for which a payment is made, before predicting the associated severity. Also, with automobile insurance data, (Shi *et al.*, 2016) account for the cross-sectional and temporal dependence among multilevel claims with a copula regression for multivariate longitudinal claims. Closer to what we intend to do in this paper, (Shi et Shi, 2022) use a multinomial logistic model for the occurrence of correlated risks. (Yang et Shi, 2019) apply a methodology similar to that of (Shi *et al.*, 2016) and model multilevel property insurance claims with a Gaussian copula while considering zero inflation. (Frees *et al.*, 2010) model the frequency and severity of multilevel property claims with, respectively, logistic regressions and Gaussian copulas. Finally, (Frees *et al.*, 2013) employ a similar approach for healthcare insurance multilevel claims and use multivariate binary regressions to model the different payment types, then Gaussian copulas again for the severities. More recently, (Côté *et al.*, 2022) extend the work of (Frees et Valdez, 2008) and (Frees *et al.*, 2009) by introducing a Bayesian model for multivariate and multilevel claim amounts, therefore facilitating the treatment of open claims which are of crucial importance in insurers' datasets.

This paper analyses a recent automobile dataset from a major Canadian insurer that provides multiple insurance coverages for each policy. Our goal is two-fold : first, we want to account for the multilevel feature of insurance claims, i.e. single claims that can impact multiple insurance coverages, in a reserving framework. We seek to model the total severity of the claim by breaking it down into the severities at the level of the different coverages that it impacts. We model the dependence between the different coverages included within a policy of the portfolio at hand. Second, we seek to separately model the main events that take place during the lifetime of a claim, namely the activation of a coverage and the time of payment which we consider as two separate events. As such, we contribute to the efforts to predict the development of the claims already illustrated in (Crèvecoeur *et al.*, 2023). We begin in Section 1.2 by introducing the dataset. Section 1.3.2 then presents the statistical model for reported but not settled (RBNS) claims based on activation patterns in which dependence between insurance coverages is captured via a multinomial logistic regression. In Section 1.4, we present the model's results applied to the Canadian dataset and compare them to the results of two additional models. We conclude our analysis in Section 1.5.

1.2 Data

This section presents an overview and some exploratory analysis of our dataset.

Our data contain 656,153 automobile insurance claims which have occurred between the 1^{st} of January 2015 and the 30^{th} of June 2021 in Canada or the United States. Each of these can impact one of four different insurance coverages provided for each policy by the insurer. In addition, we have information related either to the insured, the car driven, or the claim itself.

Note that, as is typically done in claims reserving, we work in a discrete-time framework and group the data in distinct development periods. Due to the limited number of calendar years available in the data and the short-tailed nature of the portfolio, we choose to split each calendar year into two periods of six months and use these as a basis for development periods. We illustrate this in Figure 1.1.

Using half-years as time steps will allow us, in the rest of this paper, to gain a more refined understanding of the development of claims throughout their lifetime. However, because academics and practitioners typically use one-year rather than six-month time steps in claims reserving, we will ensure that the model we develop in this paper for this specific portfolio remains general enough to be easily applied to other time units.

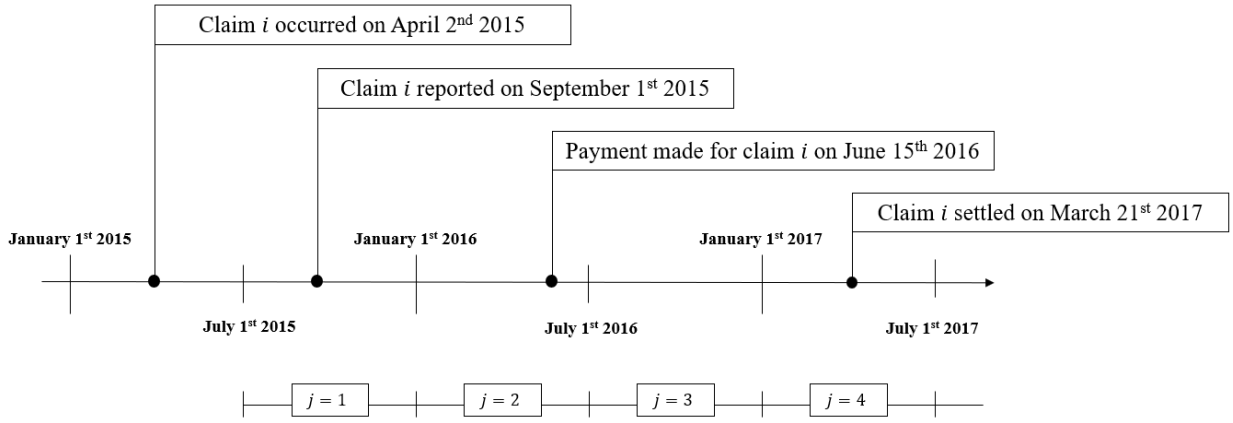

Figure 1.1 – Data grouping in periods of six months. In this illustration, claim i is reported on September 1^{st} 2015 after a delay of one period since its occurrence. It is then in its first ($j = 1$) development period. Payment is made for that claim in development period $j = 2$. The claim settles on March 21^{st} 2017, i.e., in its 4^{th} development period.

We first explore the four insurance coverages the insurance company offers for each policy in force before taking a closer look at the risk factors we will include in the analysis in Section 1.3.2.

1.2.1 Insurance coverages

Each of the 656,153 claims in our dataset impacts one or up to four coverages the insurer offers to the policyholders : the Accident Benefits coverage (loss of revenue, funeral expenses for the insured), the Bodily Injury coverage (loss revenue or medical expenses for a third party), the Vehicle Damage coverage (damages incurred to the insured's or a third party's vehicle) and the Loss of Use coverage (temporary replacement vehicle and alternative means of transportation).

To better understand the dynamics of our portfolio, Table 1.1 shows the importance of each coverage in terms of the proportion of claims, total cost, and total reserve, taking January 1^{st} 2019 as the valuation date. Unsurprisingly, the Vehicle Damage coverage is the most frequently observed, with 96.39% of the 656,153 claims activating it, while only 5.70% of the claims trigger the Bodily Injury coverage.

The proportions are slightly different when we look at the repartition of the total portfolio cost over the four coverages. In particular, Loss of Use is the coverage that weights the least cost-wise, representing less than 4% of the total cost of the portfolio. Despite being activated by less than 6% of all claims, the Bodily

Injury coverage still represents around 13% of the total portfolio cost. However, the most striking difference arises from the proportion of the total reserve attributed to each coverage. We calculate the percentages in the last column of Table 1.1 with the 1^{st} of January 2019 chosen as valuation date. Even though only 15% of all claims activate the Accident Benefits and/or Bodily Injury coverages, they amount to 85% of the total reserve. In comparison, the Loss of Use coverage activated by more than half the claims represents less than 1% of the total reserve.

Table 1.1 – Weight of each coverage in the portfolio. The percentages with respect to the total reserve are calculated by taking the 1^{st} of January 2019 as valuation date

Coverage	$\%$ of claims	$\%$ of total cost	$\%$ of total reserve
Accident Benefits	9.42	12.82	30
Bodily Injury	5.70	13.13	55
Vehicle Damage	96.39	70.44	14
Loss of Use	51.89	3.61	< 1

Table 1.2 shows descriptive statistics related to the payments made for each coverage. Bodily Injury claims are those with the largest average payments, while the Loss of Use coverage presents the lowest one, as was already hinted at in Table 1.1. The Accident Benefits, Bodily Injury, and Vehicle Damage coverages present large values for the payments in the higher quantiles, indicating that their corresponding distributions are probably heavy-tailed.

Table 1.2 – Descriptive statistics for the severity of payments of the four insurance coverages

Coverage	Mean	Std. dev.	Quantiles				Max.
			0.5	0.75	0.95	0.99	
Accident Benefits	12,386	53,561	3,215	6,909	47,757	127,896	2,435,334
Bodily Injury	23,271	76,027	4,000	15,150	98,449	322,612	2,039,570
Vehicle Damage	5,040	8,121	2,605	5,830	17,984	40,611	149,399
Loss of Use	545	620	419	714	1,000	2,336	52,777

Activation and payment delays. We define each insurance coverage's activation and payment delays. The activation delay corresponds to the delay (in period units) between the reporting date of the claim and the date the insurer first realises that this claim triggers the coverage. The payment delay is the time (in period units) between the activation date of a coverage and the date the insurer makes the first payment towards the claim related to that specific coverage. Note that we assume here that the activation of a coverage does not necessarily imply a payment. As means of illustration, consider again the example provided in Figure 1.1 and suppose that claim i only activates the Accident Benefits coverage directly upon reporting. Since the activation of the coverage takes place in the same development period as the reporting date, the activation delay for that coverage is equal to 0. The first payment is then made in development period $j = 2$, indicating a payment delay equal to 1 for the Accident Benefits coverage triggered by claim i.

Table 1.3 displays the average activation delay per coverage. The Accident Benefits, Vehicle Damage, and Loss of Use coverages are typically activated in the same period as the reporting date for the majority of claims (corresponding to the *No delay* column). However, for the Bodily Injury coverage, we observe that around 15% of the claims activate it with a delay of at least one development period after the reporting date.

Appendix A.1 further illustrates the dynamics between the different coverages by presenting the activation patterns observed in the dataset for the first development period of the claims.

Table 1.3 – Percentage of claims with different activation delays for the four coverages

	Activation delays				
Coverage	No delay	1 period	2 periods	3 periods	≥ 4 periods
Accident Benefits	93.84	5.73	0.29	0.08	0.06
Bodily Injury	85.86	9.86	1.46	1.13	1.69
Vehicle Damage	94.14	5.65	0.13	0.04	0.04
Loss of Use	92.10	7.73	0.13	0.03	0.01

Once a claim activates a coverage, we are interested in knowing when the first payment occurs. The payment delay, illustrated for our data in Figure 1.2, refers to the time (in six-month units) that elapsed between

the moment a claim activates a coverage and the date the insurer records a first payment for that same coverage. We see from Figure 1.2 that the Accident Benefits and Bodily Injury coverages, on the one hand, and the Vehicle Damage and Loss of Use coverages, on the other hand, display very similar shapes. The Loss of Use coverage is the one for which payments occur the fastest after activation, with over 80% of the claims related to that coverage receiving a payment in the same development period as the activation period. Almost all Loss of Use claims and Vehicle Damage claims will have received their first payment by the end of the second development period after activation of that coverage. Conversely, the Bodily Injury coverage is the one for which time to the first payment is the longest. Only around 20% of the claims that activate that coverage receive a first payment in the same development period as the activation of the coverage. Several Bodily Injury claims will receive their first payment only after the 8^{th} development period. The situation for the Accident Benefits coverage is very similar.

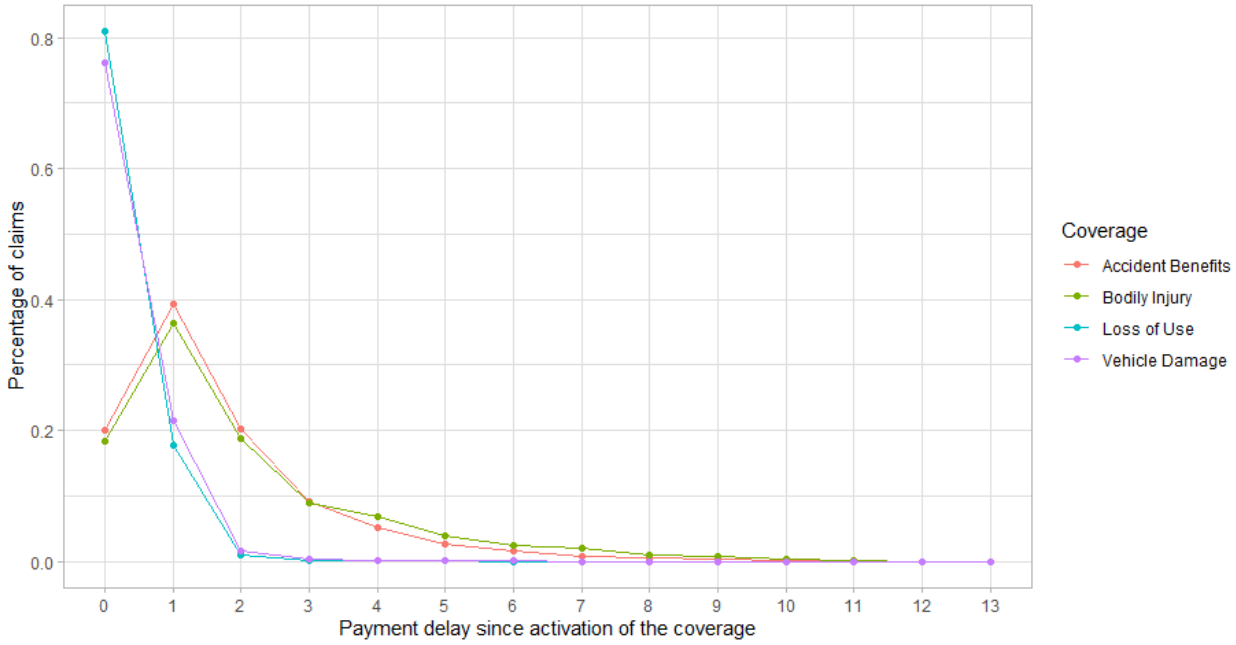

Figure 1.2 – Payment delays per coverage. A delay of 0 period means that the payment occurs in the same period the coverage becomes active

1.2.2 Risk factors

Table 1.4 presents the risk factors that we will use to build the reserving model. We provide further insights on these covariates in Appendix A.2. Note that the variables `GENDER` and `YOB` contained a significant proportion of missing values. We used the `R` package `mice` to fill them in. This package uses Fully Conditional

Specification (FCS) where binary data (GENDER) and unordered categorical data (YOB) are imputed using, respectively, logistic and polytomous logistic, regressions (see (Van Buuren et Groothuis-Oudshoorn, 2011) for more details). Note that this data-filling procedure did not cause any significant changes in the results we will discuss later.

Table 1.4 – Description of risk factors

Risk factors	
GENDER	Gender of the insured.
YOB	Year of birth : decade during which the insured was born.
VU	Use of the vehicle made by the insured.
AM	Annual distance driven (in km).
PROV	Place of occurrence of the claim : one of the Canadian provinces or the USA.
FR	Fault rating : evaluation of the insured's level of responsibility in the accident.
REP_DELAY	Reporting delay (in 6-months periods) between the occurrence and reporting dates of the claim.
ACT_DELAY	Activation delay per coverage (in 6-month periods) between the reporting date of a claim and the first date at which the insurer realises the claim triggers the corresponding coverage.
PAY_DELAY	Delay (in 6-month periods) between the activation date of a coverage by a claim and the date at which a first payment is made towards that claim relating to that specific coverage.

1.3 An activation pattern model for claims reserving

This section introduces the model we built to predict the granular reserves for the portfolio of claims described in Section 1.2 considering the dependence between the insurance coverages. Note that even though we tailor the model to the specific needs of our dataset, our goal is to provide the most general description of the model so that an interested party may easily apply it to other studies.

We first specify the notation and the model components before presenting the model more formally for

different development periods and proposing a simulation routine.

Figure 1.3 shows the typical development process for a single claim. When a claim occurs, the policyholder reports it to the insurance company. As we explained in Section 1.2, this can happen with or without delay. Once reported, the insurer records the claim and opens a new file in his claims management system. The reporting of the claim activates at least one of the coverages provided with the policy. The insurer can then start making payments to the policyholder. During its development, new information related to the claim can be brought to the insurer, which can result in the activation of one or more additional coverages. Then the insurance company will continue to make payments until the settlement of the claim.

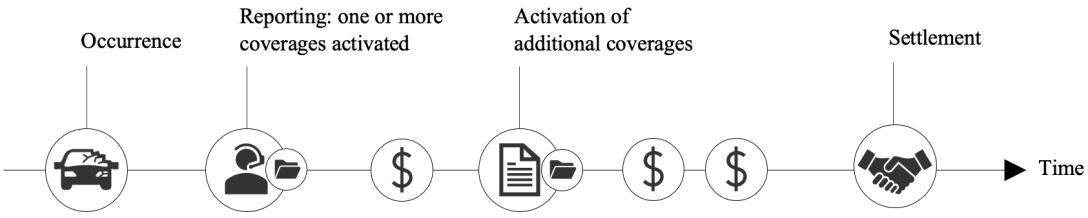

Figure 1.3 – Development process of a claim

We seek to model how a claim can activate multiple coverages upon reporting or later and the underlying dependence between them.

1.3.1 Notation

Let c with $c = 1, \ldots, C$ denote a coverage provided within the policy by our insurer. Each policy in force can incur a claim that will impact one or more of the C insurance coverages that the policyholder benefits from. For each claim i, with $i = 1, \ldots, n$ and development period j with $j = 1, \ldots, J$, we define the following components :

— $\boldsymbol{A}_{i,j}$ is a $1 \times C$ random vector whose entries $A_{i,j,c}$ each take the value 0 or 1, i.e., $\boldsymbol{A}_{i,j} \in \{0,1\}^C$. We call $\boldsymbol{A}_{i,j}$ the activation pattern vector that indicates which of the C coverages claim i activates in its j^{th} development period. Given C insurance coverages and assuming that at least one must be activated when claim i is reported, there are $V = 2^C - 1$ different activation pattern vectors in development period j. We denote $\mathcal{V}$ the set of these V patterns. Note that at this stage, we do not assume any constraints on the evolution of $\boldsymbol{A}_{i,j}$ over time in order to keep the definition as general

as possible. Further details will however be added for the sake of our application to the dataset at hand in Section 1.3.2.

— $(P_{i,j,c}|A_{i,j,c} = 1) \in \{0, 1\}$ is a random variable that takes the value 1 if, for the corresponding coverage c, the insurer makes at least one payment for claim i in development period j, given that this coverage is active, or the value 0 otherwise. It is undefined if $A_{i,j,c} = 0$. We call $P_{i,j,c}$ the payment pattern that depends on the random vector $\boldsymbol{A}_{i,j}$.

— $(Y_{i,j,c}|P_{i,j,c} = 1) \in \mathbb{R}^+$ is a positive continuous random variable that represents the severity associated with coverage c when payments have been made for it for claim i in development period j. $Y_{i,j,c}$ is thus not defined if the corresponding $P_{i,j,c}$ is equal to 0, i.e., if there was no payment for coverage c and claim i in its j^{th} development period. For claims with longer development delays, we also define the random variable $\tilde{Y}_{i,j,c} \in \mathbb{R}^+$ that represents the total remaining severity of claim i for coverage c, from development period j until settlement of the claim. We will discuss these longer claims in Section 1.3.2.

— $\boldsymbol{x}_i'$ is a $1 \times m$ vector of covariates for claim i that contains static risk factors.

Example 1. To illustrate the notation, we consider a simple example with $C = 2$ different insurance coverages. The set of possible activation patterns is $\mathcal{V} = \{(1\ 0)\ \ (0\ 1)\ \ (1\ 1)\}$, corresponding respectively to the scenarios where a claim i activates either only the first coverage, only the second coverage or both of them in the same development period j. For each of these three scenarios, various realisations of $(P_{i,j,c}|A_{i,j,c} = 1)$ and $(Y_{i,j,c}|P_{i,j,c} = 1)$ are possible, as illustrated in Figure 1.4.

1.3.2 Statistical model

Using the notation introduced in Section 2.3.1, we focus now on presenting our model for various development periods. We begin in the first development period before moving on to the following ones up to period j^*. We then present what happens after period j^* for claims with longer settlement delays. In our notation, the insurer chooses j^* as the development period from which the activation pattern for a claim stabilises, i.e., the last development period in which the claim might still activate additional coverages. This *stabilisation point* in the activation of coverages may arise from, among others, the nature of the claims (automobile insurance claims typically present very short settlement delays) or even the legislative system in which the insurer evolves (some types of claims may not be pursued if a certain period has passed since

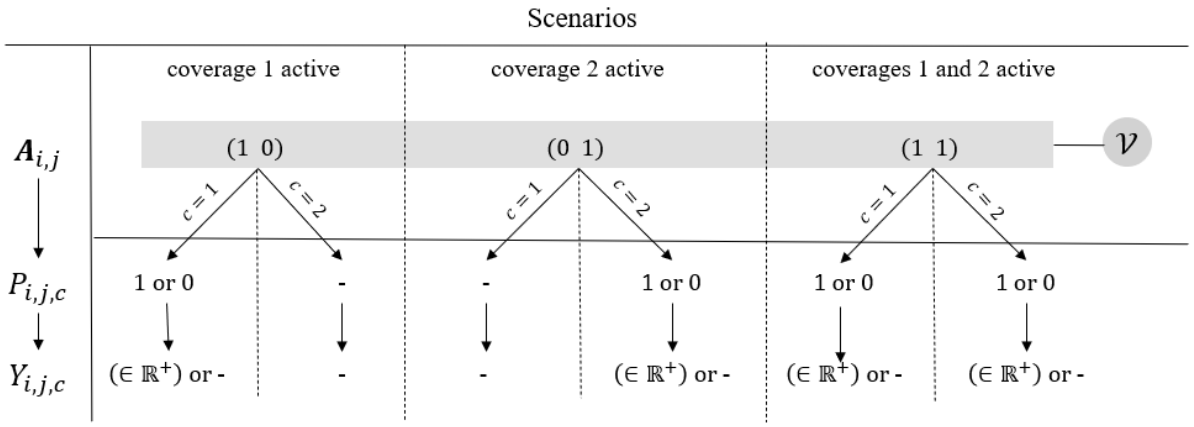

Figure 1.4 – Illustration of the possible scenarios with $C = 2$ insurance coverages. The symbol "-" stands for the cases where the patterns are not defined.

occurrence). We finally describe the severity models used for each insurance coverage.

Knowing the number of claims reported in any given period, our model allows us to determine which coverages these claims activate and to predict their future development until development period j^*. Thus, we could apply the model to reported but not paid (RBNP) claims and reported but not settled (RBNS) claims. The model does not provide a complete framework for predicting incurred but not reported (IBNR) claims. It lacks a process predicting the number and characteristics of the IBNR claims. However, such processes have been widely discussed in the literature (see, for example, Antonio and Plat (2014) (Antonio et Plat, 2014) where the authors use a Position Dependent Marked Poisson Process) and can easily be used as a preliminary step to the model presented here. Once the insurer knows the number of IBNR claims, he may use the model to estimate their development from the first period onward.

Development period $j = 1$ For claim i in development period $j = 1$ and given C insurance coverages, there exists $V_0 = 2^C - 1$ possible realisations of the random activation pattern vector $\boldsymbol{A}_{i,1}$, all included within the set $\mathcal{V}_0$. This is the initial set of all possible activation patterns once a claim has entered the reporting system of the insurer. We write these realisations $\boldsymbol{v}$ and use a multinomial logit regression to model the probability that $\boldsymbol{A}_{i,1}$ takes one of them :

$$\mathrm{P}[\boldsymbol{A}_{i,1} = \boldsymbol{v}] = \frac{\exp\left(\boldsymbol{x}_i'\boldsymbol{\beta}_{1,v}\right)}{\sum_{\nu\in\mathcal{V},}\exp\left(\boldsymbol{x}_i'\boldsymbol{\beta}_{1,v}\right)}, \tag{1.1}$$

where $\boldsymbol{x}_i'$ is the vector of covariates introduced in Section 2.3.1. $\boldsymbol{\beta}_{1,v}$ is the $1 \times m$ vector of parameters that can vary with the different coverages and thus depends on the v^{th} pattern.

Knowing which coverages claim i activates in the first development period thanks to $\boldsymbol{A}_{i,1}$, the insurer can move on to the next step of the claim development process depicted in Figure 1.1 and determine which of the active coverages will incur a payment within the period. For each insurance coverage c, we assume that

$$(P_{i,1,c}|A_{i,1,c}=1) \sim \text{Bernoulli}\big(\pi_{1,c}(\boldsymbol{x}_i, \boldsymbol{\gamma}_{1,c})\big). \tag{1.2}$$

For development period $j=1$, the probability $\pi_{1,c}(\boldsymbol{x}_i, \boldsymbol{\gamma}_{1,c})$ is given by

$$\pi_{1,c}(\boldsymbol{x}_i, \boldsymbol{\gamma}_{1,c}) = \frac{\exp\left(\boldsymbol{x}_i'\boldsymbol{\gamma}_{1,c}\right)}{1+\exp\left(\boldsymbol{x}_i'\boldsymbol{\gamma}_{1,c}\right)}.$$

The vector of covariates $\boldsymbol{x}_i'$ is the same as the one used for the activation patterns in Equation (1.1) and $\boldsymbol{\gamma}_{1,c}$ is the $1 \times m$ vector of coefficients for the period $j=1$ varying with each coverage c.

Once we know which coverages are active in period 1 and which have incurred a payment, we can calculate the corresponding severity per coverage for claim i. The actuarial literature contains many examples of the use of Generalized Additive Models for the Location, Scale, and Shape (GAMLSS) for this purpose, thanks to the high level of flexibility that these models provide (see (Stasinopoulos *et al.*, 2017) for more details on the use of GAMLSS). We use them to predict the expected severity incurred by claim i :

$$\mathrm{E}[Y_{i,1,c}|P_{i,1,c}=1] = g^{-1}(\boldsymbol{x}_i'\boldsymbol{\alpha}_{1,c} + \alpha^*_{1,c}1), \tag{1.3}$$

where $g(.)$ is the link function and $\boldsymbol{\alpha}_{1,c}$ is the $1 \times m$ vector of coefficients for the first development period that depend on the insurance coverages. $\alpha^*_{1,c}$, or more generally for any period j, $\alpha^*_{j,c}$, is the specific coefficient attached to the development period in which claim i is at the valuation date. As such, the insurer can apply this severity model in all development periods j. The predicted severity will depend on the moment the claim is in its lifetime at the valuation date.

With the activation patterns, payment patterns, and corresponding severities, we know which coverages claim i activates in development period j and the severity associated with each coverage.

Note that this model implies that the dependence between the different insurance coverages is accounted for solely by the multinomial logistic regression in Equation (1.1). We further assume the different claims in our dataset to be independent. We only consider dependence between coverages, arising from the activation patterns, while payment patterns and associated severities are all modelled independently.

Development periods $j = 2, 3, \ldots, j^*$ Knowing what happened in at least one development period, we

now describe how the insurer can keep modelling the development of his RBNS claims until they reach development period j^*, the period from which he assumes that no more coverage can become active.

We assume that once an insurance coverage is active, it remains active until settlement of the claim. For claim i and coverage c in development period $j = 2, 3, \ldots, j^*$, we then have

$$(A_{i,j,c}|A_{i,j-1,c} = 1) = 1, \quad 1 < j \leq j^*. \tag{1.4}$$

As such, we now depict the set of possible patterns for $\boldsymbol{A}_{i,j}$ as $\mathcal{V}_j \subset \mathcal{V}_0$, containing all the possible realisations $\boldsymbol{v}$. The subset $\mathcal{V}_j$, the possible realisations that it contains, and the number V_j of these possible realisations depend on the activation pattern in development period $j-1$, $\boldsymbol{A}_{i,j-1}$. Knowing this, we use a multinomial logit regression again and write

$$\begin{aligned} &\mathrm{P}[\boldsymbol{A}_{i,j} = \boldsymbol{v}|\boldsymbol{A}_{i,j-1}] \\ &\quad = \begin{cases} \frac{\mathrm{P}[\boldsymbol{A}_{i,j}=\boldsymbol{v}]}{\sum_{\boldsymbol{v}\in\mathcal{V}_j}\mathrm{P}[\boldsymbol{A}_{i,j}=\boldsymbol{v}]}, & \text{if } \boldsymbol{v} \in \mathcal{V}_j \\ 0, & \text{otherwise,} \end{cases} \end{aligned} \tag{1.5}$$

where we further assume a Markovian property for the activation pattern vectors and build them using only the information available in the previous period.

Moreover, we assume the Markovian property for the payment patterns, and we have

$$(P_{i,j,c}|A_{i,j,c} = 1) \sim \text{Bernoulli}\big(\pi_{j,c}(\boldsymbol{x}_i, \boldsymbol{\gamma}_{j,c})\big), \tag{1.6}$$

where

$$\pi_{j,c}(\boldsymbol{x}_i, \boldsymbol{\gamma}_{j,c}) = \frac{\exp\left(\boldsymbol{x}_i'\boldsymbol{\gamma}_{j,c}\right)}{1 + \exp\left(\boldsymbol{x}_i'\boldsymbol{\gamma}_{j,c}\right)},$$

with the parameter vectors $\boldsymbol{\gamma}_{j,c}$ depending once again on the specific insurance coverage c and the vector of risk factors $\boldsymbol{x}_i'$ is the same $1 \times m$ vector of covariates as before.

Knowing which coverages are active in period j and which incurred a payment, we can finally determine the severity of these payments using a similar model to the one used in the first development period.

$$\mathrm{E}[Y_{i,j,c}|P_{i,j,c} = 1] = g^{-1}(\boldsymbol{x}_i'\boldsymbol{\alpha}_{j,c} + \alpha^*_{j,c}j). \tag{1.7}$$

Development period $j > j^*$ For a given claim i in development period $j > j^*$, the insurer should not use again the model described in Section 1.3.2 to estimate the activation and payment patterns for the new

period since we assume that these will not change anymore after development period j^*. Instead, he can directly model the claim settlement by calculating its remaining severity per coverage, $\tilde{Y}_{i,j^*,c}$, using again GAMLSS as described in Equation 1.7. $\tilde{Y}_{i,j^*,c} > 0$ encapsulates the remaining amounts for coverage c at later dates for claims with longer lifetimes and whose activation pattern is not expected to change anymore.

Figure 1.5 illustrates the full model described in Section 1.3.2, starting in development period $j = 1$ until period $j > j^*$.

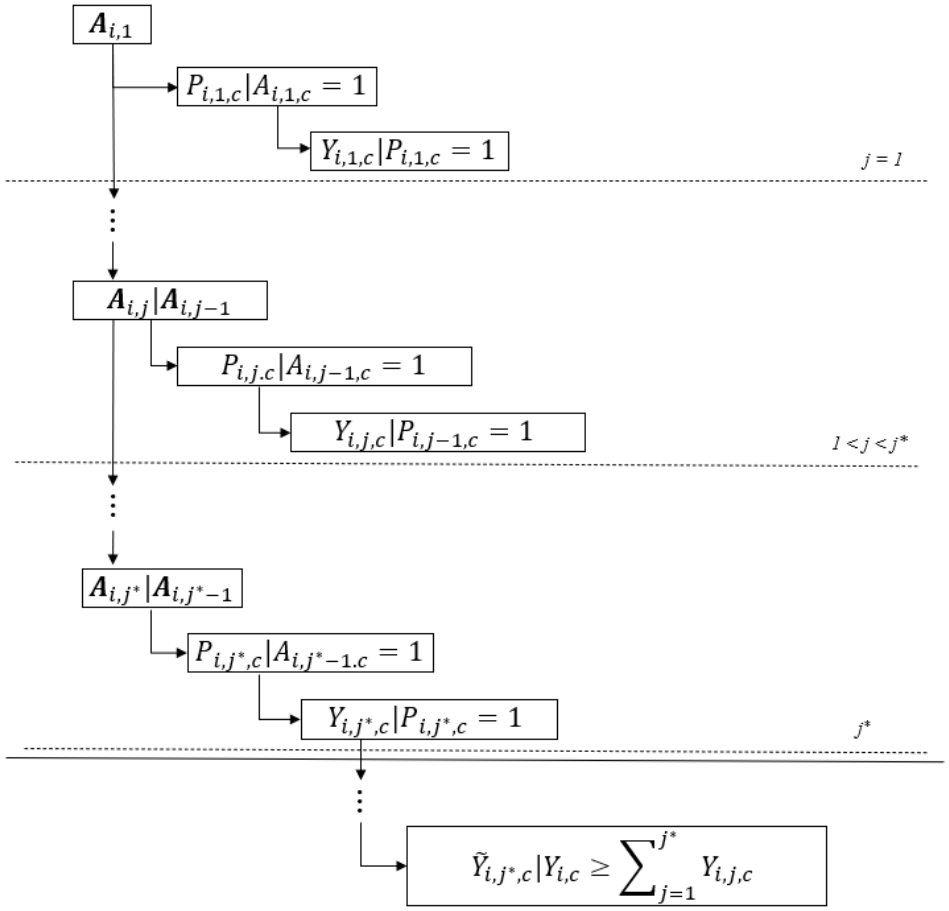

Figure 1.5 – Model illustration.

1.3.3 Simulation routine

To properly consider all open claims in a given dataset, we create a simulation routine that we can tailor to the specific development stage of different types of claims. Even though we will solely focus on the RBNS claims for our case study, we present the routine for both IBNR and RBNS claims, including those with longer

development times for which we assume that they will not activate any additional coverages.

IBNR claims. Let i denote an IBNR claim such that $i \in \mathcal{I}_{\text{IBNR}}$, where the insurer already obtained the set $\mathcal{I}_{IBNR}$ using one of the claims occurrence processes available in the literature. For ease of notation, we drop the subscript i in the description of the routine.

1. For development period $j = 1$:
   (a) Simulate the first activation pattern $\boldsymbol{A}_1$, using the parameters estimated for the multinomial logit model in Section 1.3.2.
   (b) For the inactive coverages, i.e. $A_{1,c} = 0$, then the corresponding $P_{1,c}$ is not defined. For the active coverages only, simulate the payment patterns using the Bernoulli models with parameters $\hat{\pi}_{1,c}$.
   (c) If $(P_{1,c}|A_{1,c} = 1) = 0$, then the corresponding severity for the IBNR claim related to that coverage is not defined (we will set it as zero in the simulations). If $(P_{1,c}|A_{1,c} = 1) = 1$ from the previous step, we simulate the severity from the distribution selected for the specific coverage.
2. For development periods $1 < j \leq j^*$:
   (a) Knowing the activation pattern $\boldsymbol{A}_{j-1}$ for the previous development period, only keep the activation patterns that are now still possible. Re-normalize the probabilities for the remaining activation patterns and simulate them using the multinomial logit model from Section 1.3.2 to obtain the vector $\boldsymbol{A}_j|\boldsymbol{A}_{j-1}$.
   (b) If $A_{j,c} = 0$, $P_{j,c}$ is not defined. For the active coverages only, simulate the payment patterns using the Bernoulli models with parameters $\hat{\pi}_{j,c}$.
   (c) If the payment indicator $(P_{j,c}|A_{j,c} = 1) = 0$, the severity for the corresponding coverage is not defined. If $(P_{j,c}|A_{j,c} = 1) = 1$ from the previous step, simulate the severity from the distribution selected for the specific coverage.
   (d) Repeat steps (a)-(c) until reaching development period j^*. After j^*, simulate the remaining severity per coverage, $\tilde{Y}_{j^*,c}$. The total severity simulated for this IBNR claim is given by$Y^{\text{IBNR}} = \sum_{j=1}^{j^*}\sum_{c=1}^{C} Y_{j,c} + \sum_{c=1}^{C} \tilde{Y}_{j^*,c}$.

RBNS claims. Let $i \in \mathcal{I}_{\text{RBNS}}$. At the valuation date, these claims are already in development period $j > 1$. We already know the activation patterns, payment patterns, and payments for development periods $1, \ldots, j-1$. The simulation routine for these claims is then exactly the same as that described for the IBNR claims for development periods $1 < j \leq j^*$.

RBNS claims with longer development times. To remain thorough in our analysis, we must also consider the longer RBNS claims we previously mentioned in Section 1.3.2. After a certain development period j^*, we assume that the activation pattern for claim $i \in \mathcal{I}_{\text{RBNS}}$, $\boldsymbol{A}_{i,j^*}$ remains unchanged. If claim i is still open in development period $j^* + k$ for $k \geq 1$, we simulate the remaining severity starting from period $j^* + k$ using the severity model previously described that allows the account for the specific development period in which claim i is at the valuation date. We add the condition that the final severity per coverage for that claim must be at least equal to the sum of the payments observed up to period $j^* + k$, i.e., $(\tilde{Y}_{i,j^*+k,c} | Y_{i,c} \geq \sum_{j=1}^{j^*+k-1} Y_{i,j,c}), c = 1, 2$. The total severity simulated for this RBNS claim is $Y_i^{\text{RBNS}} = \sum_{j=1}^{j^*+k-1} \sum_{c=1}^{C} Y_{j,c} + \sum_{c=1}^{C} \tilde{Y}_{j^*+k,c}$.

1.4 Numerical Application

Now that we have fully described the statistical model, we dedicate this section to its application to our dataset. We first provide some insights into how we estimate the different model parameters before presenting the results obtained by simulations for the total reserves. We finally compare these reserves to those obtained using two other models that we describe in detail below.

1.4.1 Estimation

We apply our model to the data introduced in Section 1.2. We present the estimation results performed for the three components of the model : the activation patterns, payment patterns, and payment severities.

From our observations in Section 1.2, we assume that the activation patterns in our dataset do not change anymore after development period $j^* = 4$. Considering the information provided in Table 1.3, where we see that almost all claims activate the different insurance coverages no later than four periods after their reporting date, this seems to be a valid assumption.

Appendix A.3 illustrates how we separate our dataset into a training and a valuation set, choosing the 1^{st} of January 2019 as valuation date of the reserves.

Since we have $j^* = 4$ and only consider RBNS claims, we use the model starting from Section 1.3.2 for claims in development periods $j = 2$ to $j = 4$ and from Section 1.3.2 for claims in development periods $j > 4$. Before going further with the model fitting and reserves simulations, we illustrate in Example 2 and Figure

1.6 how the model handles a given claim from our dataset.

Example 2. We consider claim i, illustrated in Figure 1.6, for which we want to estimate the reserves on January 1^{st}, 2019. Claim i was reported in the period ranging from January 1^{st} 2018 to June 30^{th} 2018. This is its first development period, $j = 1$. At the valuation date, we observed the activation patterns, payment patterns, and payment severities for the first and second development periods, as claim i enters development period $j = 3$. We now have to predict the activation patterns, payment patterns, and payment severities for development periods $j = 3$ and $j = 4$. Since we assume that the activation patterns do not change anymore after the 4^{th} development period, we can then predict the total remaining severity as of development period $j > 4$.

Activation patterns. Using the same notation as in Section 1.3.2, we have $C = 4$ insurance coverages, leading to $V_0 = 15$ possible activation patterns. We fit the multinomial logit model for the activation patterns using maximum likelihood estimation with the likelihood function given by

$$\mathcal{L}(\boldsymbol{A}_{i,j}) = \prod_{i=1}^{n} \frac{\exp\left(\boldsymbol{x}_i' \boldsymbol{\beta}_{j,v}\right)}{\sum_{k=1}^{V} \exp\left(\boldsymbol{x}_i' \boldsymbol{\beta}_{j,k}\right)}. \tag{1.8}$$

In Appendix A.4, we provide estimated parameters for each of the $V_0 = 15$ possible activation patterns.

We observe that the impact of some risk factors moves away from zero for the 5^{th} pattern, i.e., the one in which a claim simultaneously activates the Bodily Injury and Loss of Use coverages. In particular, due to the legislation regarding automobile insurance in Quebec and Saskatchewan, the probability of observing that specific activation pattern decreases substantially if the claim occurred in one of these provinces. In the province of Quebec, the Bodily Injury coverage does not exist : a person injured in a car accident will benefit from the public automobile insurance plan provided by the Société de l'assurance automobile du Québec (SAAQ) rather than from a private insurance company. In Saskatchewan, the insured must choose between a *no-fault* and a *tort* auto injury insurance coverage offered by the Saskatchewan driver's licensing and vehicle registration. Almost all residents choose the *tort* coverage under which they are insured regardless of whether they are at fault or not. As such, the Bodily Injury coverage rarely appears for claims in that province. Consequently, we see that the parameter estimates for Quebec and Saskatchewan are always under 0 for all activation patterns, including the Bodily Injury coverage.

Moreover, the birth year is also a big driver for this particular pattern : the probability of observing it increases for the younger insureds born in the 1990s and after 2000. We commonly observe in automobile insurance data that younger drivers tend to cause more frequent and more severe accidents. Consequently, we can expect more accidents with injuries or even casualties and loss of use in the younger population.

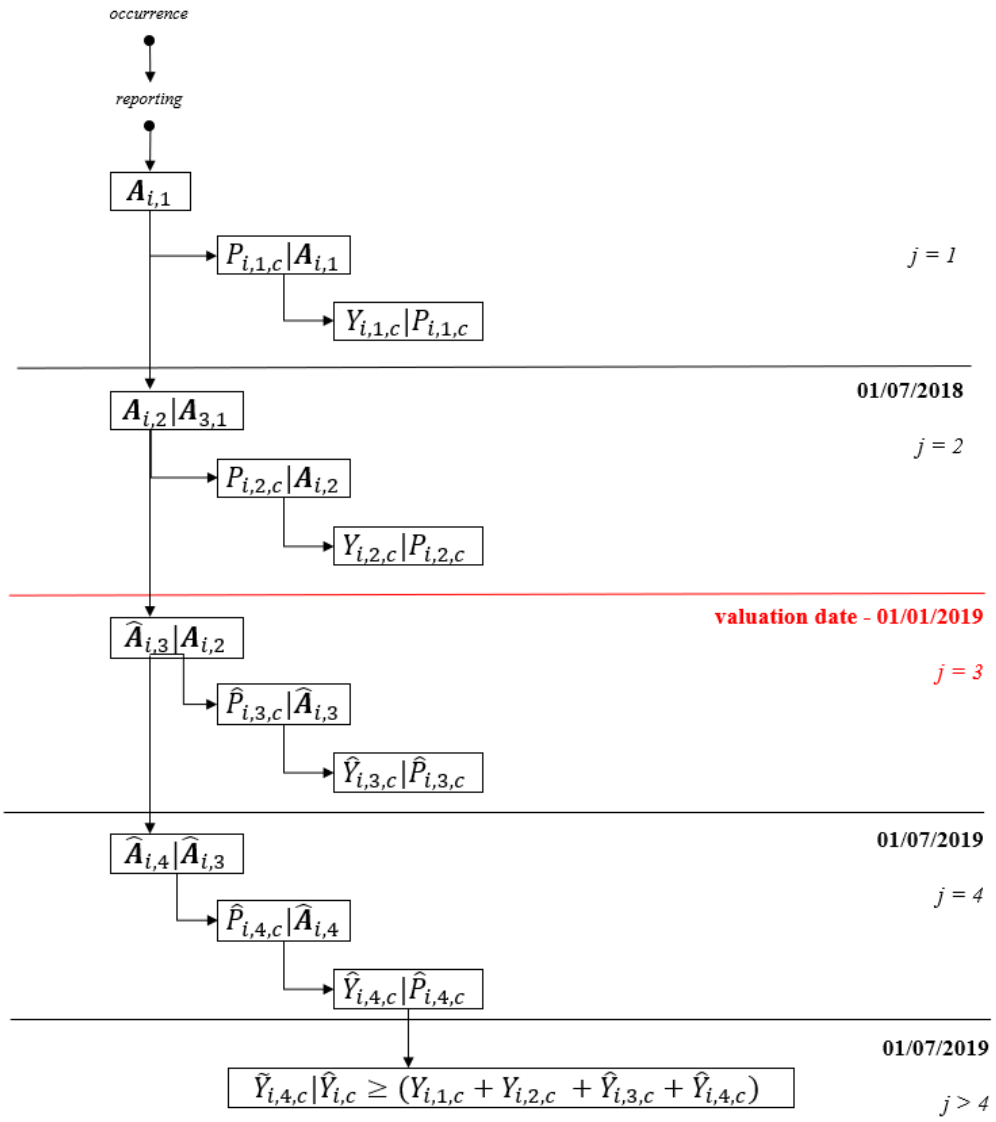

Figure 1.6 – Illustration of how the model handles a specific claim from the dataset, taking the 1st of January 2019 as valuation date.

Appendix A.4 illustrates this well, where the insureds born after 2000 are the only cohort that always has an increasing impact on the probability of observing each activation pattern, even though they only represent 1.52% of all the insureds.

Payment patterns. We fit twelve Bernoulli regressions for the payment patterns : one for each of the four coverages in development periods $j = 2$, $j = 3$, and $j = 4$ and use again maximum likelihood optimisation to obtain parameter estimates for each of the models. Table 1.5 displays the results of these estimations.

Table 1.5 – Fitted average probabilities of observing a payment

Probability	$j = 2$	$j = 3$	$j = 4$
$\pi_{j,\text{AB}}$	0.2367	0.2194	0.1267
$\pi_{j,\text{BI}}$	0.1426	0.1220	0.0695
$\pi_{j,\text{VD}}$	0.6391	0.1030	0.0311
$\pi_{j,\text{LoU}}$	0.4228	0.0368	0.0087

Payment severities. From Table 1.2, we note the importance of considering heavy-tailed distributions to model the severity of payments, as is usually done in the actuarial literature. Among others, Frees et al. (2009) (Frees *et al.*, 2009) opt for the Generalized Beta of the second kind distribution to accommodate the long-tailed nature of claims. Figure 1.7 shows the histogram of the payments made in the different development periods considered in the model for the Accident Benefits and Bodily Injury coverages. Note that for both of them, we cut out most of the tail of the histograms in the right part of the graph to ease the readability.

For each coverage, we consider five commonly used distributions : the Log-Normal, Gamma, Pareto, Generalized Beta of the second kind, and Weibull distributions. As explained in Section 1.3.2, we fit each of these models for all possible development periods, adding a covariate that specifies the development period in which a given claim is at the valuation date. Appendix A.5 presents the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for all models and all four insurance coverages. We conclude that the preferred distribution for the Accident Benefits, Vehicle Damage, and Loss of Use coverages is the Generalized Beta of the Second Kind. We select the Weibull distribution for the Bodily Injury coverage.

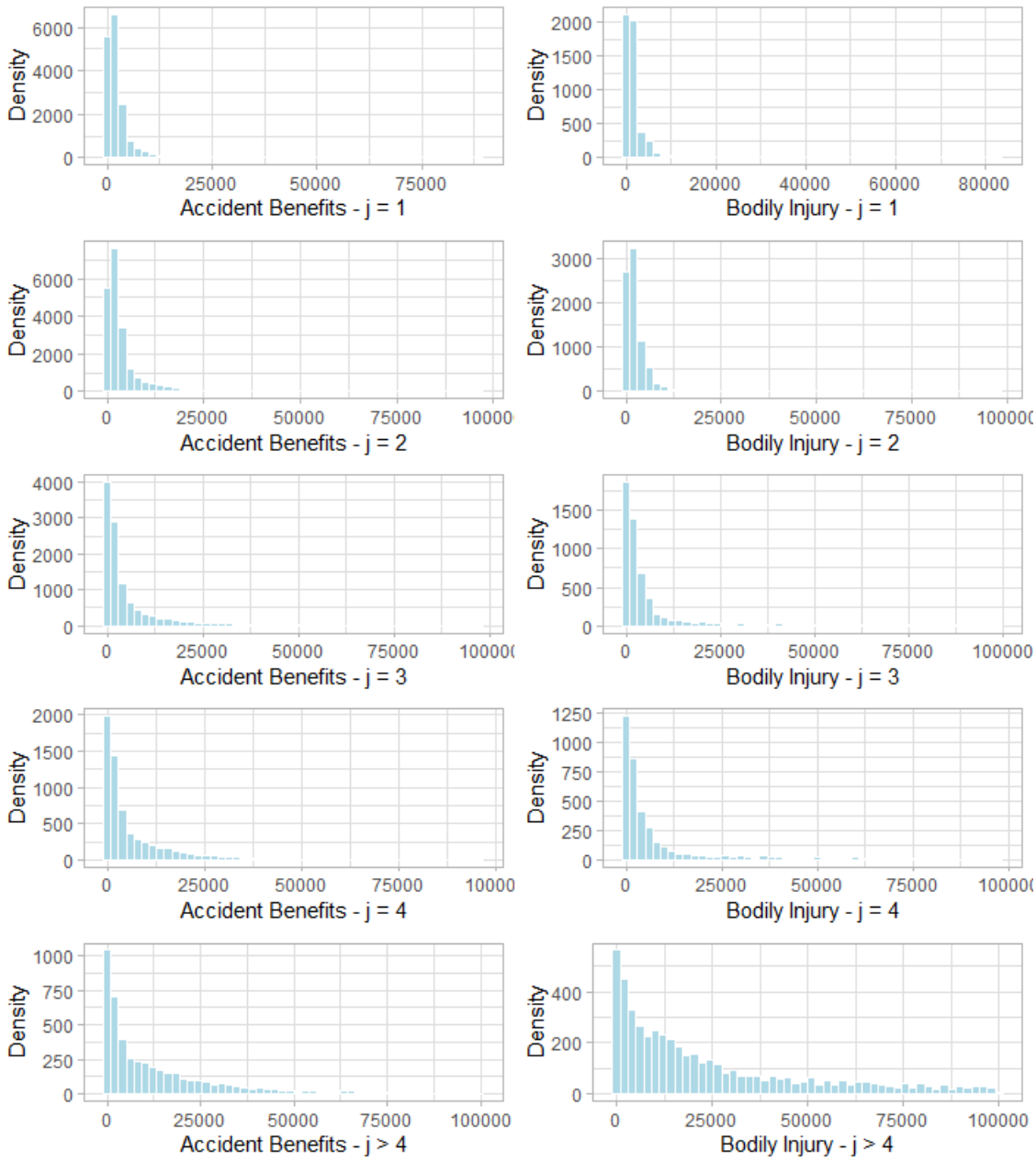

Figure 1.7 – Histograms of the observed payments for the Accident Benefits and Bodily Injury coverages

1.4.2 Predictive distributions and model comparisons

We focus in this section on presenting the results obtained for the estimation of the reserves when we apply the model described in Section 1.3.2 to the dataset described in Section 1.2, choosing the 1^{st} of January 2019 as the valuation date. We present the predictive distribution of the reserves for the portfolio as a whole and the different insurance coverages. We compare these results to the observed reserve amount and the one

obtained from fitting some more classical and commonly used reserving models.

For all models we consider in this section, we perform at least 5,000 simulations using the simulation routine described in Section 1.3.3. As shown in Appendix A.6, this large number brings stability to our results. We first present the estimations from the activation patterns model before comparing them to other reserving models.

Figure 1.8 displays the predictive distribution of the total RBNS reserve of our portfolio. In this graph and all those that will follow, the red line marks the observed reserve, the blue dotted line shows the average value of the predictions, and the continuous blue line depicts the 0.95 quantile of the distribution.

For these claims, the observed reserve amount is 524.91M CAD. Note that the true amounts that we provide henceforth are minimum amounts because more than 1% of the claims are still open in the portfolio at the valuation date. The final severity for these claims will probably be greater than the one observed now once they become settled.

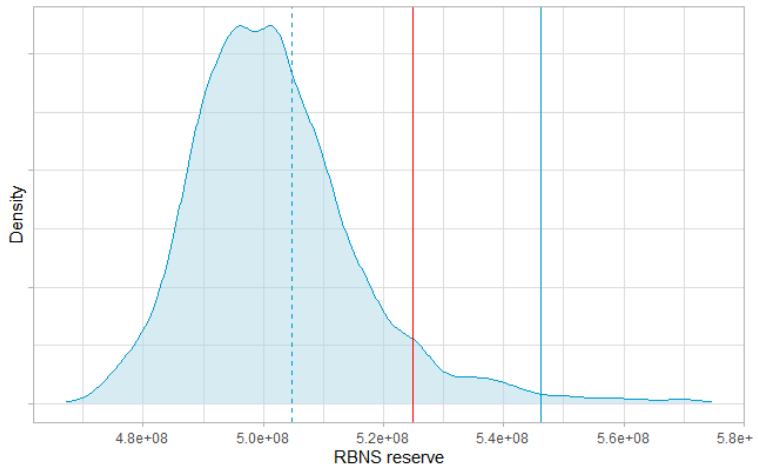

Figure 1.8 – Simulated RBNS reserves for the full portfolio as of the 1st of January 2019. The red line, dashed blue line and continuous blue lines depict, respectively, the observed reserve amount, the average and the 0.95 quantile of the simulations

Figure 1.9 presents the predictive distributions obtained for the reserves of the Accident Benefits and Bodily Injury insurance coverages. We provide further detail on these results in Table 1.6. In particular, we display the predicted reserves for the four coverages and the whole portfolio using six-month and one-year periods,

the latter being more commonly used by academics and practitioners alike. In the remainder of this section, we will only use the results obtained with the six months periods for the activation patterns model and other models used for comparison. Since financial legislation around the world typically requires insurers to set aside an amount based on a high quantile of the predictive distribution of their reserves, we provide the 95% Value-at-Risk (VaR) in addition to the mean for each of the four coverages and the portfolio as a whole.

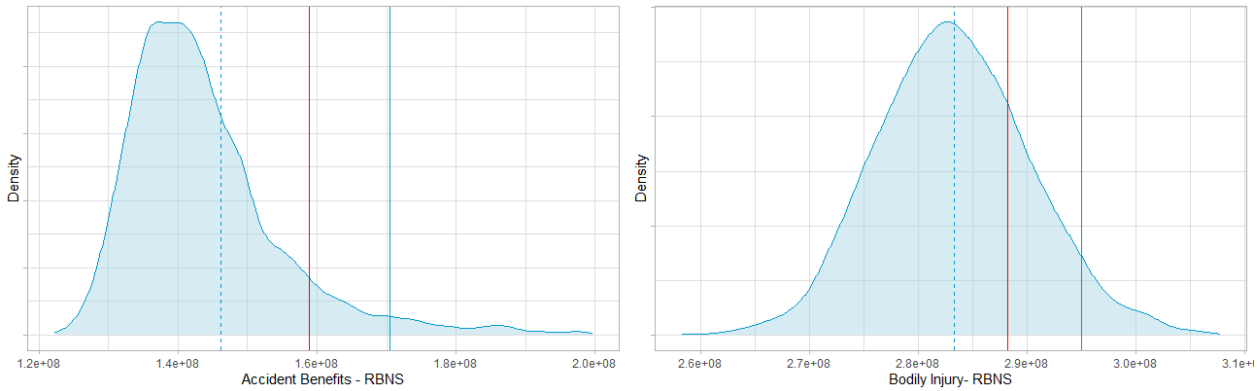

Figure 1.9 – Simulated RBNS reserves for the Accident Benefits (left) and Bodily Injury (right) coverages. The red lines, dashed blue lines and continuous blue lines depict, respectively, the observed reserve amounts, the average and the 0.95 quantile of the simulations

Table 1.6 – Simulated RBNS reserves (in M CAD) using the activation patterns with development periods of 6 months (columns 3 and 4) and 1 year (columns 5 and 6).

Coverage	True reserves	Act. Patterns (6 mo.)		Act. Patterns (1 y.)	
		Mean	$\text{VaR}_{0.95}$	Mean	$\text{VaR}_{0.95}$
Accident Benefits	>158.90	146.61	170.46	162.15	196.64
Bodily Injury	>288.17	283.27	295.02	300.02	312.58
Vehicle Damage	>73.63	70.77	76.43	70.80	72.95
Loss of Use	>4.20	4.17	4.24	3.41	4.01
Global reserve	>524.91	504.82	546.15	536.86	574.58

For each coverage and the portfolio, the average predictions underestimate the observed reserves, while the predictions obtained for the 95% VaR are all greater than the observed amounts. This is particularly

the case for the Accident Benefits and Bodily Injury coverages where the VaR amounts are, respectively, 170.46M CAD and 295.02M CAD, while the observed reserves are equal to 158.90M CAD and 288.17M CAD, respectively. For the Loss of Use coverage that represents less than 1% of the total reserve, the quantile of the predictions is close to the observed amount.

Table 1.6 illustrates that for a short-tailed claims dataset such as ours, using 6-month rather than one-year development periods leads to more accurate predictions, both for the portfolio as a whole and for each separate coverage.

Next, we pursue the analysis of our results by comparing them to those obtained using a classic aggregate model and a replica of the activation patterns model that does not take dependence into account.

Activation patterns model vs aggregate models. Table 1.7 compares the reserve estimates obtained with the proposed model based on the activation patterns to those of the classical Overdispersed Poisson (ODP) Chain Ladder and to the observed reserves, both for the portfolio and the four insurance coverages taken separately. Since we only focus our analysis on RBNS claims and for the sake of comparison with the other models, we use reporting years instead of occurrence years when building the claims triangles. We further assume that all underlying assumptions of the ODP model are verified but do not validate them here since this is a standard model in claims reserving and our aim is to focus on a comparison with the activation patterns model.

Note that in Table 1.7, the global reserve for the ODP model is not equal to the sum of the reserves per coverage. It is due to building the claims development triangles and applying the model to each coverage taken separately, as well as to the portfolio, rather than just taking the sum of the reserves obtained per coverage to obtain the total reserve. Before taking a closer look at the results obtained with the Overdispersed Poisson Chain Ladder, we remind the reader that we only have a limited number of calendar years observed in our dataset with a claim history spanning from 2015 to mid-2021. Even if we consider development periods of six months rather than the more traditional one-year periods, our claims development triangles are still based on a small number of claims. We must be careful with the results obtained with an aggregate reserving method, particularly when we apply it to coverages such as the Accident Benefits and Bodily Injury, because these methods typically require large data sets for a more in-depth analysis. To compensate for this lack of data and for the longer tailed nature of these two coverages, we use tail factors obtained from

the simulations performed with the activation patterns model. This allows for a more suitable comparison with the results of the other two models discussed in this section.

At the portfolio level, the aggregate model leads to predictions for the total reserves slightly above the observed amount, with a 95% VaR equal to 529.71M CAD. If instead of applying the model to the portfolio as is usually done, we implement it to each coverage separately and then sum up the predictions obtained, we reach an average portfolio reserve and a 95% VaR equal to, respectively, 449.28M CAD and 469.58M CAD. Neither of these values are sufficient to cover the observed reserve. This highlights the disadvantages of using an aggregate model such as the ODP Chain Ladder to a more ganular level of the dataset.

The average and quantile predictions underestimate the observed reserves for all coverages except the Loss of Use coverage where the 95% VaR is slightly above the true amount of 4.20M CAD. The aggregate model provides overall lower predicted values than the activation patterns model and mostly falls short of the observed amounts.

Table 1.7 – Comparison between the activation patterns model and the Overdispersed Poisson Chain Ladder (in M CAD)

Coverage	True reserve	Activation patterns		ODP	
		Mean	$\text{VaR}_{0.95}$	Mean	$\text{VaR}_{0.95}$
Accident Benefits	>158.90	149.61	170.46	148.73	153.41
Bodily Injury	>288.17	283.27	295.02	241.43	253.23
Vehicle Damage	>73.63	70.77	76.43	55.17	58.61
Loss of Use	>4.20	4.17	4.24	3.95	4.33
Global reserve	>524.91	504.82	546.15	515.62	529.71

Activation patterns model vs independence model. We perform independence tests on our data to strengthen our argument in favour of a model that considers the possible dependence between coverages. First, we test for independence between each pair of insurance coverages using the chi-squared goodness of fit test. Each of these two-way interaction tests rejects the hypothesis of independence. We also perform likelihood ratio tests comparing models with no interaction, two, three, and four-way interactions. Once again,

the tests show the presence of dependence between the coverages and reject the simpler models in favour of the more complex ones in each scenario.

We build an independence model to assess the impact of modelling this dependence on the reserve estimates. We reproduce the activation patterns model but rather than modelling the activation patterns in development periods $j = 2$ to $j = 4$ with the multinomial logit models described in Sections 1.3.2 and 1.3.2, we model the activation of the coverages using separate and independent Bernoulli regressions for development periods $j = 2$ to $j = 4$. We fit twelve Bernoulli regressions : one for each of the four insurance coverages and the three development periods considered for RBNS claims. The rest of the model, i.e., the estimation of the payment patterns, payment severities, and all the assumptions linked to them, remain the same. We present the results obtained with this model in Table 1.8 and compare them to the observed reserves and the estimates from our activation patterns model.

Note that since we use the same severity model in the activation patterns and independence models, the average predictions of both models are very close to each other for all insurance coverages and the portfolio. The differences arise in the tails of the predictive distributions where VaRs obtained with the independence model are not always higher than observed amounts, contrary to the ones obtained with the activation patterns model. It is the case for the Loss of Use and, more importantly, Accident Benefits coverages. With an average predicted amount and 95% VaR of, respectively, 507.8M CAD and 533.90M CAD, the independence model provides overall lower estimates for the portfolio Value-at-Risk than the activation patterns model compared to the observed reserve. We also note that predictive distributions for the Accident Benefits coverage, the Bodily Injury coverage, and the whole portfolio are less heavy-tailed than those from the activation patterns model.

Results summary. Table 1.9 summarises the results discussed in Sections 1.4.2 and 1.4.2.

The predictions from a reserving model should be sufficient to cover all future liabilities for incurred but not reported claims and, as in this paper, reported but not settled claims at the chosen valuation date.

With average predictions for the total reserves of 504.82M CAD, 515.62M CAD, and 507.8M CAD for, respectively, the activation patterns, ODP Chain Ladder, and independence models, none of these models provide estimates that are sufficient to cover the observed global portfolio reserve of 524.91M CAD. However, all three models allow us to obtain 95% VaRs that seem large enough to cover this amount. When applied to

Table 1.8 – Comparison between the activation patterns model and the independence model (in M CAD)

Coverage	True reserve	Activation patterns		Independence	
		Mean	$\text{VaR}_{0.95}$	Mean	$\text{VaR}_{0.95}$
Accident Benefits	>158.90	149.61	170.46	151.72	158.23
Bodily Injury	>288.17	283.27	295.02	283.03	292.95
Vehicle Damage	>73.63	70.77	76.43	68.99	78.59
Loss of Use	>4.20	4.17	4.24	4.06	4.13
Global reserve	>524.91	504.82	546.15	507.8	533.90

the portfolio, the ODP Chain Ladder provides the lowest global VaR with a value that is just above the observed reserve. The independence model predicts a 95% VaR of 533.90M CAD, and the activation patterns model finally presents the highest value for the quantile with a reserve amount of 546.15M CAD.

When looking at the reserves per coverage, the activation patterns model is the only one that provides Values-at-Risk higher than the observed reserves in every case. The quantiles of the predictive distributions in the case of the independence model fall short of the observed amounts for the Accident Benefits and Loss of Use coverage, and quantiles of the aggregate model underestimate the reserve for all coverages except Vehicle Damage.

1.5 Conclusion

In this paper, we analyze a Canadian automobile dataset in which each claim can impact one or more of four insurance coverages provided by the company, namely Accident Benefits, Bodily Injury, Vehicle Damage, and Loss of Use. We seek to predict the claims reserve for this portfolio while considering the underlying dependence between these coverages. More specifically, we analyze how a single claim can simultaneously activate multiple coverages and the potential impact this can have on the final amount of the reserve.

To do this, we build an individual model based on activation patterns for the insurance coverages to estimate claims reserves. Based on the predictions of the so-called activation patterns of the four coverages in a given development period j, we then predict whether an active coverage will incur a payment and its

Table 1.9 – Model comparison - summary (in M CAD)

Coverage	Observed	Simul.	Act. pat.	ODP	Ind.
Accident Benefits	>158.90	Mean	149.61	148.73	151.72
		$VaR_{0.95}$	170.46	153.41	158.23
Bodily Injury	>288.17	Mean	283.27	241.43	283.03
		$VaR_{0.95}$	295.02	253.23	292.95
Vehicle Damage	>73.63	Mean	70.77	55.17	68.99
		$VaR_{0.95}$	76.43	58.61	78.59
Loss of Use	>4.20	Mean	4.17	3.95	4.06
		$VaR_{0.95}$	4.24	4.33	4.13
Global reserve	>524.91	Mean	504.82	515.62	507.80
		$VaR_{0.95}$	546.15	529.71	533.90

corresponding amount. We capture the dependence between the coverages in the activation patterns with multinomial logit regressions. Even though we tailor our model to our specific study, we present it in the most general way possible such that one can easily extend it to other applications. For instance, the number of insurance coverages, the number of development periods after which the activation patterns remain stable, the length of the development period, or the severity distributions are all features of the model that one can adapt to other situations. Further developments of this model could include the addition of the age of the claim as a risk factor to account for time dependence. We could also drop the assumption of independence between the claims and seek to model it in a similar way as the dependence between coverages.

We then compare the results obtained for our dataset with those from two additional models : the classical Overdispersed Poisson Chain Ladder and a replica of the activation patterns model that does not consider dependence. Specifically, this replica, named the *independence model* in the paper, predicts the activation patterns of the four coverages using four independent Bernoulli regressions per period rather than a multinomial logit model. It guarantees that the activation of a coverage does not impact the activation of the other coverages in a given period.

We observe in the results that, while all three models provide predictions for the 95% portfolio VaRs of the global reserve larger than the observed amount, only the predicted values from the activation patterns model appear to be above the observed amounts as well for each coverage taken separately. While the other methods also allow us to obtain large enough predictions for the total reserves, they seem less efficient at the coverage level.

With the activation patterns model, we aim to better highlight and understand the underlying dynamics of the claims portfolio at hand through the activation of the different insurance coverages. It is a good illustration of the challenge faced nowadays by many practitioners in the insurance industry that are required, either by regulators or their own internal processes, to predict their claims reserves with increasing levels of granularity.

CHAPITRE 2

SIMULATIONS OF ARCHIMEDEAN COPULAS FROM THEIR NON-PARAMETRIC GENERATORS FOR LOSS RESERVING UNDER FLEXIBLE CENSORING

Résumé

Les assureurs bénéficiant de quantités de données de plus en plus larges et de complexité croissante, nous explorons dans cet article une méthode axée sur ces données pour modéliser la dépendance au sein de réclamations multivariées. Plus précisément, nous étendons l'estimateur non paramétrique pour les fonctions génératrices de copules archimédiennes, ainsi que la procédure de sélection de copules graphiques introduits par Genest et Rivest (1993), à des scénarios de censure flexibles, en utilisant des techniques dérivées de l'analyse de survie. Nous proposons ensuite une méthode alternative permettant de se passer de toute hypothèse paramétrique en simulant directement de nouvelles données à partir de l'estimateur de la fonction génératrice. Nous utilisons à cet effet des algorithmes basés sur les transformées de Laplace-Stieltjes inverses. Nous illustrons la performance des deux approches par le biais de plusieurs études de simulation et les comparons avec une application à un jeu de données récent d'assurance automobile d'une compagnie canadienne pour lequel nous modélisons la dépendance entre les délais d'activation des couvertures d'asssurance dépendante. Nous montrons que, bien que les deux approches fonctionnent dans divers scénarios de censure et pour de grandes quantités de données, simuler de nouvelles observations directement à partir du modèle non paramétrique peut conduire à des prédictions plus précises qui peuvent ensuite être utilisées dans le cadre plus large du calcul de réserves.

Abstract

With insurers benefiting from ever-larger amounts of data of increasing complexity, we explore a data-driven method to model dependence within multilevel claims in this paper. More specifically, we extend the non-parametric estimator for Archimedean copula generators and graphical copula selection procedure introduced by Genest and Rivest (1993), to flexible censoring scenarios, using techniques derived from survival analysis. We then propose an alternative method that allows to forego any parametric assumption by directly simulating from the estimator of the generator function, using algorithms based on inverse Laplace-Stieltjes transforms. We illustrate the performance of both approaches with multiple simulation studies and

compare them with an application to a recent Canadian automobile insurance dataset for which we model the dependence between the activation delays of correlated coverages. We show that although both approaches work well under various censoring scenarios and for large amounts of data, performing simulations directly from the non-parametric model can lead to more accurate predictions that can then be used as part of a larger claims reserving framework.

2.1 Introduction

With the increasing ease in data collection and greater availability of computational resources for financial institutions, actuarial practitioners and academics face everyday larger quantities of complex data. All this data grants insurers access to detailed information about their insureds and the claims that are typically collected over long periods of time, allowing for a better understanding of their portfolio and, among other, more accurate loss modelling.

This increase in quantity, however, also leads to an increase in data complexity. Actuaries must sort through large datasets to retrieve relevant information and derive predictions for premiums or the development of claims, using data that is more often than before unstructured, missing or untimely. One particular complexity frequently encountered in insurance is the issue of incomplete or censored data, arising when part of the information is missing for a specific observation. This situation can arise, for example, in a model for coupled lifetimes when death occurs for one of the individuals in the study, as illustrated in (Deresa *et al.*, 2022). Claims severities can be censored as well if the insurer benefits from a reinsurance treaty that covers a predefined portion of the losses. We can also consider the case where a given prescription period on the payments made by the insurer has passed, thereby stopping further payments towards the claimant even if the claim is still open. Multiple factors can also cause censoring. In the example with the coupled lifetimes, observations can be censored either by the death of one individual or if they simply revoke their contract with the insurance company. Insurers can thus observe a multitude of censoring scenarios in which one or more variables can be censored or where multiple factors can cause censoring.

Additionally, a second complexity often encountered in insurance data is the correlation between different - potentially censored - variables of interest to the insurers. This dependence can take the form, for example, of temporal dependence between the occurrence of a specific event and a claim's development delay as in (Zhou et Zhao, 2010), (Shi *et al.*, 2016) and (Shi et Yang, 2018). (Frees et Valdez, 2008), (Frees *et al.*, 2009) or (Yang et Shi, 2019) also illustrate hierarchical dependence within multi-features claims. We can even find

cross-sectional dependence between losses and expenses, as in (Frees et Valdez, 1998), or between a claim's lifetime and its final amount, as in (Lopez, 2019).

To account for this, many authors in the actuarial literature have used the framework of copulas, i.e., multivariate distribution functions whose marginals are uniformly distributed on the unit interval. Copulas allow a representation of the joint distribution function of multivariate data by separating their marginal distributions from the inherent dependence structure. Examples of their application abound in the statistical literature, and since their introduction by (Sklar, 1959), many authors have extensively studied and analyzed their properties.

Among the different classes of copulas, (Schweizer et Sklar, 1983) introduced the Archimedean family, widely applied today in many fields beyond insurance. Contrarily to other copula families, Archimedean copulas are not derived from Sklar's theorem, as highlighted in (Größer et Okhrin, 2021). They are characterized instead by a function called the *generator*,$\psi(.) : [0, \infty] \to [0, 1]$ with $\psi(0) = 1$ and $\lim_{\nu\to\infty} \psi(\nu) = 0$, defined with one or more so-called dependence parameters. The $d-$variate Archimedean copula $C(.)$ then takes the following form :

$$C(u_1, ..., u_d) = \psi(\psi^{-1}(u_1) + ... + \psi^{-1}(u_d)).$$

Given a copula $C(.)$, we can retrieve Kendall's tau using the relation

$$\tau = 4 \int_0^1 ... \int_0^1 C(u_1, ..., u_d) dC(u_1, ..., u_d) - 1.$$

It is clear that finding an estimator for the generator allows us to directly capture both the shape and strength of dependence for a given set of multivariate observations.

For this purpose, (Genest et Rivest, 1993) propose a non-parametric approach relying on a study of the bivariate probability integral transformation for Archimedean copulas. Their estimator is, however, only applicable to complete data, seldom encountered in insurance. In addition, this estimator has only been used so far to identify the Archimedean copula best fitted to a given dataset but has not been directly used itself in, for example, simulation studies.

The motivation of our paper is thus two-fold. First, considering the large amounts of data available to insurers and the highly diverse censoring schemes that these datasets can contain, we present a graphical selection procedure for Archimedean copulas based on a non-parametric estimator of the generator. The

contribution that we bring to this graphical approach is that we extend it to different flexible censoring scenarios for the data, thereby rendering it well suited to, among others, insurance applications. For this purpose, we base ourselves on the method proposed in (Genest et Rivest, 1993) and modify it, using results from (Wang et Wells, 2000) and an estimator for the joint distribution proposed by (Akritas et Van Keilegom, 2003) in survival analysis in order to account for various censoring schemes. To the best of our knowledge, this estimator has never been used in an actuarial context nor in the application of a non-parametric estimator for Archimedean copulas. In addition, it contrasts with the more naive copula modelling approach that usually consists of fitting a certain number of copulas and comparing their performance using information criteria such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). When using such criteria, we are required to fit all the candidate copula models to the data as a first step, including estimating the dependence parameters. When estimating the generator function, no prior fitting of parametric models is required, and the dependence parameters for various candidate models can be directly derived from the generator, even in the presence of censoring, as we will illustrate in this paper.

Second, we go a step further and, instead of finding the most fitted Archimedean copula, we show how to directly perform simulations from the non-parametric generator. This novel approach allows to use the Archimedean family without assigning a specific copula to the data and thereby limiting ourselves to a named and pre-defined model. This method readily applies to the large amounts of data faced by insurers and frees them from the constraint of defining a parametric model that might miss some of the finer complexities of their datasets. In addition, we show how, in a claims reserving context, performing analysis and predictions directly from the non-parametric estimator of the generator function can lead to more accurate reserves calculations and thereby financial gains for the insurers. Without the constraints of a parametric model that, even though it might be well suited to the data, is unlikely to be a perfect fit, we decrease the risk of prediction inaccuracies arising from model misspecification.

We structure our paper as follows : first, we provide a brief literature review in Section 2.2. In Section 2.3, we describe the statistical model for the non-parametric estimator using methods borrowed from survival theory. Section 2.4 describes the graphical comparison procedure and presents three model validation approaches applicable to different censoring scenarios. In Section 2.5, we show how to directly generate correlated observations starting from the non-parametric estimator, without formulating any parametric model assumptions. We present an application to a claims reserving framework and a comparison between the graphical and simulation approaches in Section 2.6, using a Canadian automobile dataset in which po-

licyholders benefit from multiple insurance coverages. This dataset provides a good illustration of a more complex censoring scheme. We conclude our work in Section 2.7.

2.2 Literature review

In recent years, a lot of attention has been paid to Archimedean copulas, particularly in finance and insurance. Allowing for very flexible dependence structures, this class of copulas has become popular thanks to their simple form and statistical tractability, among other interesting properties. Notably, many of them allow high-dimensional dependence modeling using only one parameter governing the strength of dependence. Although initially more widely used in life insurance where they arise from frailty models, Archimedean copulas have grown popular in non-life insurance as well.

Thanks to the large quantities of data more easily available, benefits can be reaped from using a non-parametric estimator such as the one proposed by (Genest et Rivest, 1993) to directly retrieve the generator of Archimedean copula models. We discuss two interesting examples of the application of this estimator in the case of censored data. First, the work of (Denuit *et al.*, 2006) combines an estimator for the generator from (Wang et Wells, 2000) to an estimator of the joint distribution proposed by (Akritas, 1994) to model the dependence between the allocated loss adjustment expenses and right-censored losses. Their methodology is, however, only applicable to cases where only one variable is subject to censoring, thereby limiting possible replications to more flexible scenarios.

To the best of our knowledge, the only non-parametric approach for more flexible censoring scenarios is the work proposed in (Gribkova et Lopez, 2015). In their paper, the authors propose a discrete estimator and two smooth estimators for Archimedean copulas and any type of copula applicable to various censoring scenarios. Their discrete estimator extends the empirical copula estimator proposed in (Deheuvels, 1997) by using random weights in the empirical joint distribution to account for the bias caused by censored observations. For the two smooth copula estimators, (Gribkova et Lopez, 2015) introduce kernel functions in addition to the weights of the empirical joint distribution estimator.

In the present paper, we use an estimator situated between those of (Denuit *et al.*, 2006) and (Gribkova et Lopez, 2015). As in (Denuit *et al.*, 2006), we focus only on Archimedean copulas and use a non-parametric estimator specifically for Archimedean generators. We, however, seek to allow for the same level of flexibility in the choice of censoring scenarios as (Gribkova et Lopez, 2015). As such, we extend the work of (Denuit

et al., 2006) to the case where any or all variables can be subject to censoring by replacing the estimator for the joint distribution with that proposed in (Akritas et Van Keilegom, 2003).

We obtain a non-parametric estimator for Archimedean copula generators that we then propose to use in two different approaches. First, in a graphical comparison similar to the one used in (Denuit *et al.*, 2006) to select the most appropriate parametric copula model for the data. Second, in a simulation routine to generate correlated observations with the same strength and shape of dependence as the original data.

In the first approach, since the choice of the copula model is driven solely by a graphical comparison, we want to use robust approaches to validate it. Goodness-of-fit tests for copulas have been discussed by some authors in the literature, notably (Genest *et al.*, 2009) who present a review of available tests or (Größer et Okhrin, 2021) who account for all the latest developments in copula theory, including estimation and testing methods.

However, all these tests only apply to complete data. The literature on goodness-of-fit tests for copulas with censored data is much scarcer. For Archimedean copulas, (Lakhal-Chaieb, 2010) proposes a test based on a Cramer-von-Mises type distance and Kendall distribution. (Wang, 2010) presents a test based on a multiple-imputation procedure of the censored data, and (Emura *et al.*, 2010) use a cross-ratio function. Some authors have also proposed tests for other families of copulas, such as the likedlihood and pseudo-likelihoog tests of (Yilmaz et Lawless, 2011). More recently, (Zhou, 2021) and (Sun *et al.*, 2023) presented general tests based on information matrices applicable to various copula families. In addition to their three non-parametric copula estimators, (Gribkova et Lopez, 2015) also propose a test similar to that of (Lakhal-Chaieb, 2010) that can be used with different kinds of distances such as Kolmogorov-Smirnov, Cramer-von-Mises or the square-root of a quadratic integrated distance.

In this paper, we use three approaches to validate the non-parametric estimation of the Archimedean copula generator. More specifically, we use an omnibus procedure based on the likelihood function, a method based on the L^2-norm distance, inspired by (Lakhal-Chaieb, 2010) and (Gribkova et Lopez, 2015), and the test proposed by (Wang, 2010).

Our second approach, namely direct simulations from the non-parametric estimator of the generator function, is inspired by the work of (Hofert, 2008) who described various sampling schemes for nested Archimedean copulas. In the various algorithms presented in his work, the author however assumes a known

Archimedean copula family with a well-defined parametric generator, or generators for which the Laplace-Stieltjes transforms are known. (Ridout, 2009) proposes an algorithm that can generate random observations from a distribution specified by its inverse Laplace-Stieltjes transform. In this work, we adapt (Ridout, 2009)'s algorithm to the case where the Laplace-Stieltjes transform is defined by a non-parametric function, then use this combined with some algorithms presented by (Hofert, 2008) to sample from our original data, without specifying a parametric copula model.

2.3 Statistical Model

In this section, we describe an approach to retrieve the Archimedean copula generator $\psi(.)$ in the case of flexible censoring. More specifically, we present a non-parametric estimator building on the work of (Akritas et Van Keilegom, 2003) in survival analysis. The advantage of using such a non-parametric estimator lies in the possibility of directly tailoring the model to the data, even in the presence of complex censoring.

We begin by defining the notation before describing the estimator. For clarity, we focus solely on the bivariate case, but the notation and model described hereafter can easily be extended to a multivariate framework.

2.3.1 Notation

We define the following model components :

- $\boldsymbol{T} = (T_1, T_2)$ with $T_i \in \mathbb{R}^+$ for $i = 1, 2$ is the initial vector of variables of interest. They can represent times-to-event, losses, etc. To avoid identifiability issues with the copula, we assume these variables to be continuous ;
- $\boldsymbol{X} = (X_1, X_2)$ with $X_i \in \mathbb{R}^+$ for $i = 1, 2$ is the vector of censoring variables, such that $\boldsymbol{T}$ and $\boldsymbol{X}$ are independent ;
- $\boldsymbol{\omega} = (\omega_1, \omega_2)$, with $\omega_i > 0$ for $i = 1, 2$ are constants acting as additional censoring factors on $\boldsymbol{T}$. We call these the *limits*.
- $\boldsymbol{Y} = (Y_1, Y_2)$ where $Y_i = \min(T_i, X_i, \omega_i)$ for $i = 1, 2$ is the observed bivariate vector, taking censoring of $\boldsymbol{T}$ into account ;
- $\boldsymbol{\Delta} = (\mathbb{1}_{[Y_1=T_1]}, \mathbb{1}_{[Y_2=T_2]})$ is the vector of censoring indicators.

To illustrate the difference between the censoring variable X and the limit ω, consider a study of the time to recovery after administration of a treatment of patients suffering from a medical condition. The time to

recovery, denoted by T, is censored if the patient dies, with the time to death being the censoring variable X. In addition, health professionals might consider the treatment ineffective if the patient does not recover after a given amount of time, even if he is still alive. Further treatments will then need to be considered. In this example, the time elapsed until the treatment is deemed to have failed is the limit ω. Even if the patient is still alive, i.e., $T > X$, the imposition of ω will censor his time to recovery. The observed delay $Y = \min(T, X, \omega)$ is of interest to medical practitioners. Y represents the delay until the first event, being either full recovery, death, or need for additional treatments.

Further applications can be found in finance, insurance, and other fields. Section 2.6 focuses on an application to bivariate claims delays.

2.3.2 Non-parametric estimator of the generator

We now describe the non-parametric approach leading to the identification of the copula linking the pairs of censored variables (Y_1, Y_2).

Assuming flexible censoring scenarios where either one or both variables can be subject to random censoring, we follow the methodology laid out in (Akritas et Van Keilegom, 2003) and work with the following extension of (Beran, 1981)'s estimators for the conditional distributions of Y_1 (resp. Y_2) given $Y_2 = y_2$ (resp. $Y_1 = y_1$) :

$$\hat{F}_{1|2}(y_1|y_2) = 1 - \prod_{Y_{i1} \leq y_1, \Delta_{i1}=1} \left(1 - \frac{W_{ni2}(y_2; h_n)}{\sum_{j=1}^{n} W_{nj2}(y_2; h_n) \mathbb{1}_{Y_{j1} \geq Y_{i1}}}\right), \tag{2.1}$$

where y_2 must be uncensored and

$$W_{ni}(y; h_n) = \begin{cases} \dfrac{k\left(\frac{y - Y_{i2}}{h_n}\right)}{\sum_{\Delta_{j1}=1} k\left(\frac{y - Y_{j2}}{h_n}\right)}, & \text{if } \Delta_{i2} = 1 \\ 0, & \text{if } \Delta_{i2} = 0. \end{cases}$$

Here, $k(.)$ is a known kernel function and $\{h_n\}$ is the bandwidth : a sequence of positive constants such that $h_n \to 0$ as $n \to \infty$. We work with a similar estimator for $\mathrm{P}[Y_2 \leq y_2 | Y_1 = y_1] = \hat{F}_{2|1}(y_2|y_1)$.

The estimator of the joint distribution proposed in this context by (Akritas et Van Keilegom, 2003) is then given by :

$$\hat{F}(\boldsymbol{y}) = w(\boldsymbol{y}) \int_0^{y_2} \hat{F}_{1|2}(y_1|z_2) d\tilde{F}_2(z_2) + (1 - w(\boldsymbol{y})) \int_0^{y_1} \hat{F}_{2|1}(y_2|z_1) d\tilde{F}_1(z_1), \tag{2.2}$$

where $\tilde{F}_1$ and $\tilde{F}_2$ are the marginal estimators of Kaplan and Meier (1958) :

$$\tilde{F}_j(y_i) = 1 - \prod_{Y_{i,j} \leq y_i, \Delta_{ij}=1} \Big(1 - \frac{1}{n-i+1}\Big), \quad j = 1, 2.$$

The weights $w(\boldsymbol{y})$ minimise the mean-squared error of $\hat{F}(\boldsymbol{y})$. Equation (2.2) is the estimator that, combined with (Genest et Rivest, 1993)'s approach, allows us to retrieve the generator of the copulas in the presence of different censoring scenarios. We start from (Genest et Rivest, 1993)' proposition laid out just below.

Proposition 2.1 ((Genest et Rivest, 1993)) *Let $S_1(t_1)$ and $S_2(t_2)$ be marginal survival functions whose dependence function $C(.)$ is of the form*

$$C(S_1(t_1), S_2(t_2)) = \psi\Big\{\psi^{-1}(S_1(t_1)) + \psi^{-1}(S_2(t_2))\Big\},$$

with $\psi(.) : [0, \infty] \to [0, 1]$ with $\psi(0) = 1$ and $\lim_{\nu\to\infty} \psi(\nu) = 0$. Let

$$U = \frac{\psi^{-1}(S_1(t_1))}{\psi^{-1}(S_1(t_1)) + \psi^{-1}(S_2(t_2))} \quad \textit{and} \quad V = \psi\Big\{\psi^{-1}(S_1(t_1)) + \psi^{-1}(S_2(t_2))\Big\} = C(S_1(t_1), S_2(t_2)).$$

Then,

1. *U is uniformly distributed on $(0, 1)$;*
2. *V follows the Kendall distribution defined as $K(\nu) = \nu - \frac{\psi^{-1}(\nu)}{[\psi^{-1}]^{(1)}(\nu)}$, for $0 < \nu \leq 1$; and*
3. *U and V are independent.*

(Genest et Rivest, 1993) suggest the following estimator for the Kendall distribution in the presence of complete data :

$$\hat{K}_n(\nu) = \frac{1}{n} \#\{i | \nu_i \leq \nu\},$$

where $\#$ denotes the cardinality of a set.

We cannot use this estimator in the presence of censored data. Instead, we choose to apply the extension proposed by (Wang et Wells, 2000) to estimate the Kendall distribution by

$$\hat{K}_n(\nu) = \int_0^\infty \int_0^\infty \mathbb{1}_{[\hat{F}(\boldsymbol{y}) \leq \nu]} \hat{F}(d\boldsymbol{y}), \tag{2.3}$$

in which we insert (2.2) as the estimator of the joint distribution, allowing for all possible censoring scenarios.

Finally, we insert $\hat{K}_n(\nu)$ in the estimator for the Archimedean generator proposed by (Genest et Rivest, 1993) :

$$\hat{\psi}_n^{-1}(\nu) = \exp\left\{\int_{\nu_0}^{\nu} \frac{1}{t - \hat{K}_n(t)} dt\right\}, \tag{2.4}$$

with $0 < \nu_0 < 1$ an arbitrarily chosen constant. Thanks to the use of Equation (2.2), we can retrieve the generator non-parametrically, regardless of the level of censoring present in the data.

In Section 2.4 and 2.5, we now present two different methods to use $\hat{\psi}_n(\nu)$ to perform further analysis and predictions.

2.4 Graphical comparison

Given the empirical estimator $\hat{K}_n(\nu)$, we can select the parametric copula model best fitted to the data at hand using the graphical procedure introduced by (Genest et Rivest, 1993). Let the univariate function $\lambda(\nu)$ be defined from the Kendall's distribution as

$$\lambda(\nu) = \nu - K(\nu) = \frac{\psi^{-1}(\nu)}{[\psi^{-1}]^{(1)}(\nu)}.$$

An estimator of $\lambda(\nu)$ is easily retrieved by

$$\hat{\lambda}_n(\nu) = \nu - \hat{K}_n(\nu). \tag{2.5}$$

The idea is now to graphically compare the empirical estimator $\hat{\lambda}_n(\nu)$ to the corresponding $\lambda_{\hat{\alpha}}(\nu)$ of different competing copula models under consideration, where $\hat{\alpha}$ is the dependence parameter of the copula. Knowing $\hat{\lambda}_n(\nu)$, we proceed as described below :

1. Estimate Kendall's tau using the following equality proposed by (Genest et Rivest, 1993) :

$$\hat{\tau} = 4\int_0^1 \hat{\lambda}_n(\nu)d\nu = 3 - 4\int_0^1 \hat{K}_n(\nu)d\nu \tag{2.6}$$

2. Given the relationship between Kendall's tau and the dependence parameter α for some Archimedean copulas, retrieve these copula parameters using

$$\hat{\alpha} = g^{-1}(\hat{\tau}). \tag{2.7}$$

Table B.1 in B.1 provides the functions $g(.)$ as well as other specific information for some popular Archimedean copulas.

3. Given $\hat{\alpha}$, trace the curves of $\lambda_{\hat{\alpha}}(\nu)$ for some competing copula models, using again the information provided in B.1.
4. Select the parametric copula model m for which the curve of $\lambda_{\hat{\alpha}_m}(\nu)$ is closest to that of $\hat{\lambda}_n(\nu)$.

One important benefit of this graphical procedure is that, contrarily to the more classical model selection using information criteria, it only requires to estimate $\hat{\lambda}_n(\nu)$. From this estimator, we can directly evaluate on the graph what type of Archimedean copula might best fit our data by observing the shape of the $\hat{\lambda}_n(\nu)$ curve. We also easily retrieve the copula parameter using Equation (2.7). There is no need to fit different copula models beforehand and compare them using AIC and BIC. Estimating $\hat{\lambda}_n(\nu)$ is enough to identify the most appropriate model, even in the presence of incomplete data.

2.4.1 Results validation

While practical and easy to implement, the estimator described in Section 2.3.2 is non-parametric, and the choice of the copula model is driven at this stage by a graphical analysis only. Goodness-of-fit tests for copulas are rare in the literature, even more so in the presence of censoring. To validate our graphical choice and increase the level of confidence in our selected model, we briefly describe three approaches in this section. More details on each of them can be found in B.3.

A first way to verify whether the selected copula model is appropriate for the data at hand is to compare the dependence parameter $\hat{\alpha}$ estimated from Equation (2.7) to its equivalent $\hat{\alpha}^*$, obtained from an optimisation procedure. This is the omnibus procedure, or maximum pseudo-likelihood procedure that consists in a semi-parametric optimization that substitutes empirical versions of the marginal distributions in the (parametric) likelihood function of the copula model, $L(.)$. This procedure has been used, for example, in (Genest *et al.*, 1995) and (Denuit *et al.*, 2006).

The second approach is based on the L^2-norm. Following the original idea of (Wang et Wells, 2000), we verify that the distance between the non-parametric estimator of the Kendall distribution $\hat{K}_n(\nu)$ and $K_{\hat{\alpha}_m}(\nu)$ for the selected model m is the smallest in comparison with the distances to the Kendall distributions of all other candidate models. We propose the following parametric boostrap procedure, inspired by (Lakhal-Chaieb, 2010) and (Gribkova et Lopez, 2015) to select the most appropriate copula model based on a pseudo $p-$value :

Step 1. Compute $\hat{K}_n(\nu)$ for the data at hand and, for the M candidate models under consideration, estimate $\hat{\alpha}_m$, for $m = 1, ..., M$.

Step 2. For each candidate model $C_{\hat{\alpha}_m}$, generate B censored samples of size n.

Step 3. For each of the B samples, estimate $\hat{\alpha}_{m,b}$ and $D(\hat{\alpha}_{m,b}) = \sum_{i=1}^{n} \left(\hat{K}_n(\nu_{(i)}) - K_{\hat{\alpha}_{m,b}}(\nu_{(i)})\right)^2 (\nu_{(i)} - \nu_{(i-1)})$.

Step 4. For each model m, obtain the pseudo p-value as

$$p_m = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}[\min_{l} D(\hat{\alpha}_{b,l}) > D(\hat{\alpha}_{b,m})]$$

with $m \neq l$. The copula model m best fitted to the data is the one for which the pseudo p-value is the smallest. In this case, we can interpret p_m as the number of times among B replications that copula model m is selected as most fitted for the data at hand.

The third and final approach to validate the chosen copula model is a more robust statistical test proposed by (Wang, 2010). This goodness-of-fit test bases its premise on Proposition 2.1 from (Genest et Rivest, 1993) : under the null hypothesis that an Archimedean copula can model the dependence between the pairs of observations (T_1, T_2) with generator $\psi(.)$, the correlation coefficient between the variables U and V is null. (Wang, 2010) describes the test in details and extends it to censored data using a multiple imputation procedure.

2.4.2 Simulation study

Using simulation studies, we now illustrate the graphical approach described in this section. We provide a few examples to showcase the graphical model selection procedure and the validation approaches discussed in Section 2.4.1.

Independence case. The first way to validate the results from the estimator for the generator in Section 2.3 is to analyze how it handles independent data. Setting Kendall's tau equal to 0, each of the four copulas should converge to the independence copula. This is the case if $\alpha = 0$ for both the Clayton and Frank copulas and $\alpha = 1$ for the Gumbel and Joe copulas.

We simulate samples of size $n = 1000$ from each of the four copulas with $\tau = 0$. We then use the methodology laid out in Section 2.3 to estimate $\hat{K}_n(\nu)$, $\hat{\tau}$ and, finally, $\hat{\alpha}$ for each copula. Figure 2.1 presents the graphical comparison between the average of the 1000 non-parametric curves $\hat{\lambda}_n(\nu)$ and the correspon-

ding $\lambda_{\hat{\alpha}(\nu)}$ for four candidate models, namely the Clayton, Frank, Gumbel and Joe copulas. We also present the average estimates of $\hat{\alpha}$ over the 1000 simulations in Table 2.1.

Figure 2.1 shows that the curves for the four candidate parametric models as well as the non-parametric curve, all converge towards the same curve on the plot, i.e. the curve of the independence copula, and it is hard to differentiate between them. Table 2.1 confirms these results by showing that the parameter estimates of the Clayton and Frank copula tend to 0, while those of the Gumbel and Joe copulas tend to 1. Our approach using (Akritas et Van Keilegom, 2003)'s estimator for the joint distribution provides adequate estimates of the dependence parameters when $\tau = 0$.

Table 2.1 – Average estimates of $\hat{\alpha}$ for 1000 simulations when $\tau = 0$.

	Clayton	Frank	Gumbel	Joe
$\hat{\alpha}$	0.0249	0.0352	1.0163	1.0150

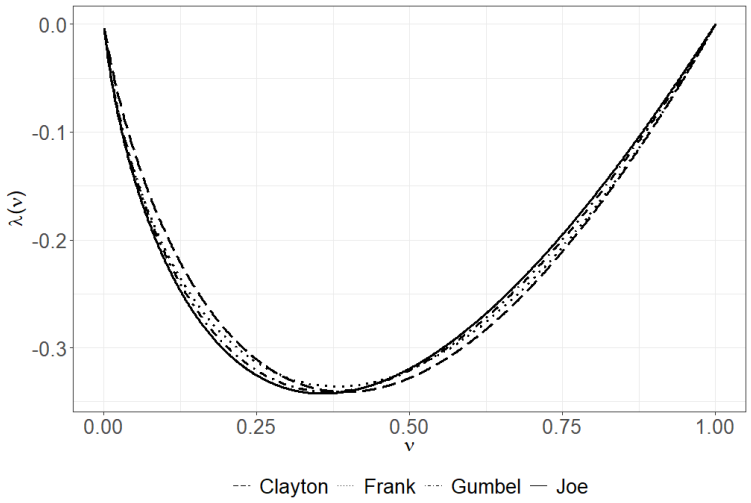

Figure 2.1 – Graphical comparison of the estimated $\hat{\lambda}(\nu)$ functions for the independent samples.

Graphical comparison. We now move away from the independence scenario and focus on the non-parametric estimator of $\hat{\lambda}_n(\nu)$ used for the graphical comparison.

We simulate bivariate samples (T_1, T_2) of $n = 1000$ observations from the Clayton, Gumbel, Frank, and Joe copulas. In each case, we select the dependence parameter such that Kendall's tau equals 0.4, and we work with unit-exponential distributions for the marginals. We simulate the vector of censoring variables

(X_1, X_2) from exponential distributions with their parameters set such that 20% of observations have at least one censored component. We are in a double-censoring scenario where one or both variables can be censored.

Figure 2.2 displays the empirical estimator $\hat{\lambda}_n(\nu)$ and the curves of $\lambda_{\hat{\alpha}}(\nu)$ for the four copulas under consideration, for each of the four simulated samples. In each plot, the continuous curve depicts the non-parametric estimator presented in Equation (2.5). For all, $\hat{\lambda}_n(\nu)$ appears to be closest to the curve of the correct copula, i.e., the copula from which the sample was simulated. These graphical comparisons suggest that the graphical approach proposed in this section performs quite well under flexible censoring scenarios.

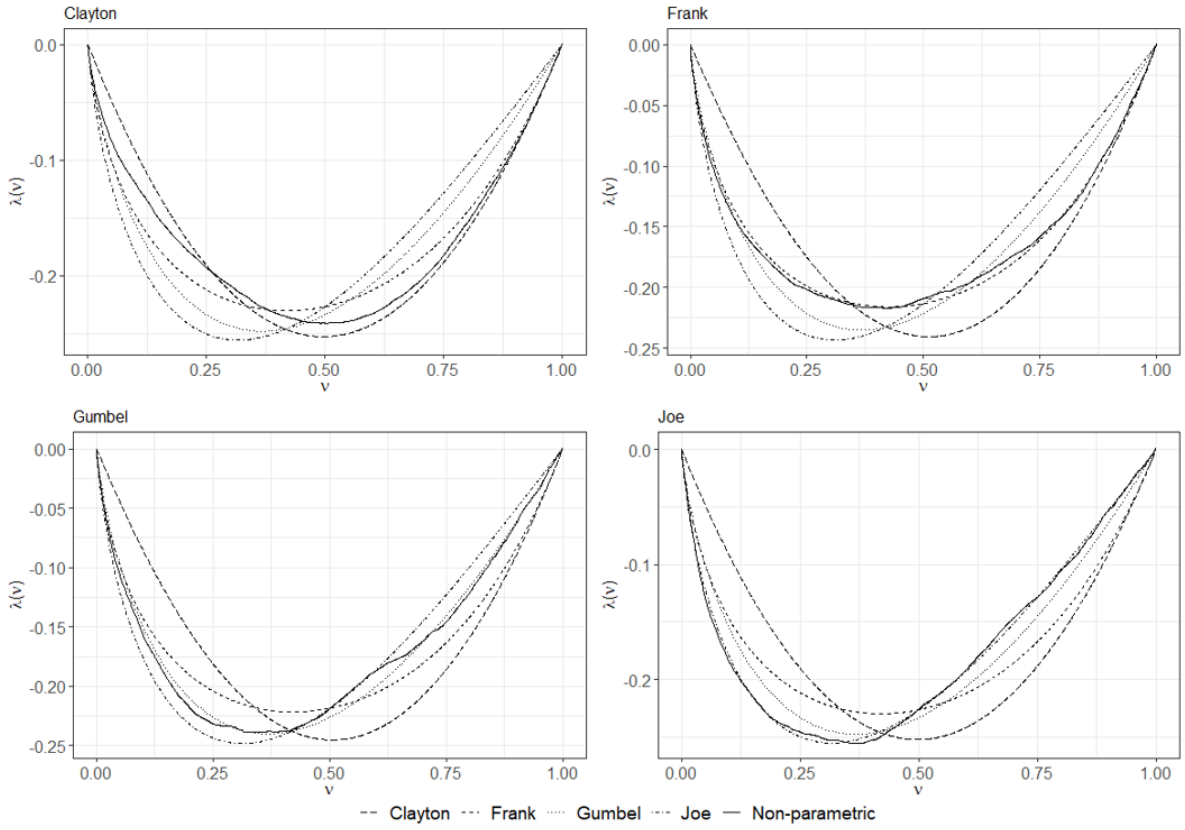

Figure 2.2 – Graphical comparison of the estimated $\hat{\lambda}(\nu)$ functions for the four simulated samples. We simulate $n = 1000$ bivariate observations from a Clayton (top left), Frank (top right), Gumbel (bottom left) and Joe (bottom right) copulas.

Omnibus procedure. We begin validating the results from Figure 2.2 with the omnibus procedure. In addition to the double-censoring scenario considered here, we add two other censoring schemes : one with

complete data (no censoring) and one where we allow for only one censored variable. We work again for these two additional scenarios with simulated samples of size $n = 1000$, Kendall's tau equal to 0.4, and unit-exponential margins. For the single-censoring scenario, we simulate the censoring variable following an exponential distribution such that around 20% of observations have a censored component, similar to the double-censoring scheme.

Table 2.2 shows the results of performing the omnibus comparison between $\hat{\alpha}$ and $\hat{\alpha}^*$ one thousand times. Each value in this table represents the percentage of simulations in which the competing models were not considered best-fitted for the sample in terms of the difference between the estimated dependence parameters. We observe, for example, that when we simulate a sample from the Clayton copula with $\tau = 0.4$ in a double-censoring scenario, only 276 simulations out of a thousand reject the Clayton model as best-fitted. Among these 276 simulations, 188 opt for a Joe model, 71 for a Gumbel model, and 17 for a Frank model.

Table B.3 in B.4 shows the results of one of these simulations when $\tau = 0.4$ by comparing the dependence parameters $\hat{\alpha}$ estimated with Equation 2.7 to the corresponding $\hat{\alpha}^*$ for the omnibus procedure.

The results displayed in both Table B.3 and B.4 illustrate that the omnibus procedure is a good first indicator of the appropriateness of the copula model selected via the graphical procedure.

Table 2.2 – Percentage of simulations in which different candidate copula models are rejected with the omnibus procedure.

Scenario	True copula	Candidate model			
		Clayton	Frank	Gumbel	Joe
No censoring	Frank	0.876	0.322	0.853	0.949
	Joe	0.897	0.680	0.845	0.578
Single-censoring	Frank	0.880	0.274	0.884	0.962
	Joe	0.892	0.844	0.868	0.396
Double-censoring	Clayton	0.276	0.983	0.929	0.822
	Frank	0.865	0.392	0.947	0.796
	Gumbel	0.906	0.992	0.370	0.732
	Joe	0.936	0.942	0.962	0.160

L^2-norm. Next, we perform a second study in which we simulate bivariate samples of $n = 500$ observations from various copulas with Kendall's tau equal to $0.2, 0.4$, and 0.6. We use the same exponential distributions of parameters equal to one for the marginal distributions of (T_1, T_2). We again consider three censoring schemes : one with no censoring, one where we allow only one variable to be censored, and one where both variables are subject to censoring. For the double-censoring scheme, we also analyze the independence case.

When working with complete data, we use (Genest et Rivest, 1993)'s estimator $\hat{K}_n(\nu)$ to compute $D(\hat{\alpha})$:

$$\hat{K}_n(\nu) = \frac{1}{n}\#\{i|\nu_i \leq \nu\},$$

with

$$\nu_i = \frac{1}{n-1}\#\{(t_{1,(j)}, t_{2,(j)})|t_{1,(j)} < t_{1,(i)}, t_{2,(j)} < t_{2,(i)}\}.$$

For the single-censoring scenario, we use a (Wang et Wells, 2000)'s estimator $\hat{K}_n(\nu)$ similar to Section 2.3.2, but with a different estimator for the joint distribution $\hat{F}(\boldsymbol{y})$. Namely, we compute the estimator proposed

by (Akritas, 1994) for single-censoring scenarios. Assuming that T_1 is the variable that can be censored by X_1, we have :

$$\begin{aligned}\hat{F}(y_1, t_2) &= \int_0^{t_2} \hat{F}_{Y_1|T_2}(y_1|z) d\hat{F}_{T_2}(z) \\ &= \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}[0 \leq t_{2,k} \leq t_2] \hat{F}_{Y_1|T_2}(y_1|t_{2,k}),\end{aligned}$$

where $\hat{F}_{Y_1|T_2}(y_1|z)$ is (Beran, 1981)'s estimator from Equation (3.4).

For the two censoring scenarios, we simulate the censoring variables using exponential distributions with parameters such that at least 20% of observations have one (or two) censored components. For each sample, we then perform the bootstrap procedure described earlier to obtain estimates of the L^2-norm distance between $\hat{K}_n(\nu)$ estimated non-parametrically for the data and the corresponding $K_{\hat{\alpha}}(\nu)$ for different candidate models.

Table 2.3 presents the results of 1000 simulations for each sample. Each value in this table corresponds to the pseudo p-value from the bootstrap procedure and should be understood as the percentage of bootstrap simulations in which the candidate model was not selected as the best model for the simulated sample. That is the percentage of simulations for which the L^2-norm distance between $\hat{K}_n(\nu)$ and $K_{\hat{\alpha}}(\nu)$ for the candidate model was not the smallest one compared to the other models under consideration. We use samples from both Frank and Joe copulas in the no-censoring and single-censoring schemes. At the same time, in the double-censoring scenario, we analyze samples from the Clayton, Frank, Gumbel, and Joe copulas.

We observe that in all three scenarios for each copula, the pseudo p-value is the smallest for the candidate model corresponding to the true copula. For example, in the double-censoring scenario, when we simulate the data from a Frank copula with Kendall's tau equal to 0.4, we observe that the Frank copula is rejected as an appropriate candidate model in just 1.5% of the bootstrapped simulations performed. The Gumbel model is rejected in 98.5% of simulations, and the Clayton and Joe copulas are rejected in all simulations. In addition, we observe that as the value of Kendall's tau increases, the percentage of rejection for the correct model decreases. For the Frank sample in the double-censoring scenario, 8.7% of the bootstrap simulations reject the Frank model when $\tau = 0.2$. This percentage decreases to 1.5% when $\tau = 0.4$ and to 0.9% when $\tau = 0.6$. We observe similar results and trends in all scenarios for each copula considered. This showcases the strength of our test.

In addition, for the doubly-censored scenario, we simulate samples from each of the four copulas with Kendall's tau equal to 0. This is equivalent to simulating four bivariate samples from the independence copula. We observe that the results of the bootstrap simulations are much more random. For the Joe sample, for example, 45% of simulations reject the Clayton model, all reject the Frank model, 80% reject the Gumbel model, and 75% reject the Joe model. This is to be expected since, when $\tau = 0$, all four copulas converge towards the same independence copula. The estimates that we thereby get for $K_{\hat{\alpha}}(\nu)$ and $\lambda_{\hat{\alpha}}(\nu)$ all converge towards the same functions. The choice of the smallest distance between $\hat{K}_n(\nu)$ and $K_{\hat{\alpha}}(\nu)$ for the different candidate models thus becomes much more random. This was already well illustrated in Figure 2.1 that shows the plot of $\hat{\lambda}(\nu)$ and those of $\lambda_{\hat{\alpha}}(\nu)$ for the four candidate models considered—the curves for the Clayton, Frank, Gumbel, and Joe copula overlap.

Table 2.3 – Pseudo p-values for different candidate copula models under different censoring scenarios.

Scenario	True copula	τ	Candidate model			
			Clayton	Frank	Gumbel	Joe
No censoring	Frank	0.2	1.000	0.091	0.909	1.000
		0.4	1.000	0.067	0.948	0.985
		0.6	1.000	0.001	0.999	1.000
	Joe	0.2	1.000	0.999	0.924	0.077
		0.4	1.000	1.000	0.998	0.002
		0.6	1.000	1.000	0.999	0.001
single-censoring	Frank	0.2	1.000	0.078	0.923	0.999
		0.4	1.000	0.054	0.946	1.000
		0.6	1.000	0.012	0.999	0.989
	Joe	0.2	1.000	1.000	0.928	0.072
		0.4	1.000	1.000	0.987	0.013
		0.6	1.000	1.000	0.999	0.001
double-censoring	Clayton	0.0	0.470	0.890	0.790	0.850
		0.2	0.170	0.831	0.999	1.000
		0.4	0.079	0.921	1.000	1.000
		0.6	0.000	1.000	1.000	1.000
	Frank	0.0	0.230	1.000	0.770	1.000
		0.2	0.986	0.087	0.928	0.999
		0.4	1.000	0.015	0.985	1.000
		0.6	1.000	0.009	0.991	1.000
	Gumbel	0.0	0.620	1.000	1.000	0.038
		0.2	1.000	0.970	0.168	0.862
		0.4	1.000	0.975	0.135	0.890
		0.6	1.000	0.986	0.057	0.957
	Joe	0.0	0.450	1.000	0.800	0.750
		0.2	1.000	1.000	0.843	0.157
		0.4	1.000	1.000	0.929	0.071
		0.6	1.000	1.000	0.949	0.051

Goodness-of-fit test. Using the same censoring scenarios, marginal distributions, and copulas as for the L^2-norm with Kendall's tau equal to 0.2, 0.4, and 0.6, we now apply (Wang, 2010)'s goodness-of-fit test described in B.3.3. We present the results of 1000 simulations for simulated samples of size $n = 200$ in Table 2.4. For the single-censoring and double-censoring scenarios, we use $M = 5$ imputed datasets.

We observe similar results to those obtained by (Wang, 2010) and to those from Table 2.3. When dependence is weak, the test does not always accurately predict the best model for the data. In particular, we see that in the single-censoring scenario when the sample comes from the Frank copula with $\tau = 0.2$, even though the lowest percentage of rejection of the null hypothesis is correctly observed when the candidate model is the Frank copula, this percentage is also relatively low for the three alternative models. As already observed in (Wang, 2010), the power of the test, however, increases when we increase the sample size. To illustrate this, we present in Figure 2.3 what happens to the p-values in the no-censoring scenario when the true model is the Frank copula with $\tau = 0.4$ when we progressively increase the sample size from $n = 200$ to $n = 2000$. We work with increasing steps of 200 observations; in each case, we perform 1000 simulations. The strength of the test increases quite steeply with the sample size. For the Frank candidate model, the resulting p-value starts at a low level of approximately 5%, as was reported in Table 2.4, and quickly tends to 0 as we increase the sample size. The Gumbel copula displays a low p-value of less than 10% for the smallest sample size. However, it greatly increases such that for a sample of 2000 observations, it is correctly rejected by over 85% of simulations. The Joe and Clayton candidate models display similar increasing trends, even though they started with a higher p-value for small sample sizes than the Gumbel copula.

Similarly to the results of the L^2-norm in Table 2.3, we also observe in Table 2.4 that the strength of the test increases with the level of dependence. In all scenarios and for all copula models, the percentage of rejection of the null hypothesis increases with τ when the null hypothesis is incorrectly specified and decreases with τ when the true and null copula coincide.

Impact of the limit. In Section 2.3.1, we introduced a limit ω_i for $i = 1, 2$ acting as an additional censoring layer. Depending on the context, these limits can be imposed by practitioners, authorities, etc. They can be the maximum time before treatment is considered ineffective, even if the patient is still alive, or the maximum payment delay authorized for a claim for example.

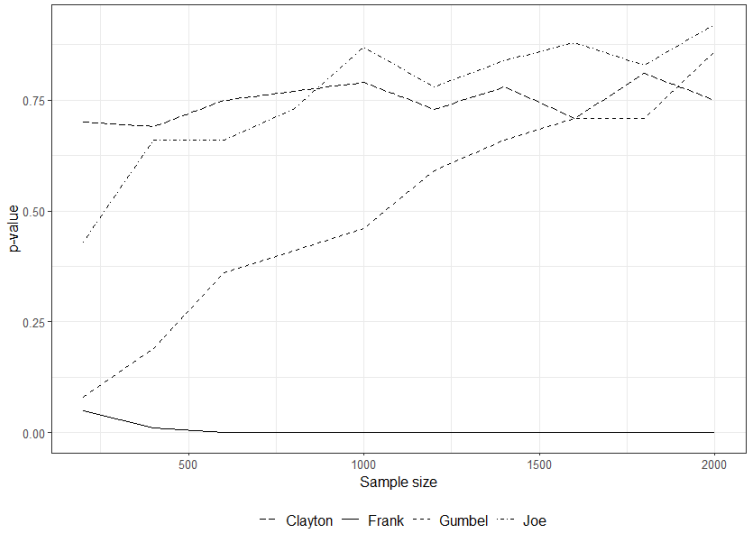

Figure 2.3 – Evolution of the p-values of competing models based on 1000 simulations for varying sample sizes simulated from the Frank copula with Kendall's tau equal to 0.4 in a no censoring scenario.

In this section, we conduct a simulation study to assess the impact of ω on the strength and form of dependence between two censored variables.

Consider a bivariate vector of time-to-events $\boldsymbol{T} = (T_1, T_2)$ with log-normal distributions such that $\mu_{T_1} = 8$, $\mu_{T_2} = 7$, $\sigma_{T_1} = 1$ and $\sigma_{T_2} = 3$. We set the correlation ρ between T_1 and T_2 at 0.35. Let $\boldsymbol{X} = (X_1, X_2)$ be the censoring vector with X_1 independent from X_2. Both X_1 and X_2 are log-normally distributed, with parameters μ_{X_1} and μ_{X_2} chosen to allow for different levels of censorship. We also set $\sigma_{X_1} = 1$ and $\sigma_{X_2} = 1$. We perform three simulation studies : one with low levels of censorship for both variables (approx. 2% and 1% for, respectively, T_1 and T_2), one with medium levels of censoring (approx. 25% and 40%) and one with higher levels of censorship (approx. 65% and 75% for, respectively, T_1 and T_2). Finally, consider the limits ω_1 and ω_2 such that we observe the bivariate vector $\boldsymbol{Y} = (Y_1, Y_2) = \big(\min(T_1, X_1, \omega_1), \min(T_2, X_2, \omega_2)\big)$.

We use the strategy laid out in Section 2.3 to derive the joint distribution (2.2) from Beran's estimators. As in (Akritas et Van Keilegom, 2003), we choose the weights $w(\boldsymbol{y}) = 0.5$ for the sake of simplicity although better results could be obtained by optimizing them. Using the estimated joint distribution, we then derive Kendall's tau and estimate $\hat{\lambda}_n(\nu)$ for a sample of $n = 500$ observations. We use different quantiles from the distributions of Y_1 and Y_2 to set the values of the limits ω_1 and ω_2.

Figure 2.4 displays the results of decreasing the limit as well as the level of censoring on the form of depen-

Table 2.4 – Percentage of rejection of the null hypothesis for different copulas $n = 200$

Scenario	True copula	τ	Copula under H_0			
			Clayton	Frank	Gumbel	Joe
No censoring	Frank	0.2	0.351	0.082	0.030	0.074
		0.4	0.728	0.058	0.086	0.679
		0.6	0.859	0.022	0.196	0.956
	Joe	0.2	0.814	0.206	0.248	0.022
		0.4	0.922	0.776	0.724	0.018
		0.6	0.944	0.970	0.942	0.003
single-censoring	Frank	0.2	0.142	0.032	0.229	0.142
		0.4	0.712	0.084	0.800	1.000
		0.6	0.760	0.000	0.968	0.990
	Joe	0.2	0.720	0.030	0.052	0.016
		0.4	0.970	0.232	0.081	0.007
		0.6	0.980	0.400	0.131	0.004
double-censoring	Clayton	0.2	0.097	0.031	0.512	0.239
		0.4	0.021	0.203	0.835	0.945
		0.6	0.016	0.645	0.968	0.958
	Frank	0.2	0.391	0.022	0.057	0.029
		0.4	0.810	0.012	0.850	0.999
		0.6	0.944	0.001	0.975	1.000
	Gumbel	0.2	0.456	0.082	0.010	0.012
		0.4	0.848	0.172	0.030	0.051
		0.6	0.962	0.458	0.025	0.775
	Joe	0.2	0.990	0.468	0.535	0.081
		0.4	1.000	0.530	0.727	0.002
		0.6	1.000	0.880	0.843	0.000

dence via the function $\hat{\lambda}_n(\nu)$ and the strength of dependence via Kendall's tau. We observe that neither the limit nor the level of censoring seem to have an impact on the form of dependence : in all cases, the shape

of the generator function remains similar. As we increase the level of censoring and the limit, we notice that the shape of dependence becomes even more pronounced. This is in line with the results from Tables 2.3 and 2.4 where we noted that, even though both tests were able to correctly choose the most appropriate copula for a given sample and low dependence levels, their strength increased with Kendall's tau.

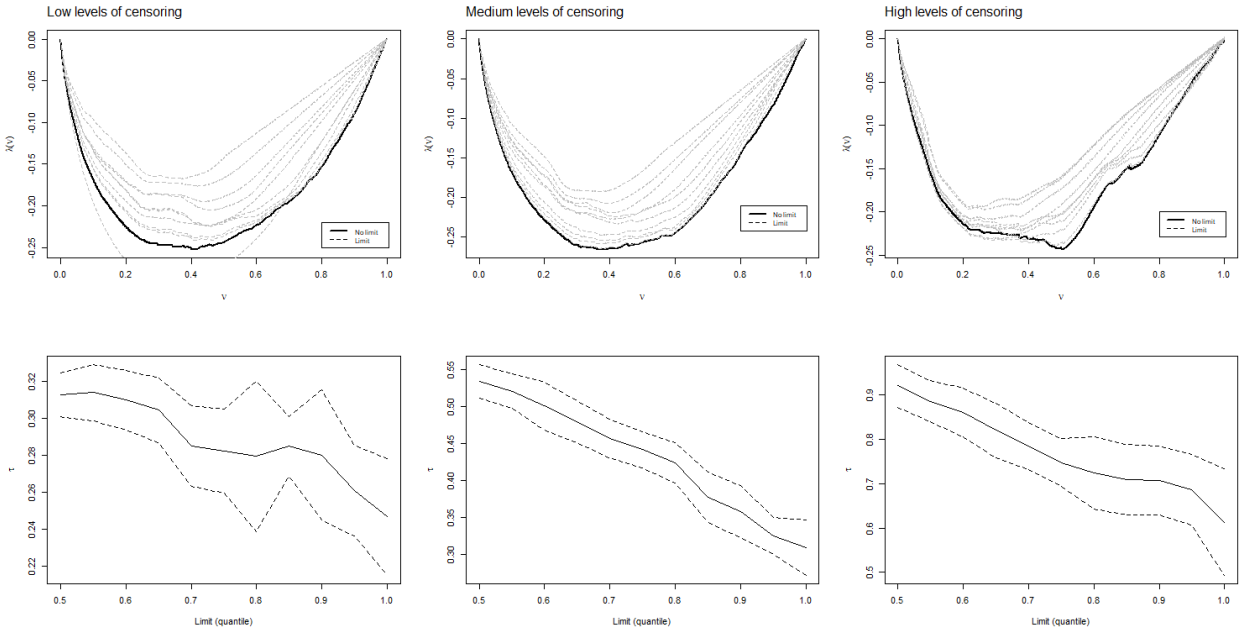

Figure 2.4 – Plots of $\lambda(\nu)$ and Kendall's tau for the joint distribution using different limits and different levels of censoring. Low, medium and high levels of censoring correspond to, approximately, 5%, 30% and 75%.

2.5 Simulation from $\hat{\psi}_n(.)$

Although the graphical procedure described in Section 2.4 can lead to the selection of an adequate parametric copula model for a given dataset, we might be interested in working directly from the non-parametric estimator of the generator $\hat{\psi}_n(\nu)$ obtained in Section 2.3. This could be particularly interesting in cases where the graphical comparison does not lead to a clear copula, for example if the curve of $\hat{\lambda}_n(\nu)$ lies somewhere between the curves of two candidate models.

When this occurs, an attractive alternative would consist in foregoing any parametric model and avoid labelling the copula. Besides the obvious advantage of removing any parametric constraints linked to the already-defined models, working directly from $\hat{\psi}_n(\nu)$ implies that we do not have to select a pool of candidate models nor perform any goodness-of-fit tests or model validations. Instead, we would directly simulate from the generator to retrieve the starting correlated variables. This is possible when considering Bernstein's

theorem :

Theorem 2.2 (Bernstein) *A function $\psi(.)$ is strictly monotonic and $\psi(0) = 1$ if and only if it can be written as*

$$\psi(\nu) = \mathcal{L}_\Theta(\nu), \tag{2.8}$$

where $\mathcal{L}(.)$ is the Laplace-Stieltjes transform of the strictly positive random variable Θ.

As highlighted by (Hofert, 2008), if we are able to sample from the inverse Laplace-Stieltjes transform of the generator, then we can use existing algorithms such as the one proposed in (Marshall et Olkin, 1988) to simulate d-variate vectors of observations $(U_1, ..., U_d)$ from the Archimedean copula from which the data originates.

Algorithm 1: Marshall, Olkin

1: Generate a random observation θ from the distribution with Laplace transform ψ.
2: For $i = 1, ..., d$, generate i.i.d. $X_i \sim \mathsf{U}(0, 1)$.
3: Return $(U_1, ..., U_d)$ where $U_i = \psi(-\log(X_i)/\theta)$, for $i = 1, ..., d$.

(Ridout, 2009) proposes a method for the first step of Algorithm 1, namely the RLAPTRANS Algorithm. This algorithm is a standard modification of the classic Newton-Raphson method to ensure convergence when generating random numbers with an inverse Laplace transform. We slightly modify it in order to be able to use it with our non-parametric estimator function $\hat{\psi}_n(\nu)$. The main steps are summarized below but more details can be consulted in (Ridout, 2009).

Algorithm 2: RLAPTRANS

1: Generate n independent $U(0, 1)$ observations and sort them : $u_{(1)} < ... < u_{(n)}$.
2: Find a value θ_{max} to serve as upper bound, i.e. $\psi^{-1}(\theta_{max}) \geq u_{(n)}$.
3: Set the lower bound $\theta_L = \theta_{(i-1)}$ and upper bound $\theta_U = \theta_{max}$. Repeat the modified Newton-Raphson procedure for $i = 1, ..., n$ to obtain the ordered sample $\theta_{(1)}, ..., \theta_{(n)}$.
4: Permute the ordered sample randomly to obtain the unordered sample $\theta_1, ..., \theta_n$.

Thanks to this approach, we can relax any assumption related to a parametric model and are able to directly use the data to perform new simulations and, in a next step, predictions. We also note that since any censo-

ring schemes present in the original data is already accounted for in the non-parametric estimator $\hat{\psi}_n(\nu)$, this method works well regardless of the censoring scenario at hand.

2.5.1 Simulation study

We now illustrate the use of the non-parametric generator function $\hat{\psi}_n(\nu)$ instead of a pre-defined parametric copula model by means of some simulation studies.

In a first example, we generate a dataset of $n = 500$ bivariate observations from a Clayton copula with Kendall's tau equal to 0.4, marginal exponential distributions of parameter equal to 1 for (T_1, T_2) and with the censoring variables (X_1, X_2) also following exponential distributions with parameters set such that around 20% of observations have at least one censored component. Figure 2.5 displays the original non-parametric estimator $\hat{\lambda}_n(\nu)$ for the Clayton data in black compared to the results of direct simulations from $\hat{\psi}_n(\nu)$ in blue. In total, we simulated 100 new sets of size $n = 500$, i.e. $50\,000$ new pairs of observations (Y_1, Y_2). For each of these new sets, we used again the approach laid out in Section 2.3 to retrieve the estimator $\lambda_{i,\hat{\psi}_n}(\nu)$, for $i = 1, ..., 1000$. The dashed blue curve depicted in Figure 2.5 is the average of these 1000 curves, i.e. $\bar{\lambda}_{\hat{\psi}_n(\nu)}$ and the two dotted curves represent the 95% confidence interval of these simulations.

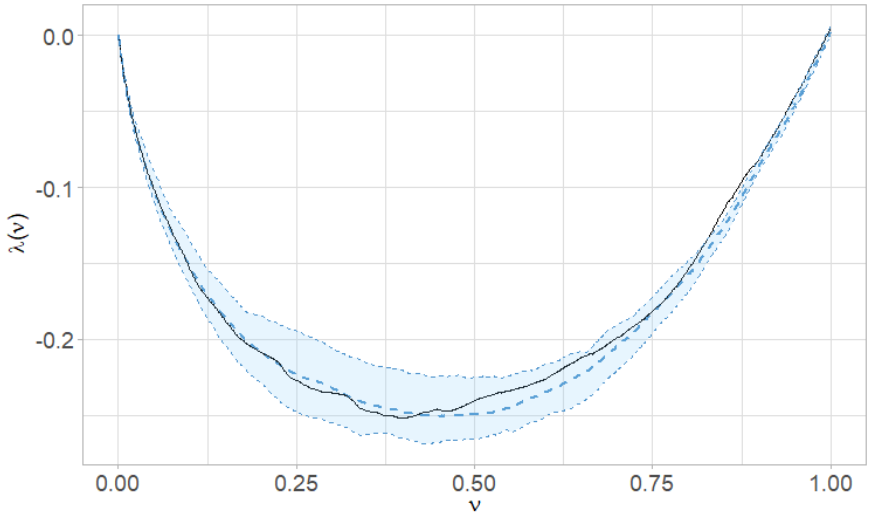

Figure 2.5 – Comparison between $\hat{\lambda}_n(\nu)$ (black continuous curve) initially obtained for the original data and the average $\bar{\lambda}_{\hat{\psi}_n}(\nu)$ (dashed blue curve) obtained across 1000 simulations using Algorithms 2 and 1. The shaded area represents the 95% confidence interval for the simulations.

We observe that our non-parametric simulations provide results that are very close to the original data. This is further confirmed by the plots displayed in Figure 2.6 that show the scatter plots of the original Clayton data in black (top left) and of the average simulations in blue (top right), as well as a comparison between the original and simulated densities for Y_1 (bottom left) and Y_2 (bottom right). On the density plots, the black continuous curves represent the original densities and the black vertical curves the average observed values for both variables. The dashed blue curve and vertical blue curves show, respectively, the simulated densities and simulated average values.

From the curves in Figure 2.5 and the scatterplots in Figure 2.6, we see that the simulated sets have the same shape and strength of dependence as the original data. Both scatterplots display the typical shape expected for data issued from a Clayton copula, with a clear concentration of points on the bottom left corner of the plot. The similar levels of concentration of points observed in both plots, as well as the similar shapes of the curves in Figure 2.5 further show that the dependence between our simulated bivariate observations is also of similar strength to that of the original couples. In addition, we observe in Figure 2.6 that the simulated marginal densities for Y_1 and Y_2 in, respectively, the bottom left and bottom right plots, are similar to the original densities shown in black. This further illustrates that directly simulating from the non-parametric generator allows to generate very similar sets of observations, without the need to specify a parametric model nor perform any model validation tests.

In Figure 2.7, we show similar simulation results when the data originates from a Frank, Gumbel and Joe copulas for different values of Kendall's tau, using heat maps for the simulated scatterplots. In each case, we observe that the sets simulated using the non-parametric generator display the same shapes and strength of dependence as the original data. When $\tau = 0.25$, the heat maps of the simulated data points are rather diffuse and it is harder to recognise the copula from which the data originates. As τ increases, however, the shapes of the heat maps become clearer and look more and more like the typical shapes expected for the different copula models. For the data issued from the Frank copula on the first row of plots, we observe that the points gather more on the diagonal as the level of dependence increases. For the Gumbel copula, we clearly see the higher densities in the bottom left and top right corners of the plots becoming more apparent as τ increases. Finally, for the Joe copula, we see the density becoming more important on the top right corner of the plot as Kendall's tau becomes larger.

These simulations illustrate how using the RLAPTRANS algorithm in combination with an algorithm such

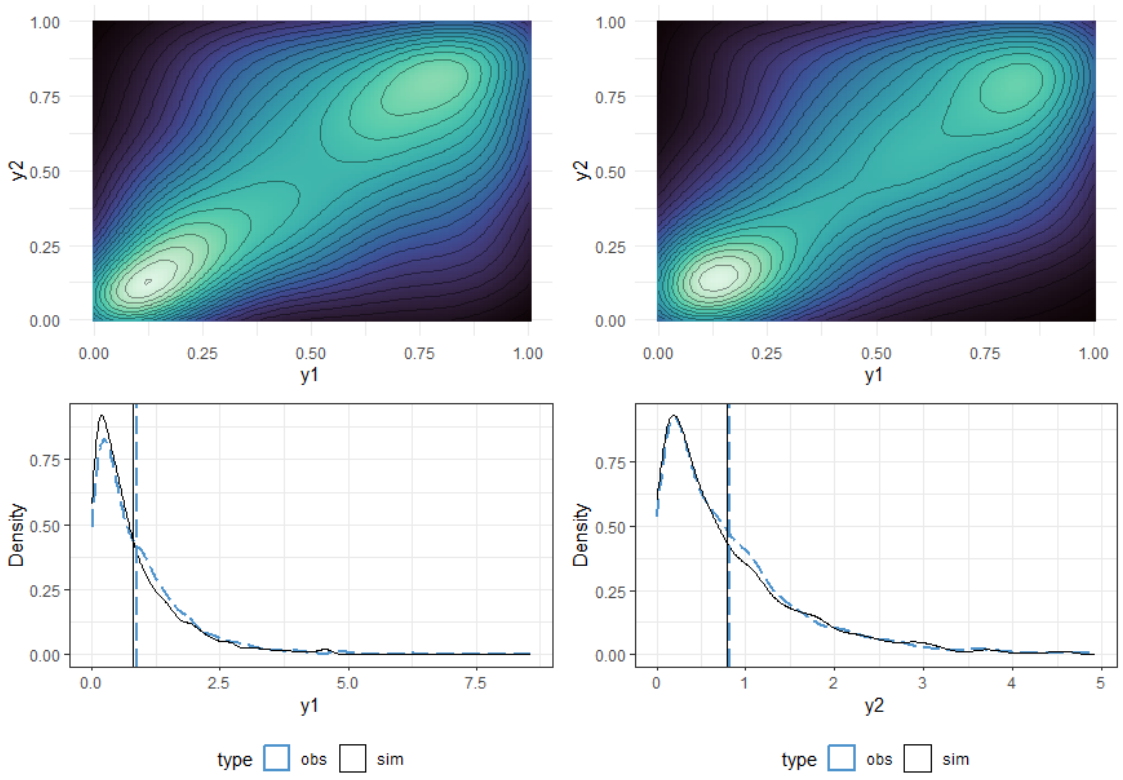

Figure 2.6 – Scatterplots of the original data issued from a Clayton copula (top left) and of the re-simulated sample (top right), and comparison between the original and simulated densities of Y_1 (bottom left) and Y_2 (bottom right). On the density plots, the black continuous curves show the original densities and the black vertical line the observed average values, while the dashed blue curves and vertical blue lines represent, respectively, the simulated densities and simulated average values.

as the one proposed by (Marshall et Olkin, 1988) allows to efficiently simulate correlated pairs of observations from an Archimedean family without the need to define a parametric copula model. Using the non-parametric estimator $\hat{\psi}_n(\nu)$, we are easily able to simulate new pairs of observations with the same shape and strength of dependence as the original sample. These new simulations can then be used in prediction models, as we illustrate in the next Section.

2.6 Application to automobile insurance claims

In this section, we apply the approach described in Section 2.3 to claims reserving, using a recent Canadian automobile insurance dataset.

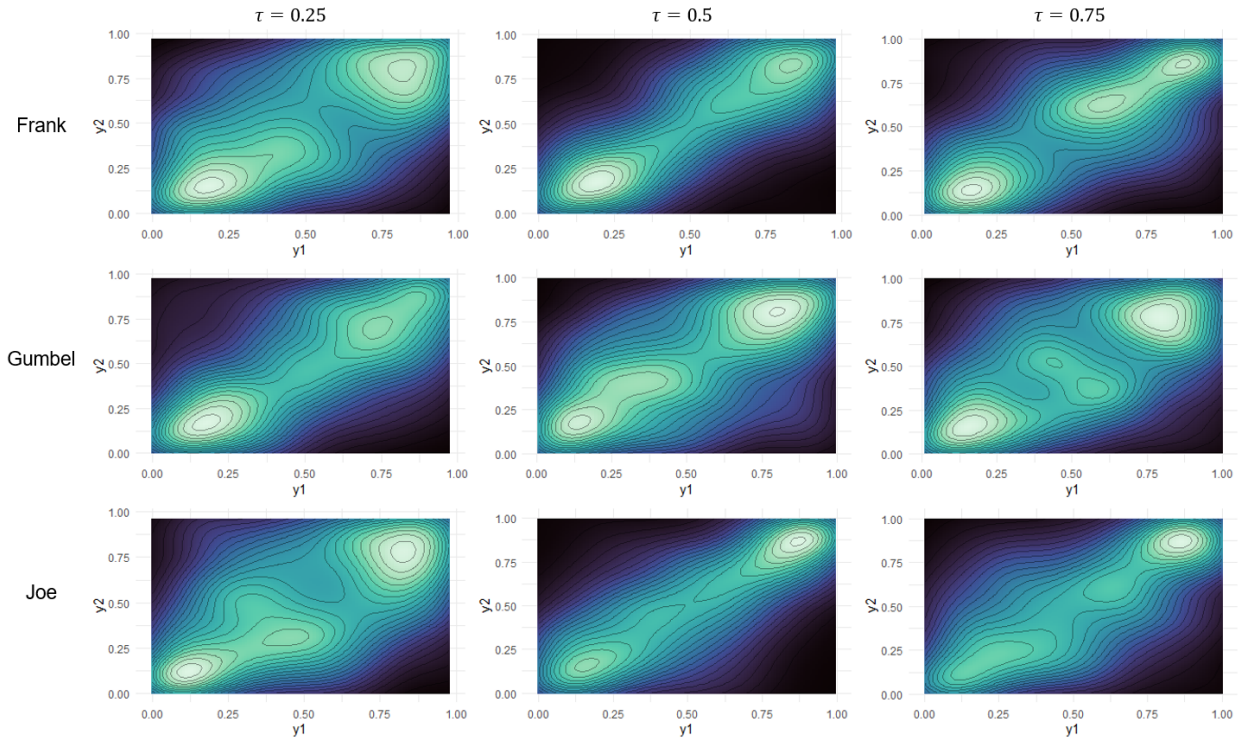

Figure 2.7 – Heat maps for the scatterplots of data simulated from the non-parametric estimator of the generator function, when the original sample comes from either the Frank, Gumbel or Joe copulas for different values of Kendalls's tau.

2.6.1 Data description

The data comprises over $600\ 000$ claims that occurred between January 2015 and June 2021. For each of these claims, we have information related to the policyholder, the vehicle driven or the claim itself. In addition, each policy in force in this dataset provides four different coverages to its holder, and each claim can fall under one or more of these coverages : the Accident Benefits coverage (loss of revenue, funeral expenses for the insured), the Bodily Injury coverage (loss revenue or medical expenses for a third party), the Vehicle Damage coverage (damages incurred to the insured's or a third party's vehicle) and the Loss of Use coverage (temporary replacement vehicle and alternative means of transportation). Table 2.5 provides some insights into the dynamics of our portfolio by presenting the relative importance of each of these four coverages in terms of the proportion of claims, total cost and total reserve that they represent. We observe that even though the Vehicle Damage and Loss of Use coverages are triggered by much more claims than the Accident Benefits and Bodily Injury coverages, they barely represent together 15% of the total reserve, taking January 1^{st}, 2019 as valuation date. Although less than 6% of all claims trigger the Bodily

Injury coverage, this coverage alones represent more than half of the total portfolio reserve.

Table 2.5 – Weight of each coverage in the portfolio. The percentages with respect to the total reserve are calculated by taking the 1^{st} of January 2019 as valuation date

Coverage	$\%$ of claims	$\%$ of total cost	$\%$ of total reserve
Accident Benefits	9.42	12.82	30
Bodily Injury	5.70	13.13	55
Vehicle Damage	96.39	70.44	14
Loss of Use	51.89	3.61	< 1

Table 2.6 further illustrates the importance of the Accident Benefits and Bodily Injury coverages for the insurer by presenting some descriptive statistics for the severity of payments of all four coverages. Bodily Injury and Accident Benefits claims have the largest average payments and present large values in the higher quantiles, indicating heavy-tailed distributions. In the rest of this section, we thereby choose to focus on these two coverages only, which are clearly the most important cost-wise in the portfolio.

Table 2.6 – Descriptive statistics for the severity of payments of the four insurance coverages

Coverage	Mean	Std. dev.	Quantiles				Max.
			0.5	0.75	0.95	0.99	
Accident Benefits	12,386	53,561	3,215	6,909	47,757	127,896	2,435,334
Bodily Injury	23,271	76,027	4,000	15,150	98,449	322,612	2,039,570
Vehicle Damage	5,040	8,121	2,605	5,830	17,984	40,611	149,399
Loss of Use	545	620	419	714	1,000	2,336	52,777

2.6.2 Notation

Our goal is to model the dependence between them using their activation delays. These are defined as the time elapsed between the reporting date of the claim and the date at which the insurer first triggers the coverage in the claims management system. This is the date the insurer records that the claim falls under

that specific coverage. We denote the vector of activation delays by $\boldsymbol{T} = (T_1, T_2)$. If the claim settles before one of the coverages is triggered, then the delay for this coverage is censored. The censoring variable, that we denote by $\boldsymbol{X} = (X_1, X_2)$, represents the claim settlement delay, i.e. the delay between the reporting date and settlement date of the claim.

In this specific application, we choose not to consider claims that settle with no payment emitted from the insurance company, nor claim re-openings. As such, we might need to be careful with the independence assumption between $\boldsymbol{T}$ and the censoring times. However, empirical analyses not shown here have demonstrated that the correlation between the activation delays and settlement delays for the Accident Benefits and Bodily Injury coverages is close to zero. In more classical reserving applications, actuaries typically use datasets with claims that might reopen after being settled the first time, that could settled without payment or even that could become settled for external reasons. The assumption that the activation delays are independent from the settlement delays chosen as censoring times thereby holds in general reserving frameworks.

In addition, policyholders may be unable to receive compensation for a specific coverage if a certain amount of time has passed since reporting. This limit on the delays can depend on the nature of the claims, local or governmental regulations, or even on company-level rules of the insurer. Both internal and external agents can set it. This is our limit ω_i, for $i = 1, 2$ introduced in Section 2.3.1. Suppose a coverage has not been triggered yet when this limit passes. In that case, its activation delay becomes censored, even if the claim remains open.

Figure 2.8 illustrates the typical development of a claim and the notation introduced above. In this example, a claim occurred on June 1^{st} and was reported the following day. On June 5^{th}, the insurer records that the claim has impacted coverage 1. Recording the delays in days, the activation delay for this coverage is then $T_1 = 3$. Assuming a limit of $\omega_1 = \omega_2 = 30$ days on the activation delays, coverage 2 becomes censored on July 2^{nd}, even if the claim is still open. It settles two days later, leading to $X_1 = X_2 = 32$. The observed delays are then given by $(Y_1, Y_2) = \big(\min(T_1, X_1, \omega_1), \min(T_2, X_2, \omega_2)\big) = \big(\min(3, 32, 30), \min(\infty, 32, 30)\big) = (3, 30)$.

Considering the observed activation delays shown (in periods of six months) in Table 2.7 for the Accident Benefits and Bodily Injury coverages in the portfolio at hand, we set the limit at $\omega_1 = \omega_2 = 730$ days,

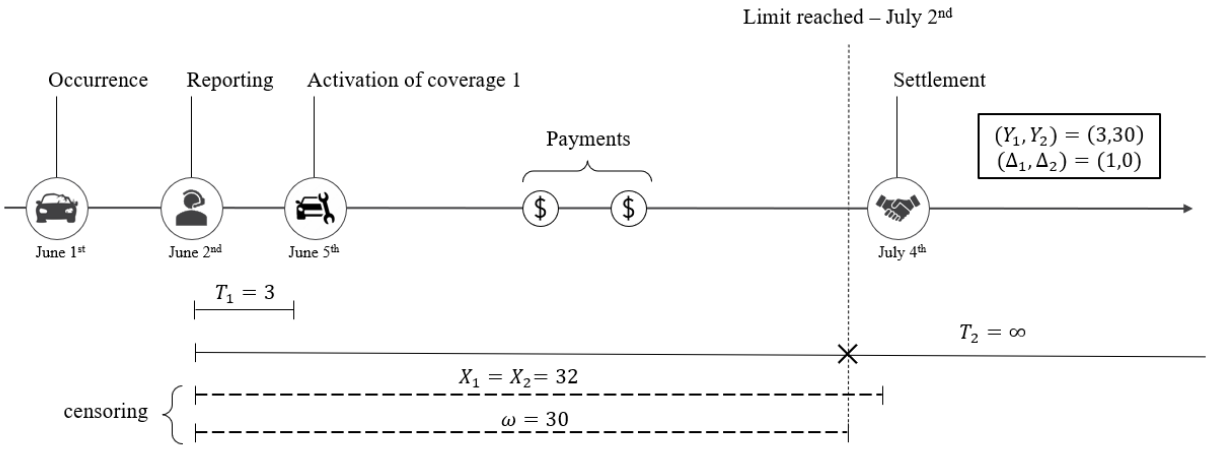

Figure 2.8 – Typical claim development and illustration of the notation used.

equivalent to two calendar years. Almost 94% of all Accident Benefits claims trigger that coverage within six months of reporting, and most of the remaining claims will be labelled as Accident Benefits no later than a year after the reporting date. The activation delays are longer for Bodily Injury claims, with 1.69% of them taking two years or more after reporting to be labelled as such. In our portfolio, a claim always triggers at least one coverage. As such, 22.4% of observed claims trigger both the Accident Benefits and Bodily Injury coverages, 55.4% of the claims only activate the Accident Benefits coverage and the remaining 22.2% of claims are solely Bodily Injury claims. In other words, the levels of censoring observed for both coverages are of 22.2% and 55.4% for, respectively, Accident Benefits and Bodily Injury.

Table 2.7 – Percentage of claims with different activation delays, shown in periods of six months.

	Activation delays				
Coverage	No delay	1 period	2 periods	3 periods	≥ 4 periods
Accident Benefits	93.84	5.73	0.29	0.08	0.06
Bodily Injury	85.86	9.86	1.46	1.13	1.69

2.6.3 Analysis

After using the approach described in Section 2.3 to find the non-parametric estimator of $\hat{K}_n(\nu)$ and $\hat{\psi}_n(\nu)$, we first find the parametric copula best fitted to our portfolio using the graphical procedure from Section 2.4. We present the plots of the non-parametric estimators $\hat{K}_n(\nu)$ and $\hat{\lambda}_n(\nu)$ in Figure 2.9, along with the corresponding fitted curves for four candidate Archimedean models. In both plots, the non-parametric estimator is closest to the curves of the Joe copula. With an estimated $\hat{\tau} = 0.2705$, the Joe copula with $\hat{\alpha}_J = 1.6652$ best fits our data.

The omnibus procedure first confirms this result. We estimate the dependence parameters for the competing models using the likelihood in Equation (B.1). The results shown in the third and fourth columns of Table 2.8 indicate that the Joe copula is most appropriate since it presents the smallest difference between $\hat{\alpha}$ and $\hat{\alpha}^*$. Similarly, the results of 1000 bootstrapped simulations for the L^2-norm validation approach shown in the fifth column of Table 2.8 show that $D(\hat{\alpha})$ is the smallest for the Joe copula. We finally apply (Wang, 2010)'s goodness-of-fit test and find again in the last column of Table 2.8 that the Joe model presents the smallest percentage of rejection when we perform 1000 simulations for each competing model.

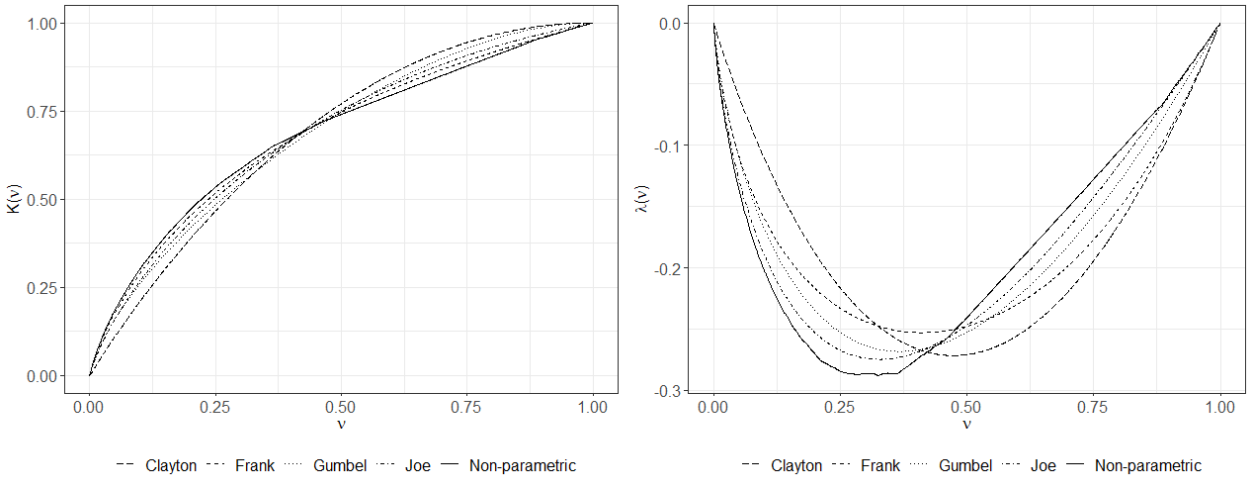

Figure 2.9 – $K(\nu)$ and $\lambda(\nu)$ for the different copulas and for the Canadian automobile insurance dataset with a limit at 730 days.

Table 2.8 – Results validation, based on 1000 bootstrapped simulations for the L^2-norm and 1000 simulations for (Wang, 2010)'s test.

Copula	$\hat{\tau}$	Omnibus procedure		L^2-norm	(Wang, 2010)
		$\hat{\alpha}^*$	$\hat{\alpha}$	$D(\hat{\alpha})$	% rejection H_0
Clayton	0.2705	0.2432	0.7417	0.00399	0.9994
Frank		1.1166	2.5612	0.00180	0.9265
Gumbel		1.0554	1.3708	0.00084	0.8742
Joe		1.3821	1.6652	0.00033	0.0476

Even though the results from Table 2.8 indicate that the Joe copula is an appropriate fit for the data, we observe on the plot of the $\lambda(.)$ functions in Figure 2.9 that the non-parametric $\hat{\lambda}_n(\nu)$, although being indeed closest to the curve $\lambda_{\hat{\alpha}_J}(\nu)$ for the Joe model, is still a little distance away from it. In this case, it might then be more fitting to directly work with the non-parametric generator $\hat{\psi}_n(\nu)$, using the algorithms described in Section 2.5. In what follows, we compare simulation results obtained with the Joe copula selected with the graphical procedure and by working directly from the generator function for the data at hand.

In Figure 2.10, we present a comparison between $\hat{\lambda}_n(\nu)$ (black continuous curve), the non-parametric estimator of $\lambda(.)$ obtained for our insurance portfolio using the methodology laid out in Section 2.3, $\lambda_{\hat{\alpha}_J}(\nu)$ (dark blue dotted curve), the corresponding curve for the selected parametric Joe model, and $\lambda_{\hat{\psi}_n}(\nu)$ (light blue dashed curve), the curve obtained when re-simulating the data using the non-parametric generator $\hat{\psi}_n(\nu)$ and Algorithms 1 and 2. The dashed blue curve displayed on the plot is the average of performing 1000 simulations and the shaded area shows the distance between the smallest and highest curves among these simulations. We observe that the curve for the data simulated from the non-parametric generator is much closer to the curve for the original data than the one from the parametric Joe copula model. This indicates that the new simulated observations for the Accident Benefits and Bodily Injury activation delays have a dependence structure and a strength of dependence that are closest to the ones observed in the original data than the observations simulated using the parametric copula model.

In Figures 2.11 and 2.12, we compare the results obtained from simulations using the parametric and non-

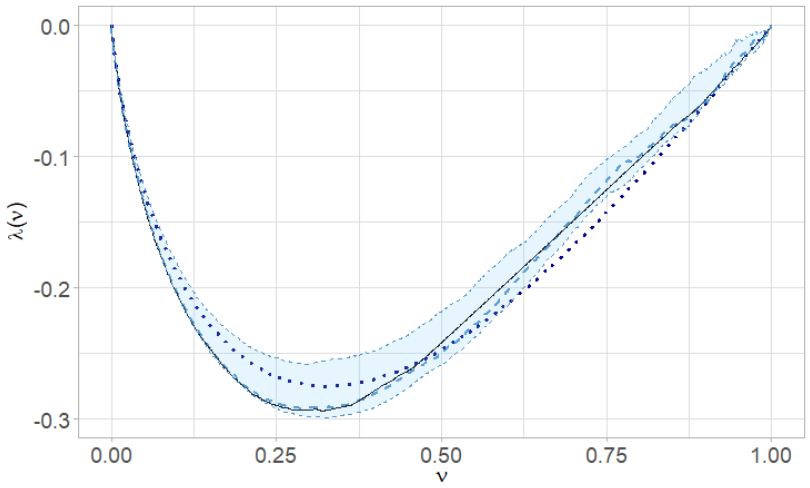

Figure 2.10 – Comparison between the non-parametric $\hat{\lambda}_n(\nu)$ estimated for the original data (blue continuous curve), the parametric $\lambda_{\hat{\alpha}_J}(\nu)$ (dark blue dotted curve) for the selected Joe copula model and $\lambda_{\hat{\psi}_n}(\nu)$ (light blue dashed curve) for the data simulated 1000 times using the generator $\hat{\psi}_n(\nu)$.

parametric models. Figure 2.11 shows the heat maps for the scatterplots of the Accident Benefits and Bodily Injury activation delays, simulated from the parametric Joe copula model (left) and using the non-parametric estimator of the generator function (right). We observe that, although both heat maps display a high concentration of points in the top right corner of the plots, as is typically observed with Joe copulas, this concentration seems higher for the non-parametric model. This would indicate a closer fit with the typical Joe copula scatterplot than for the data simulated from the parametric model. This is aligned with the results observed in Figures 2.9 and 2.10. In Figure 2.9, the curve of the Joe copula model is quite close to the curve for the Gumbel model. This could explain that we also observe a small concentration of points at the bottom left corner of the scatterplot for the parametric model in Figure 2.11 : although we are indeed simulating from a Joe copula, the plots from Figure 2.9 indicate that this model is not too far from a Gumbel copula. In Figure 2.10, the curve for the non-parametrically simulated data is closer to the curve of the original data and more pulled towards the bottom left corner of the plot than the curve for the Joe copula. As such, the dependence structure usually observed in a Joe model seems more pronounced in the original data and in the non-parametric simulations than it is in the parametric simulations.

In Figure 2.12, we compare the original density functions of the Accident Benefits and Bodily Injury activation delays to the densities simulated using the parametric and non-parametric models. For both coverages, we observe that the densities for the non-parametric simulations (light blue dashed curves) are closest to

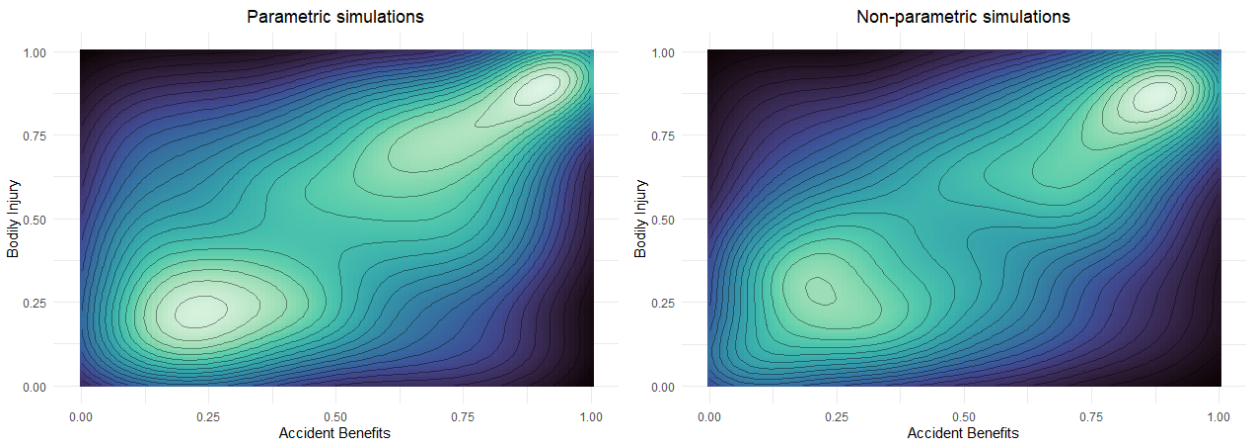

Figure 2.11 – Heatmaps of the scatterplots for some simulations of the Accident Benefits and Bodily Injury activation delays, using the parametric Joe copula model (left) and the non-parametric generator function (right).

the true densities (black continuous curves) than the densities for the parametric model (dark blue dotted curves). This is particularly true for the Bodily Injury coverage, where we see some very similar shapes between the black and light blue curves. The vertical lines represent the average observed activation delays (black), average simulated delays with the non-parametric model (light blue) and with the Joe copula (dark blue). These values are also given in Table 2.9. The Joe copula model appears to underestimate the activation delays for both coverages, particularly for Bodily Injury claims with a difference of over 60 days between an average simulated delay of 167.17 days compared to an observed average of 233.50 days. The difference if of almost 40 days for the Accident Benefits coverage, with an average simulated delay of 67.67 days compared to an observed delay of 107.12 days. The non-parametric model simulations provide better accuracy, with a close average simulated delay of 100.29 days for the Accident Benefits coverage and a slightly higher estimated delay of 245.12 days for Bodily Injury claims.

Table 2.9 – Average activation delays (observed and simulated).

Coverage	Observed	$\hat{\psi^{-1}}(\nu)$	Joe copula
Accident Benefits	107.12	100.29	67.67
Bodily Injury	233.50	245.12	167.17

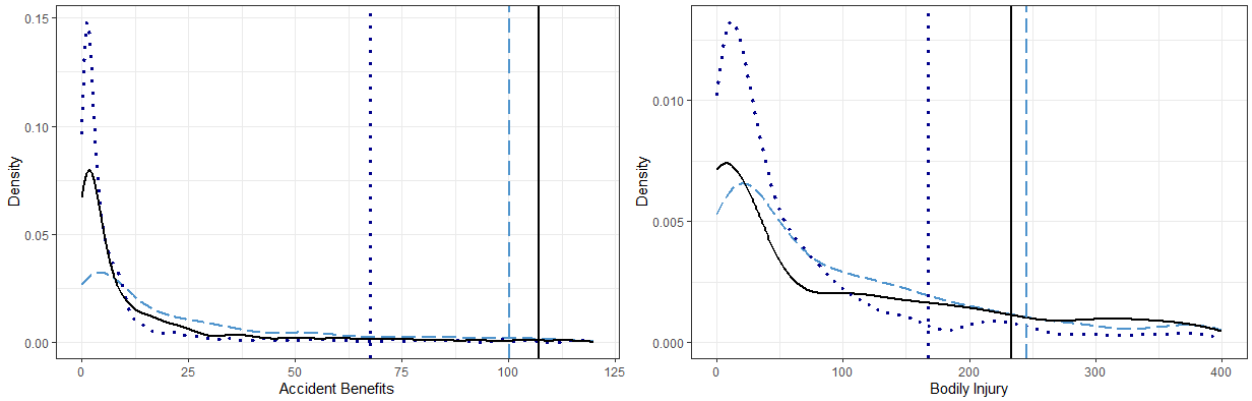

Figure 2.12 – Comparison between the original (black continuous curve) and simulated marginal densities for the Accident Benefits (left) and Bodily Injury (right) coverages. The light blue dashed curves represent the simulated density using direct estimation from the generator function, while the dark blue dotted curves show the simulated density using the Joe copula. The vertical lines represent the average of the observed and simulated delays.

2.6.4 A simple claims reserving example

To illustrate the usefulness of our model and the way it can be used as part of a larger framework, we provide in this section a simple reserves simulation scheme.

Assuming that a claim costs a fixed amount of $1\$$, we use the classic exponential growth model to evaluate the portfolio reserve for different levels of inflation, taking January 1^{st}, 2019 as valuation date. Let r denote the annual inflation rate. The annualised cost for claim i related to insurance coverage j, denoted by $Z_{i,j}$, is given by

$$Z_{i,j} = \left(1 + \frac{r}{360}\right)^{Y_{i,j}},$$

where $Y_{i,j}$ is the activation delay of coverage j for claim i.

Table 2.10 presents the results of $5\ 000$ simulations for the Accident Benefits and Bodily Injury reserves, estimated using the parametric Joe copula models and the non-parametric model. We let the inflation r vary from 2% to 8% by increments of 2%. In each scenario, we observe that the simulated reserves with the non-parametric model are closer to the true amounts than those obtained with the parametric copula. This was to be expected, considering the results observed in Figures 2.10, 2.11 and 2.12. For the Accident Benefits coverages, the simulated reserves with the non-parametric model are slightly smaller than the

true reserves but higher than the predictions obtained with the Joe copula. Note, however, than in a more realistic reserving setting where we would fit models for the claim severities, the small underestimation that we observe here would probably disappear or become insignificant. For the Bodily Injury coverage, non-parametric predictions are slightly larger than the true amounts while the predictions made from the Joe model underestimate the true reserves. These results are in line with what we observe in Table 2.9, where the predictions for the Accident Benefits delays are a bit smaller than the observed delay with the non-parametric model and even smaller with the parametric model. For the Bodily Injury coverage, the predictions for the delays are a bit higher for the non-parametric model and much lower for the Joe copula model.

Although we use here a very simplified model for illustration purposes, the results already hint that using the more accurate delay predictions from the non-parametric models will certainly lead to greater accuracy in the reserves predictions as well.

Table 2.10 – Comparison between the true reserves and predicted reserves from the non-parametric and parametric models under different inflation scenarios.

Coverage	Model	Value	Inflation rate			
			0.02	0.04	0.06	0.08
Accident Benefits	Observed		7593.67	7644.94	7697.88	7752.53
	Non-parametric	Average	7585.83	7628.75	7672.76	7717.97
		$\text{VaR}_{0.95}$	7587.33	7631.95	7677.47	7724.44
	Parametric	Average	7578.17	7613.32	7649.56	7687.01
		$\text{VaR}_{0.95}$	7579.70	7616.51	7654.29	7693.47
Bodily Injury	Observed		7640.68	7740.23	7842.75	7948.33
	Non-parametric	Average	7648.70	7756.59	7867.75	7982.11
		$\text{VaR}_{0.95}$	7650.93	7761.01	7874.65	7991.37
	Parametric	Average	7618.66	7695.55	7774.71	7856.43
		$\text{VaR}_{0.95}$	7620.71	7699.67	7780.94	7865.09

2.7 Conclusion

In this paper, we present a non-parametric estimator for the generator function of the Archimedean family of copulas, applicable to flexible censoring scenarios. Following the initial idea of (Genest et Rivest, 1993) and combining it with tools from survival analysis in order to account for the presence of incomplete data, we derive the estimator $\hat{\psi}_n(\nu)$.

We then present two approaches to use it. We begin with a graphical procedure that compares the estimator for a given dataset to different parametric copula models, and select the best fitted one among them. Thanks to our extension of the non-parametric estimator to various censoring schemes, we show that this simple graphical selection method works well with both complete and incomplete data by illustrating three results validation approaches.

Then, rather than seeking to use a well-defined parametric copula model, we suggest to directly work from the non-parametric estimator of the generator function to perform simulations and further analysis on the data. This approach is interesting, for example, when the graphical comparison does not lead to a clear resemblance between the non-parametric curve and one of the candidate copula models. It allows to forego any parametric assumptions and work instead directly from the non-parametric estimator without labelling it to any Archimedean family. Viewing the generator as the Laplace-Stieltjes transform of a strictly positive random variable, we propose to use the RLAPTRANS algorithm of (Ridout, 2009) in combination with an Archimedean copula sampling algorithm like the one suggested by (Marshall et Olkin, 1988). With just a small modification of the RLAPTRANS algorithm to use the non-parametric function rather than a parametric one, we are able to generate random observations using the inverse Laplace-Stieltjes transform. We then use these random observations to generate couples of observations with the same shape and strength of dependence as those in the original data. Considering that the non-parametric estimator we derive for the generator function takes into account the flexible censoring scenarios that can be found in the data, this method can also be used in the presence of incomplete data. We demonstrate this by means of some simulations studies.

We then apply our methods to an automobile claims insurance portfolio for which we seek to model the dependence between two insurance coverages by means of their activation delays. These correspond to the time elapsed between the reporting date of a claim and the date at which the claim triggers one of the coverages under which the policyholder is insured. Our data presents two levels of censoring. First,

if a coverage does not become active before the claim settles, then the activation delay for this specific coverage is censored. Second, considering current insurance legislation in Canada and the specific nature of our claims, we apply a limit of 730 days after which, even if a claim is still open, additional coverages can not be triggered anymore. This limit implies that after a certain amount of time has passed, policyholders are not able to seek new compensation anymore, even if their claim is still open. This acts as a second level of censoring on the activation delays of the coverages. We find non-parametrically the generator function in this specific censoring scenario and then compare the estimation results when using both the best-fitted parametric copula model, selected via the graphical procedure, and direct simulations from the generator. In this specific case, even though the Joe copula with parameter $\hat{\alpha} = 1.6652$ is deemed well-fitted to the data by the three model validation approaches, there exists a small gap on the plot of the $\lambda(.)$ functions in Figure 2.9 between this parametric model and the curve for the data. We show that in this case, directly simulating from the generator function using Algorihtms 1 and 2, leads to more accurate predictions of the activation delays than when using the Joe copula model. We conclude by a simple illustration of how these results can then be used as part of a larger claims reserving framework. We demonstrate that using the predictions from the non-parametric model can lead to reserves predictions that are closer to the true reserves amounts.

This approach could be used as part of a larger model such as the one proposed in (Michaelides *et al.*, 2023) or (Côté *et al.*, 2022) where the estimation of the activation delays can replace, respectively, the activation patterns and the multinomial estimation of the claim type. A similar methodology could also be used in different models such as that proposed in (Antonio et Plat, 2014) where, for example, the next event's exact time for reported but not settled claims could be extended to a multivariate framework by considering simultaneously, as in Section 2.6, different business lines or insurance coverages.

CHAPITRE 3

PARAMETRIC ESTIMATION OF CONDITIONAL ARCHIMEDEAN COPULA GENERATORS FOR CENSORED DATA

Résumé

Dans cet article, nous proposons une approache novatrice afin d'estimer la fonction génératrice des copules archimédiennes de façon conditionnelle, en y incorporant des variables explicatives. Notre méthode permet d'évaluer l'impact de différents niveaux de ces variables à la fois sur la force et la forme de la dépendance, en portant directement sur l'estimation de la fonction génératrice plutôt que la copule elle-même. Ce faisant, nous contribuons à assouplir l'hypothèse simplificatrice inhérente à la modélisation traditionnelle des copules. Nous démontrons l'efficacité de notre approche à travers des applications dans deux contextes différents : une étude sur la rétinopathie diabétique et un modèle de micro-réserves. Dans les deux cas, nous montrons comment la prise en compte de l'influence des variables exogènes permet une modélisation plus précise de la structure de dépendance sous-jacente dans les données, améliorant ainsi l'applicabilité des modèles de copules, en particulier dans des contextes actuariels.

Abstract

In this paper, we propose a novel approach for estimating Archimedean copula generators in a conditional setting, incorporating endogenous variables. Our method allows for the evaluation of the impact of the different levels of covariates on both the strength and shape of dependence by directly estimating the generator function rather than the copula itself. As such, we contribute to relaxing the simplifying assumption inherent in traditional copula modeling. We demonstrate the effectiveness of our methodology through applications in two diverse settings : a diabetic retinopathy study and a claims reserving analysis. In both cases, we show how considering the influence of covariates enables a more accurate capture of the underlying dependence structure in the data, thus enhancing the applicability of copula models, particularly in actuarial contexts.

3.1 Introduction

Copulas are a powerful tool when we want to address the dependence among correlated variables. Widely used since their introduction by (Sklar, 1959) in various fields from finance and insurance to medicine, they allow to separate the dependence structure within a multivariate distribution. Multiple authors have extensively studied their different families. The Archimedean copulas, introduced by (Schweizer et Sklar, 1983), are particularly useful thanks to their easy characterization via a generator function $\psi(.)$. To model the dependence for a vector of d variables $\boldsymbol{Y} = (Y_1, ..., Y_d) \in \mathbb{R}^d$, the $d-$variate Archimedean copula $C(.)$ takes the form

$$C(F_1(y_1), ..., F_d(y_d)) = \psi\big\{\psi^{-1}[F_1(y_1)] + ... + \psi^{-1}[F_d(y_d)]\big\},$$

where the generator function $\psi(.) : [0, \infty] \to [0, 1]$ must be such that $\psi(0) = 1$ and $\lim_{\nu\to\infty} \psi(\nu) = 0$. From a copula $C(.)$, one can recover Kendall's tau and as such, assess the degree of association between the variables under consideration. Archimedean copulas are typically characterized by a small number of dependence parameters, some of them even possessing only one for which a direct relationship exists with Kendall's tau. This is notably the case for the Clayton, Frank, Gumbel and Joe copulas.

Although widely popular in the literature for a few decades now, it is only in recent years, thanks to the early work of (Patton, 2006), that some authors have began to consider the impact of risk factors on the dependence between random variables. (Patton, 2006) extended Sklar's theorem to incorporate a covariate Z in the copula model. More specifically, this novel approach allows to model the dependence between the random variables $(Y_1, ..., Y_d)$, conditionally on a covariate Z. The conditional copula $C(.|Z = z)$ is given by the joint distribution function of $(F_{1|Z}(Y_1|z), ..., F_{d|Z}(Y_d|z))$, where $Z = z$. In the bivariate case, (Patton, 2006) showed that the joint conditional distribution is then uniquely defined for any z in the support of Z as

$$H_Z(y_1, y_2|z) = C\big(F_{1|Z}(y_1|z), F_{2|Z}(y_2|z)|Z = z\big),$$

for all $(y_1, y_2) \in \mathbb{R}^2$.

Very often in conditional copula models, authors rely on the so called *simplifying assumption*. This assumption states that a copula $C(.|Z)$ is independent of Z :

$$C\big(F_{1|Z}(y_1|z), F_{2|Z}(y_2|z)|Z = z\big) = C\big(F_{1|Z}(y_1|z), F_{2|Z}(y_2|z)\big) \quad \forall z \in Z.$$

The effects of the covariate Z are only captured through the marginal distributions $F_{1|Z}(.)$ and $F_{2|Z}(.)$ while the copula itself remains unchanged for all values of z. The simplifying assumption is particularly useful in pair copula constructions that allow to specify multivariate distributions using only bivariate copulas as building blocks. They allow to capture highly complex dependence structures. In addition, assuming that the conditional copulas used in the hierarchical structure do not depend on the conditioning variables keeps the resulting multivariate distribution tractable for inference and model selection. Although this assumption greatly simplifies conditional copula models, voices have been raised in the statistical literature against it, see for example (Hobæk Haff *et al.*, 2010), (Acar *et al.*, 2012) or (Stöber *et al.*, 2013). In particular, some authors like (Levi et Craiu, 2018) argue for the need of effective tests against the simplifying assumption.

Some contributions to the literature have worked towards relaxing this assumption. Most of these focus on incorporating endogenous variables in the strength of dependence, often to estimate Kendall's tau or, equivalently, the dependence parameters for Archimedean copulas, as in (Acar *et al.*, 2011). However and to the best of our knowledge, no model presented so far allows for both the strength and the structure of dependence to implicitly vary with the values of a covariate.

The approach that we present in this paper aims at relaxing the simplifying assumption in the case of Archimedean copulas. More specifically, we propose a new parametric estimator for the generator function $\psi(.)$, suitable for censored data and inspired by the non-parametric estimator presented in (Genest et Rivest, 1993). With this new parametric model, we include covariates in the estimation of the generator function, thereby considering their impact on both the strength and shape of dependence between random variables of interest. This contrasts with most approaches presented in the literature that incorporate endogenous variables directly in the copula $C(.)$ rather than in the generator function. Besides enabling the relaxation of the simplifying assumption, this approach is parametric, contrarily to a lot of the models in the existing literature. Although non-parametric frameworks help avoiding assumptions that could be misleading, incorporating covariates in these models often proves difficult, and interpreting their effects is seldom straightforward. By using parametric models, we can more easily include multiple covariates and more readily analyze their impact. Another interesting feature of our model is that it is suitable for various censoring schemes for the dependent random variables. Incomplete data is often encountered in a variety of fields and although multiple authors have proposed copula models that can accommodate different types of censoring, see for example (Gribkova et Lopez, 2015), very few so far allow to incorporate endogenous variables. Finally, while most contributions to conditional copula models are found in the medical field, this

paper contributes to the actuarial literature. To the best of our knowledge, very few authors have used such frameworks in actuarial modeling. Considering the important quantity of information available to insurers, incorporating risk factors related to the claims or the policyholders in actuarial models helps refine the predictions made by actuaries. In this paper, we present an application of a conditional copula model to micro-level loss reserving.

This paper is structured as follows. In Section 3.2, we review the existing conditional copula models as well as some tests for the simplifying assumption. Section 3.3 presents the parametric estimator for the generator function of Archimedean copulas. In Section 3.4, we apply our model to the Diabetic Retinopathy dataset used, among others, by (Geerdens *et al.*, 2018). We present an application to granular reserving in Section 3.5. Using a Canadian automobile insurance dataset in which each policy in force provides different coverages, we investigate the dependence between these coverages and the impact that some covariates may have on this dependence. Section 3.6 concludes our work.

3.2 Literature review

In this section, we present an overview of the literature on conditional copulas. We discuss the contributions made since the pioneering work of (Patton, 2006), starting with the parametric approaches before moving on to the non-parametric and semi-parametric models. We then discuss some of the works related to the simplifying assumption.

As mentioned in Section 3.1, (Patton, 2006) was the first to set the basis of conditional copula modeling by extending Sklar's theorem. He uses conditional copulas to capture the dependence between the Deutsche mark-dollar and Yen-dollar exchange rates and the effect of time on this dependence, particularly before and after the introduction of the Euro. Several applications of these conditional copulas emerged in the financial literature simultaneously, each concentrating on integrating time-varying aspects into the dependence structure of ARMA models.

Several authors followed suit, marking the start of the literature on conditional copulas. Most contributions to date fall in the non-parametric domain. Among the first, (Acar *et al.*, 2011) use likelihood estimation for covariate-adjusted copulas in a local polynomial framework. Assuming the conditional marginal distribu-

tions $F_{1|Z}$ and $F_{2|Z}$ known, the authors focus on the model

$$(U_{1i}, U_{2i})|Z_i \sim C(u_{1i}, u_{2i}|\theta(z_i)),$$

with $\theta(z_i) = g^{-1}\{\eta(z_i)\}$, for $i = 1, .., n$, the dependence parameter. Here, g^{-1} is a known inverse link function that keeps the dependence parameter in its appropriate range. η is the unknown calibration function that the authors estimate using local maximum likelihood and the framework of local polynomials.

In a similar vein, (Valle *et al.*, 2017) also opt for a non-parametric approach, using a flexible Bayesian framework for the estimation of conditional copula densities. Their method allows to overcome the issue of selecting among various types of copulas. (Bouezmarni *et al.*, 2019) use an inverse-probability-of-censoring weighting approach in a bivariate case to derive a non-parametric estimator of the conditional marginal distribution with one covariate when only one of the correlated variables is subject to random censoring. Focusing solely on modeling the strength of dependence, (Derumigny et Fermanian, 2019) directly estimate the conditional Kendall's tau in the presence of a vector of covariates by using different kernel-based methods. The authors then prove several theoretical properties of these estimators. In (Derumigny et Fermanian, 2022), the same authors prove the weak convergence of conditional empirical copulas processes under various conditioning events with non-zero probabilities, extending the work of (Segers, 2012).

In the actuarial literature, only few authors until now have delved into the early work of (Patton, 2006) and explored conditional copulas within actuarial models. Among rare examples, (Yang *et al.*, 2020) propose a non-parametric method to estimate copula regression models with discrete outcomes, introducing a perturbed version of the probability integral transform. They apply their methodology to estimate insurance claim frequencies across different business lines. Also in the context of multiple business lines in insurance, but this time in a claims reserving setting, (Yang, 2022) develops a copula estimator for mixed data. In his paper, the author accounts for the mixed nature of claims severities with a probability mass at zero and a positive continuous part.

Although the literature on conditional copula accounts for a lot of non-parametric models, some authors have also built on the work of (Patton, 2006) to propose parametric approaches, mostly using Bayesian regression copulas. This is the case, for example, in (Pitt *et al.*, 2006), with different types of outcomes and distributional assumptions for the marginals. (Fermanian et Wegkamp, 2012) propose time-dependent copulas where each dependent random variable depends on its value at the previous lag. Another interesting example, (Hans *et al.*, 2022) use boosting to estimate the coefficients of Generalized Additive Models for

the Location, Scale and Shape (GAMLSS) for the parameters of the marginal distributions and dependence parameters of Gaussian, Clayton and Gumbel copulas. The authors apply their methodology to birth cohort data to predict the correlated birth length and birth weight of newborns, using 36 covariates. Recently, (Wei *et al.*, 2023) propose to model time to graft failure and time to death since kidney transplantation with conditional survival copulas using parametric distributions for the marginals. They then evaluate the effects of different covariates by means of hazard ratios estimated from the copula models. One of the few examples handling censored data, (Geerdens *et al.*, 2018) use a local likelihood approach for the copula parameter with both parametrically and non-parametrically estimated marginal survival functions. The authors also propose a generalized likelihood ratio test to evaluate the constancy of the conditional copula parameter.

Most applications of these Bayesian regression copulas are, however, found in medicine, and only a few examples have found their way to the actuarial literature. Notably, (Shi et Lee, 2022) employ a copula regression framework with vine copulas to jointly estimate a policyholder' deductible, claim frequency and claim amounts.

Other propose models in-between the non-parametric and parametric frameworks. (Klein et Smith, 2021) use a copula decomposition of the joint distribution of a vector of values from a single response variable. The authors assume that the dependence structure is an unknown smooth (parametric) function of the covariates and use non-parametric estimators for the marginal distributions. They construct the copula by inverting a pseudo regression. (Liu *et al.*, 2021) propose a semi-parametric conditional mixture copula model, where the copula is a weighted average of different conditional copulas. They model not only the copula parameters but also the weights of each individual copula in the mixture based on a covariate and using a two-step semi-parametric estimation procedure. This allows to model highly flexible dependence structures. Very recently in the actuarial literature, (Wang *et al.*, 2023) apply conditional copulas to model the dependence between a pair of discrete and continuous outcomes, namely the number of claims and average claim amount observed for policyholders. They use the distribution regression approach to provide a semi-parametric estimation framework for the joint distribution of these outcomes. More precisely, they use regressions to estimate the marginal distribution of the number of claims using covariates. Knowing this, they then estimate the marginal distribution for the claim amount, conditional on the covariates and on the number of claims. Combining these two regressions, they are able to retrieve the joint conditional distribution.

Most contributions mentioned in this section so far rely on the simplifying assumption, i.e. endogenous variables are included in the model solely through the marginal distributions and they assume that the copula does not vary with the covariates. As mentioned in Section 3.1, some authors have however challenged this assumption and proposed formal tests to assess its validity in recent years. A nice overview can be found in (Derumigny et Fermanian, 2017). Among others, (Gijbels *et al.*, 2017) develop a test based on a Rao-type score statistic. Their test allows both for assessing the impact of a vector of risk factors on the dependence parameter but also on the specification of the copula model. Working with three-dimensional vine copulas, (Kraus *et al.*, 2017) investigate the differences between simplified and non-simplified models. (Spanhel et Kurz, 2019) question the simplifying assumption for pair copula constructions and explore cases where it does not hold, leading to a partial vine copula approximation. In (Spanhel et Kurz, 2022), the same authors present a test for the simplifying assumption, applicable to high-dimensional vine copulas by discretizing the support of the conditioning covariate and using a penalty in the proposed test statistic.

3.3 Parametric estimator for Archimedean copulas generators

In this section, we present a parametric model to estimate the generator function of Archimedean copulas in the presence of endogenous variables. We show how directly working with the generator rather than the copula itself relaxes the simplifying assumption by allowing both the dependence parameters and shape of dependence to vary with different levels of the covariates.

We consider the vector of times-to-events $\boldsymbol{T} = (T_1, T_2)$. Note that we present here the bivariate case for the sake of simplicity. Let $\boldsymbol{X} = (X_1, X_2)$ be the vector of censoring times such that we only observe

$$\boldsymbol{Y} = (Y_1, Y_2) = \big(\min\{T_1, X_1\}, \min\{T_2, X_2\} \big)$$

and the censoring indicators

$$\delta_j = \mathbb{1}_{[T_j \leq X_j]},$$

for $j = 1, 2$. Let $\boldsymbol{Z} \in \mathbb{R}^{n \times P}$ be a vector of risk factors, where $\boldsymbol{Z}_i = (Z_{i,1}, Z_{i,2}, ..., Z_{i,P})$ is the vector of risk factors for the i^{th} observation.

Our goal is to capture the dependence between Y_1 and Y_2 with an Archimedean copula, while including the effect of the covariates $\boldsymbol{Z}$. To do this, we consider the following non-parametric estimator of the Archi-

medean generator function $\psi^{-1}(.)$ first introduced by (Genest et Rivest, 1993) :

$$\psi^{-1}(\nu) = \exp\left\{ \int_{\nu_0}^{\nu} \frac{1}{t - K(t)} dt \right\}, \tag{3.1}$$

with $0 < \nu_0 < 1$ an arbitrarily chosen constant and where $K(\nu)$ is the univariate Kendall distribution, defined as

$$K(\nu) = \nu - \frac{\psi^{-1}(\nu)}{\psi^{-1^{(1)}}(\nu)}, \quad 0 < \nu \leq 1.$$

In the presence of censored data, (Wang et Wells, 2000) propose the following non-parametric estimator for the Kendall distribution

$$\hat{K}_n(\nu) = \int_0^{\infty} \int_0^{\infty} \mathbb{1}_{[\hat{F}(\boldsymbol{y}) \leq \nu]} d\hat{F}(\boldsymbol{y}), \tag{3.2}$$

where $\hat{F}(\boldsymbol{y})$ is an appropriate non-parametric estimator for the joint distribution of Y. A non-parametric estimator for the Archimedean generator function $\hat{\psi}_n^{-1}(.)$ can then be recovered by substituting (3.2) in (3.1).

When both T_1 and T_2 are subject to random censoring, (Akritas et Van Keilegom, 2003) propose the following estimator for the joint distribution

$$\hat{F}(\boldsymbol{y}) = w(\boldsymbol{y}) \int_0^{y_2} \hat{F}_{1|2}(y_1|\xi_2) d\tilde{F}_2(\xi_2) + (1 - w(\boldsymbol{y})) \int_0^{y_1} \hat{F}_{2|1}(y_2|\xi_1) d\tilde{F}_1(\xi_1), \tag{3.3}$$

where $\tilde{F}_1$ and $\tilde{F}_2$ are the marginal estimators of Kaplan and Meier (1958) :

$$\tilde{F}_j(y_i) = 1 - \prod_{Y_{i,j} \leq y_i, \Delta_{ij}=1} \left(1 - \frac{1}{n - i + 1}\right), \quad j = 1, 2,$$

and the weights $w(\boldsymbol{y})$ minimize the mean-squared error of $\hat{F}(\boldsymbol{y})$. The functions $\hat{F}_{1|2}(y_1|y_2)$ and, similarly, $\hat{F}_{2|1}(y_2|y_1)$, are the estimators of the conditional distributions of, respectively, Y_1 and Y_2. These are obtained using the following extension of (Beran, 1981)'s estimator :

$$\hat{F}_{1|2}(y_1|y_2) = 1 - \prod_{Y_{i,1} \leq y_1, \Delta_{i,1}=1} \left(1 - \frac{W_{ni2}(y_2; h_n)}{\sum_{j=1}^{n} W_{nj2}(y_2; h_n) \mathbb{1}_{Y_{j,1} \geq Y_{i,1}}}\right). \tag{3.4}$$

In this estimator, the values of y_2 must be uncensored and

$$W_{ni2}(y_2; h_n) = \begin{cases} \dfrac{k\left(\frac{y_2 - Y_{i,2}}{h_n}\right)}{\sum_{\Delta_{j,1}=1} k\left(\frac{y_2 - Y_{j,2}}{h_n}\right)}, & \text{if } \Delta_{i2} = 1 \\ 0, & \text{if } \Delta_{i2} = 0, \end{cases}$$

where $k(.)$ is a known kernel function and $\{h_n\}$ is the bandwidth : a sequence of positive constants such that $h_n \to 0$ as $n \to \infty$. A similar estimator is used for the conditional distribution of Y_2.

Although using Beran's estimator allows to incorporate a covariate in the model, the method suffers from a lack of interpretability of the effect of this covariate on the response. In addition, including more than one covariate can become tedious. We propose instead a parametric model for the conditional distribution of Y_1 (resp. Y_2) given Y_2 (resp. Y_1) and the vector of continuous or discrete covariates $\boldsymbol{Z}$. More specifically, we use the framework of generalized additive models for location, scale and shape (GAMLSS) to build censored regressions for both Y_1 and Y_2. The great flexibility of GAMLSS allows us to choose the best fitting distribution among a very large selection and to estimate the different parameters of the chosen distribution using linear, non-linear or smooth functions of the covariates. As such, we are able to build and compare various censored parametric models and better assess the impact of each covariate on the conditional distributions.

We thereby seek to model $(Y_{i,j}|\boldsymbol{Z}_i = \boldsymbol{z}_i, Y_{i,k} = t_{i,k})$, the observed time-to-event for observation i, with $i = 1, ..., n$ and $j = 1, 2$, knowing the vector of risk factors $\boldsymbol{Z}_i$ and the second time-to-event $t_{i,k}, k \neq j$. Given a candidate distribution $F_j(.|\boldsymbol{\theta}_j)$ with a vector of D parameters $\boldsymbol{\theta}_j$, we use GAMLSS as follows :

$$\theta_{i,j,d} = g^{-1}(\boldsymbol{z}_i'\boldsymbol{\beta}_{p,j} + \beta_j^* t_{i,k}), \tag{3.5}$$

where $g(.)$ is an appropriate link function, $\boldsymbol{\beta}_{p,j}$ is a $1 \times P$ vector of coefficients associated to covariate Z_p and linked to parameter $\theta_{j,d}$ that is estimated, and β_j^* is the specific coefficient attached to the (uncensored) activation delay of coverage k, for $k = 1, 2$ and $k \neq j$ and observation i. Following (Geerdens *et al.*, 2018), one can easily retrieve the coefficient estimates via maximum likelihood using

$$\hat{\boldsymbol{\beta}}_{j,d} = \arg\max \sum_{i=1}^{n} \delta_{i,j} \ln f_{j|\boldsymbol{Z},t_k}(y_{i,j}|\boldsymbol{Z}_i, t_{i,k}, \boldsymbol{\beta}_{p,j}) + (1 - \delta_{i,j}) \ln F_{j|\boldsymbol{Z},t_k}(y_{i,j}|\boldsymbol{Z}_i, t_{i,k}, \boldsymbol{\beta}_{p,j}), \tag{3.6}$$

where $f(.)$ and $F(.)$ are, respectively, the density and cumulative distribution functions of the chosen parametric model.

Once we select an appropriate model for Y_j, we estimate the model coefficients for the different distribution parameters using (3.6) and then obtain the parameters estimates $\hat{\boldsymbol{\theta}}_j$ for each observation i using (3.5). We can now easily obtain the estimators for the conditional distributions

$$\hat{F}_{j|\boldsymbol{Z},t_k}(y_j|\boldsymbol{Z}, t_k, \hat{\boldsymbol{\theta}}_j). \tag{3.7}$$

For a $1 \times P$ vector of covariates, the estimation of the joint distribution then becomes

$$\hat{F}(\boldsymbol{y}) = w(\boldsymbol{y}) \int_0^{z_1} \dots \int_0^{z_P} \int_0^{y_2} \hat{F}_{1|\boldsymbol{Z},t_2}(y_1|\boldsymbol{\zeta}, \xi_2, \hat{\boldsymbol{\theta}}_1) d\tilde{F}_{2|\boldsymbol{Z}}(\xi_2|\boldsymbol{\zeta}) dF_{Z_P}(\zeta_P) \dots dF_{Z_1}(\zeta_1)$$
$$+ (1 - w(\boldsymbol{y})) \int_0^{z_1} \dots \int_0^{z_P} \int_0^{y_1} \hat{F}_{2|\boldsymbol{Z},t_1}(y_2|\boldsymbol{\zeta}, \xi_1, \hat{\boldsymbol{\theta}}_2) d\tilde{F}_{1|\boldsymbol{Z}}(\xi_1|\boldsymbol{\zeta}) dF_{Z_P}(\zeta_P) \dots dF_{Z_1}(\zeta_1). \tag{3.8}$$

From (3.8), it is then possible to isolate the effects of one covariate, or of a subset of covariates, to assess their impact on the joint distribution. Suppose that we want to analyze the impact of the subset $\boldsymbol{Z}^Q = (Z_1, ..., Z_Q) \subset \boldsymbol{Z}$, with $Q < P$. We then have

$$\hat{F}(\boldsymbol{y}|\boldsymbol{Z}^Q) = w(\boldsymbol{y}) \int_0^{z_{Q+1}} \dots \int_0^{z_P} \int_0^{y_2} \hat{F}_{1|\boldsymbol{Z},t_2}(y_1|\boldsymbol{\zeta}, \xi_2, \hat{\boldsymbol{\theta}}_1) d\tilde{F}_{2|\boldsymbol{Z}}(\xi_2|\boldsymbol{\zeta}) d\breve{F}_{Z_P}(\zeta_P) \dots d\breve{F}_{Z_{Q+1}}(\zeta_{Q+1})$$
$$+ (1 - w(\boldsymbol{y})) \int_0^{z_{Q+1}} \dots \int_0^{z_P} \int_0^{y_1} \hat{F}_{2|\boldsymbol{Z},t_1}(y_2|\boldsymbol{\zeta}, \xi_1, \hat{\boldsymbol{\theta}}_2) d\tilde{F}_{1|\boldsymbol{Z}}(\xi_1|\boldsymbol{\zeta}) d\breve{F}_{Z_P}(\zeta_P) \dots d\breve{F}_{Z_{Q+1}}(\zeta_{Q+1}). \tag{3.9}$$

In (3.8) and (3.9), $\tilde{F}_{j|\boldsymbol{Z}}(y_j|\boldsymbol{Z})$ for $j = 1, 2$ are the marginal parametric estimators, fitted through a GAMLSS, and $\breve{F}_{Z_p}(Z_p)$ for $p = 1, ..., P$ are the marginal estimators for each covariate.

With this new parametric estimator for the joint distribution $\hat{F}(\boldsymbol{y})$, we propose the following estimator for the univariate Kendall distribution :

$$\hat{K}(\nu) = \int_0^{\infty} \int_0^{\infty} \mathbb{1}_{[\hat{F}(\boldsymbol{y}) \leq \nu]} d\hat{F}(\boldsymbol{y}) = \nu - \hat{\lambda}(\nu),$$

where $\hat{F}(\boldsymbol{y})$ is the estimator shown in (3.8). Similarly as before, we can isolate the effects of a subset of covariates $\boldsymbol{Z}^Q \subset \boldsymbol{Z}^P$, with $1 \leq Q \leq P$, using

$$\hat{K}(\nu|\boldsymbol{Z}^Q) = \int_0^{\infty} \int_0^{\infty} \mathbb{1}_{[\hat{F}(\boldsymbol{y}) \leq \nu]} d\hat{F}(\boldsymbol{y}|\boldsymbol{Z}^Q) = \nu - \hat{\lambda}(\nu|\boldsymbol{Z}^Q). \tag{3.10}$$

Thanks to the estimator shown in (3.10), we obtain a different Kendall distribution based on the values of the endogenous variables of interest. Finally, we obtain a parametric estimator for the Archimedean copula generator function :

$$\hat{\psi}^{-1}(\nu) = \exp\left\{ \int_{\nu_0}^{\nu} \frac{1}{t - \hat{K}(t)} dt \right\}, \tag{3.11}$$

or, with the isolated effects of a subset of covariates,

$$\hat{\psi}^{-1}(\nu|\boldsymbol{Z}^Q) = \exp\left\{ \int_{\nu_0}^{\nu} \frac{1}{t - \hat{K}(t|\boldsymbol{Z}^Q)} dt \right\}. \tag{3.12}$$

Thanks to the estimator of the Kendall distribution in (3.10), we can compute Kendall'a tau either across all covariates, as shown below

$$\hat{\tau} = 4\int_0^1 \hat{\lambda}(\nu)d\nu = 3 - 4\int_0^1 \hat{K}(\nu)d\nu,$$

or while taking into account the impact of the selected subset of risk factors, using the relation

$$(\hat{\tau}|\boldsymbol{Z}^Q) = 4\int_0^1 \hat{\lambda}(\nu|\boldsymbol{Z}^Q)d\nu = 3 - 4\int_0^1 \hat{K}(\nu|\boldsymbol{Z}^Q)d\nu. \tag{3.13}$$

Even more interestingly, in addition to being able to quantify the impact of a subset of covariates on the strength of dependence as in (3.13), using the estimator proposed in (3.12) allows the risk factors to impact the generator function and thereby, the shape of dependence. As such, we can relax the simplifying assumption when using $\hat{\psi}^{-1}(\nu|\boldsymbol{Z}^Q)$ since we do not work with the hypothesis that the structure of dependence is invariant to the effects of the covariates.

3.4 Diabetic Retinopathy Study

To illustrate the methodology laid out in Section 3.3, we consider the Diabetic Retinopathy Study (RDS) dataset used by (Geerdens *et al.*, 2018) and (Huster *et al.*, 1989). This dataset contains 197 entries, corresponding to 197 diabetic patients, for which we observe the effect of a treatment by laser photocoagulation in delaying the time to blindness. For each patient, one eye is randomly selected to remain untreated while we observe the effect of the laser treatment in the other eye. The objective is to model the dependence between the time to blindness of both eyes, as well as the impact of a covariate, namely the age at onset of diabetes, on this dependence. The times to blindness, in months, can be censored if blindness did not occur while the patient was observed in the study.

For each entry i in the dataset, for $i = 1, ..., 197$, we observe the following vector : $(Y_{i,1}, Y_{i,2}, \delta_{i,1}, \delta_{i,2}, Z_i)$. $Y_{i,1}$ (resp. $Y_{i,2}$) is the time to blindness for the treated (resp. untreated) eye, $\delta_{i,1}$ (resp. $\delta_{i,2}$) is the censoring indicator for the treated (resp. untreated) eye, equal to 0 if the observation is censored, i.e. if the study has stopped before the onset of blindness in that eye. The time to blindness in the treated eye is censored for 73% of the 197 patients observed in the dataset, against only 49% for the untreated eye. The covariate Z_i represents the age at onset of diabetes (in years) for the individual, ranging from 1 to 58 years old.

3.4.1 Dependence modeling without covariate

We first analyze the dependence structure present in this dataset without taking the covariate Z into account. Assuming that we can model the data with an Archimedean copula, we find an estimator of the generator function $\psi^{-1}(.)$ and use it to estimate Kendall's tau. We compare two approaches : first, we find the estimator $\hat{\psi}_n^{-1}(\nu)$ non-parametrically, combining the method proposed in (Genest et Rivest, 1993) in (3.1) and (Beran, 1981)'s estimators for the conditional marginal distributions, as in (Akritas et Van Keilegom, 2003) in (3.3). We use Epanechnikov kernels and select the optimal bandwidth by cross-validation for Beran's estimators. Second, we replace these non-parametric estimators by parametric models for the marginal distributions of the times to blindness. As such, we estimate the joint distribution using (3.8), without the terms related to the vector of covariates $\boldsymbol{Z}$. Following (Geerdens *et al.*, 2018), we opt for Weibull marginal distributions, with parameter vectors $\boldsymbol{\theta}_j = (\theta_{j,1}, \theta_{j,2})$ for $j = 1, 2$, where $\theta_{j,1}$ and $\theta_{j,2}$ are, respectively, the scale and shape parameters for the time to blindness in eye j. The joint distribution is then given by

$$\hat{F}(\boldsymbol{y}) = w(\boldsymbol{y}) \int_0^{y_2} \hat{F}_{1|2}(y_1|\xi_2) d\tilde{F}_2(\xi_2) + (1 - w(\boldsymbol{y})) \int_0^{y_1} \hat{F}_{2|1}(y_2|\xi_1) d\tilde{F}_1(\xi_1),$$

where

$$\hat{F}_{1|2}(y_{i,1}|\xi_{i,2}) = 1 - \exp\Big[- \Big(\frac{y_{i,1}}{\hat{\theta}_{i,1,1}} \Big)^{\hat{\theta}_{i,1,2}} \Big],$$

with

$$\hat{\theta}_{i,1,1} = \exp(\hat{\beta}_{0,1} + \hat{\beta}_1^* \xi_{i,2}),$$

the scale parameter for the Weibull distribution, where $\hat{\beta}_{0,1}$ is the intercept of the model for parameter $\hat{\boldsymbol{\theta}}_1$ and $\hat{\beta}_1^*$ is the coefficient linked to ξ_2. For the shape parameter, we have

$$\hat{\theta}_{i,1,2} = \exp(\hat{\gamma}_{0,1} + \hat{\gamma}_1^* \xi_{i,2}),$$

where $\hat{\gamma}_{0,1}$ is the intercept of the model for parameter $\hat{\boldsymbol{\theta}}_2$ and $\hat{\gamma}_1^*$ is the coefficient linked to ξ_2. $\tilde{F}_2(y_2)$ is the parametric marginal estimator of y_2. We proceed similarly for $\hat{F}_{2|1}(.)$.

Figure 3.1 shows the estimated Kendall distribution $\hat{K}(.)$ (left) and, equivalently, $\hat{\lambda}(.)$ (right), obtained using both the non-parametric (black curves) and parametric (grey curves) approaches from, respectively, (3.1) and (3.11).

We observe that the results of both methods are very close. The curves of the $\hat{\lambda}(.)$ functions on the right plot display similar shapes, indicating that the dependence structure is the same with both estimations,

which was to be expected. The resulting values for Kendall's tau are $\hat{\tau}_{\text{NP}} = 0.1864$ and $\hat{\tau}_{\text{P}} = 0.1859$ for, respectively, the non-parametric and the parametric models. The estimated strength of dependence is thus also almost identical. Since we do not include any risk factor in the model at this stage, we expect both approaches to give similar results.

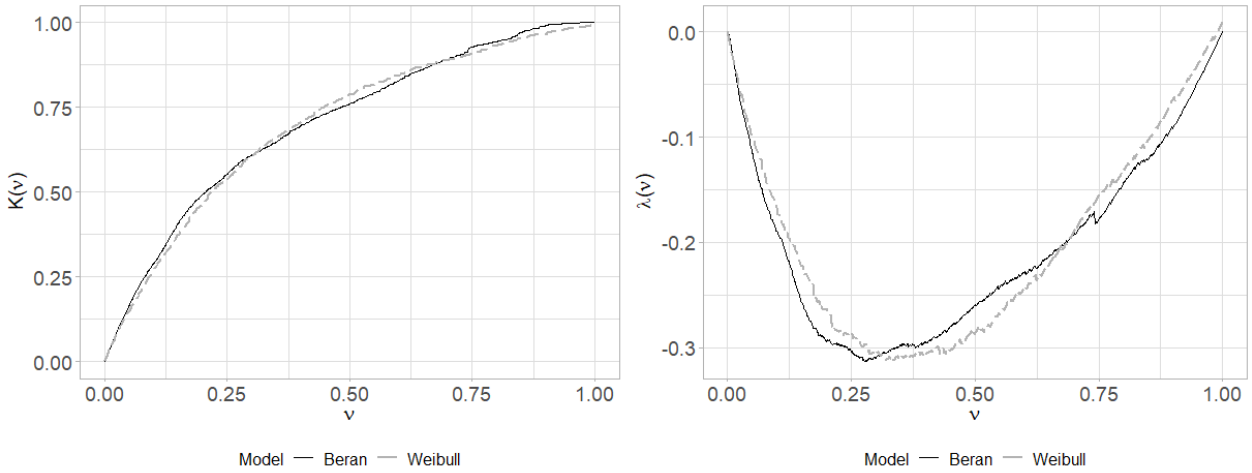

Figure 3.1 – $\hat{K}(\nu)$ and $\hat{\lambda}(\nu)$ with the non-parametric (black lines) and parametric (gray dashed lines) approaches for the RDS data.

3.4.2 Dependence modeling with a covariate

We now enter the age at onset of diabetes as covariate in the parametric model. The joint distribution thus becomes

$$\hat{F}(\boldsymbol{y}) = w(\boldsymbol{y}) \int_0^z \int_0^{y_2} \hat{F}_{1|Z,t_2}(y_1|\zeta,\xi_2) d\tilde{F}_{2|Z}(\xi_2|\zeta) dF_Z(\zeta) + (1 - w(\boldsymbol{y})) \int_0^z \int_0^{y_1} \hat{F}_{2|Z,t_1}(y_2|\zeta,\xi_1,) d\tilde{F}_{1|Z}(\xi_1|\zeta) dF_Z(\zeta),$$

with

$$\hat{F}_{1|Z,t_2}(y_{i,1}|\zeta_i,\xi_{i,2}) = 1 - \exp\Big[- \Big(\frac{y_{i,1}}{\hat{\theta}_{i,1,1}} \Big)^{\hat{\theta}_{i,1,2}} \Big],$$

where

$$\hat{\theta}_{i,1,1} = \exp(\hat{\beta}_{0,1} + \hat{\beta}^*_{1,1}\xi_{i,2} + \hat{\beta}_1\zeta_i)$$

and

$$\hat{\theta}_{i,1,2} = \exp(\hat{\gamma}_{0,1} + \hat{\gamma}^*_{1,1}\xi_{i,2} + \hat{\gamma}_1\zeta_i).$$

For $\hat{F}_{2|Z,t_1}$, we have

$$\hat{F}_{2|Z,t_1}(y_{i,2}|\zeta_i, \xi_{i,1}) = 1 - \exp\Big[- \Big(\frac{y_{i,2}}{\hat{\theta}_{i,2,1}} \Big)^{\hat{\theta}_{i,2,2}} \Big],$$

where

$$\hat{\theta}_{i,2,1} = \exp(\hat{\beta}_{0,2} + \hat{\beta}^*_{1,2}\xi_{i,1} + \hat{\beta}_2\zeta_i)$$

and

$$\hat{\theta}_{i,2,2} = \exp(\hat{\gamma}_{0,2} + \hat{\gamma}^*_{1,2}\xi_{i,1} + \hat{\gamma}_2\zeta_i).$$

Figure 3.2 shows again the curves of the estimated Kendall distribution on the left and of $\hat{\lambda}(\nu)$ on the right, this time for the parametric model with (black dotted line) and without (grey dashed lines) the covariate. We observe a clear jump in the curve, corresponding to a jump in the value of Kendall's tau, from $\hat{\tau} = 0.1859$ in the model without the age at onset of diabetes to $\hat{\tau} = 0.3001$ when including this variable in the joint distribution of the times to blindness.

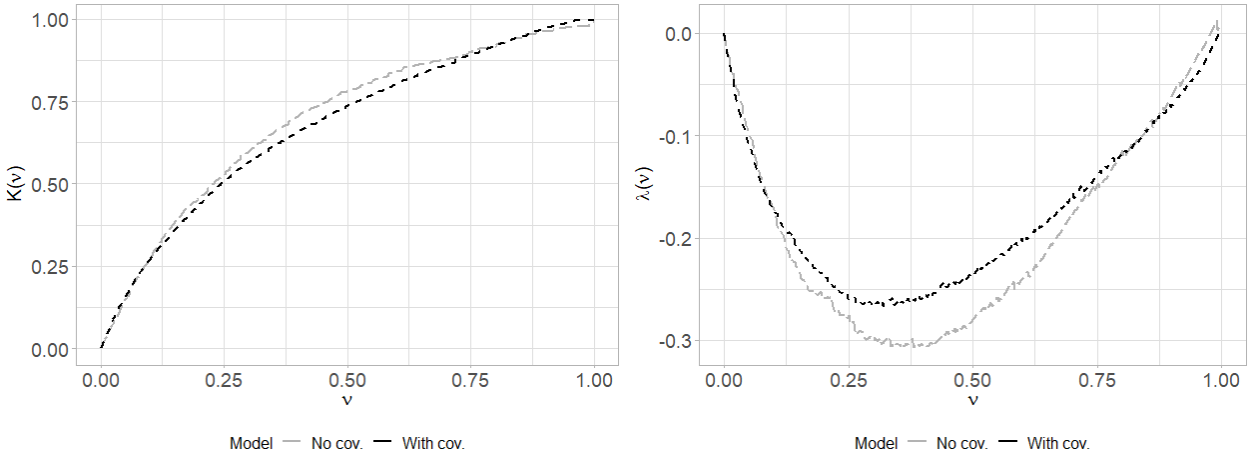

Figure 3.2 – $\hat{K}(\nu)$ and $\hat{\lambda}(\nu)$ with (black dotted lines) and without (gray dashed lines) covariate for the RDS data, under the parametric model.

This risk factor thus has an important impact on the strength of dependence when considering all the individuals in the dataset. We now investigate the impact of different values of the covariate on the dependence model. Using (3.9), we could obtain a separate estimator of the generator function for each different value of the age at onset of diabetes, that is a different curve for the Kendall distribution for each age. For illustrative purposes, we analyze the impact for individuals who were younger and older than 20 years old at

onset of diabetes. The choice of this splitting point is, at this stage, only motivated by the results derived in (Geerdens *et al.*, 2018), where the authors show that with the inclusion of the covariate in the model, the average estimate of Kendall's tau for the whole dataset is around 0.3. It is below (resp. above) this value for individuals who were first diagnosed with diabetes before (resp. after) the age of 20. This splitting point also separates the dataset into two subsets of roughly the same size. We will further discuss the choice of the splitting point for the conditioning variable in Section 3.6.

Figure 3.3 compares the curves of the Kendall distribution and the lambda function for all individuals, i.e. $\hat{K}(\nu)$, (in black) and for those who were younger or older than twenty at onset of diabetes (in grey), i.e. respectively, $\hat{K}(\nu|Z \leq 20)$ and $\hat{K}(\nu|Z > 20)$. The black curve for all individuals is the same as the one depicted in Figure 3.2. Interestingly, we observe that the shapes of the three curves on both plots are quite different, indicating that conditionally on the value of the covariate, both the strength and the shape of dependence vary.

We first show the differences on the strength of dependence via Kendall's tau in Table 3.1. While the estimated value for the population as a whole is around 30%, it it slightly smaller at 26.3% for individuals who were first diagnosed with diabetes before turning twenty, and quite higher at almost 56% for individuals who started suffering from diabetes after the age of twenty. These results are in line with those presented in (Geerdens *et al.*, 2018).

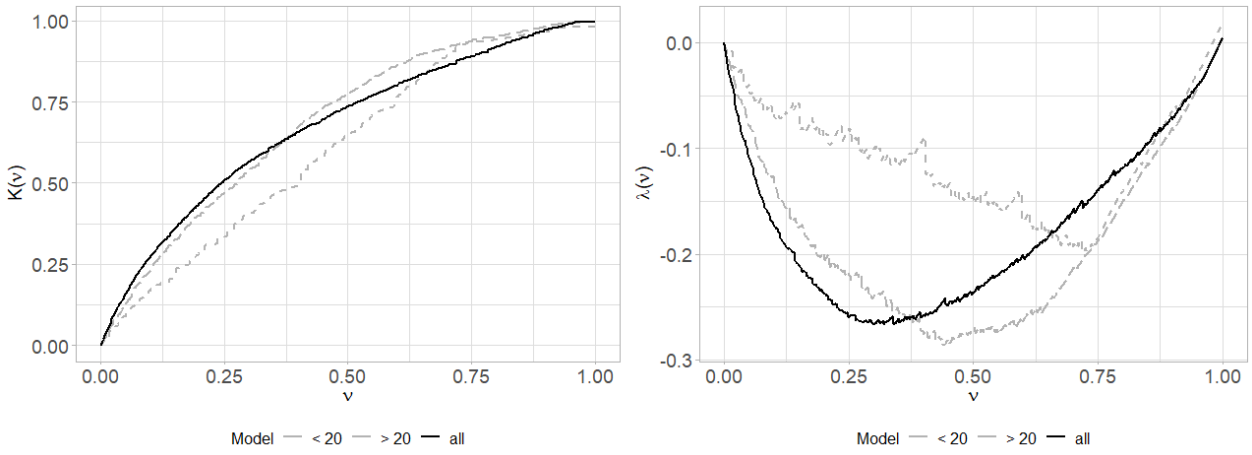

Figure 3.3 – Plots of $\hat{K}(\nu)$ (black continuous curve), $\hat{K}(\nu|Z \leq 20)$ (grey dashed curve) and $\hat{K}(\nu|Z > 20)$ (grey dotted curve) on the right and $\hat{\lambda}(\nu)$, $\hat{\lambda}(\nu|Z \leq 20)$ and $\hat{\lambda}(\nu|Z > 20)$ on the left.

Table 3.1 – Kendall's tau for different values of the age covariate Z

Model	$\hat{\tau}$
All	0.3001371
$Z \leq 20$	0.2629644
$Z > 20$	0.5591597

Next, we observe the impact of these two different age groups on the shape of dependence. From Figure 3.3, the shape of the $\hat{\lambda}(.)$ curve for the dataset as a whole points towards the bottom left corner of the plot, depicting the typical shape of a Gumbel or a Joe copula. The curve for the patients whose disease started before the age of 20 points more towards the center or the bottom right corner of the plot, indicating that a Frank or a Clayton copula might be more fitted for this age group. To determine which model best fits these two populations, we perform simulations from the estimated generator functions and apply the L^2-norm model validation method used in, among others, (Lakhal-Chaieb, 2010). We use the `RPLATRANS` algorithm described in (Ridout, 2009), together with the Marshall, Olkin algorithm from (Marshall et Olkin, 1988). The `RPLATRANS` algorithm allows to sample from a distribution specified by its inverse Laplace-Stieltjes transform. Knowing this, we can use the Marshall, Olkin algorithm to simulate bivariate vectors of observations (U_1, U_2) from the Archimedean copula from which the data originates. More details on these algorithms are provided in Appendix C.1. More specifically, we perform the simulation procedure presented below.

In Algorithm 3, the pseudo p-value p_m can be interpreted as the percentage of simulated samples for which candidate model m presents the smallest distance between its Kendall distribution and that of the data. The copula with the largest pseudo $p-$value will be the most appropriate model.

Figure 3.4 shows the results of $J = 1000$ simulations for the whole dataset (left) and for the individuals who had diabetes before the age of twenty. The black curves correspond to the $\hat{\lambda}(\nu)$ estimated parametrically from the original data. The dashed grey curves represent the average of the simulations and the shaded areas show the 95% confidence intervals.

Algorithm 3: Copula simulation and selection procedure

1: For j in $1, ..., J$ where J is the total number of simulations performed, simulate a new bivariate sample of size n from the parametric estimator $\hat{\psi}(\nu)$, using the `RLAPTRANS` and Marshall, Olkin algorithms : $(\boldsymbol{Y}_1^{(j)}, \boldsymbol{Y}_2^{(j)}) = \{(Y_{1,1}^{(j)}, Y_{1,2}^{(j)}), ..., (Y_{n,1}^{(j)}, Y_{n,2}^{(j)})\}$.

2: For each new simulated sample (j), compute again the Kendall distribution $\hat{K}^{(j)}(\nu)$ and estimate $\hat{\tau}^{(j)}$.

3: For the M candidate copula models under consideration, use $\hat{\tau}^{(j)}$ to estimate the dependence parameter $\hat{\alpha}_m$, for $m = 1, ..., M$. For each candidate model, estimate $K_{\hat{\alpha}_m}^{(j)}(\nu)$.

4: Calculate the L^2-norm between $K_{\hat{\alpha}_m}^{(j)}(\nu)$ and $\hat{K}^{(j)}(\nu)$ as

$$D^{(j)}(\hat{\alpha}) = \int_{\xi}^{1} \left(\hat{K}^{(j)}(\nu) - K_{\hat{\alpha}_m}^{(j)}(\nu)\right)^2 d\nu.$$

5: For each candidate model m, obtain the pseudo $p-$value as

$$p_m = \frac{1}{J} \sum_{j=1}^{J} \mathbb{1}[\min_{l} D^{(j)}(\hat{\alpha}_l) > D^{(j)}(\hat{\alpha}_m)],$$

with $l, m = 1, ..., M$ and $l \neq m$.

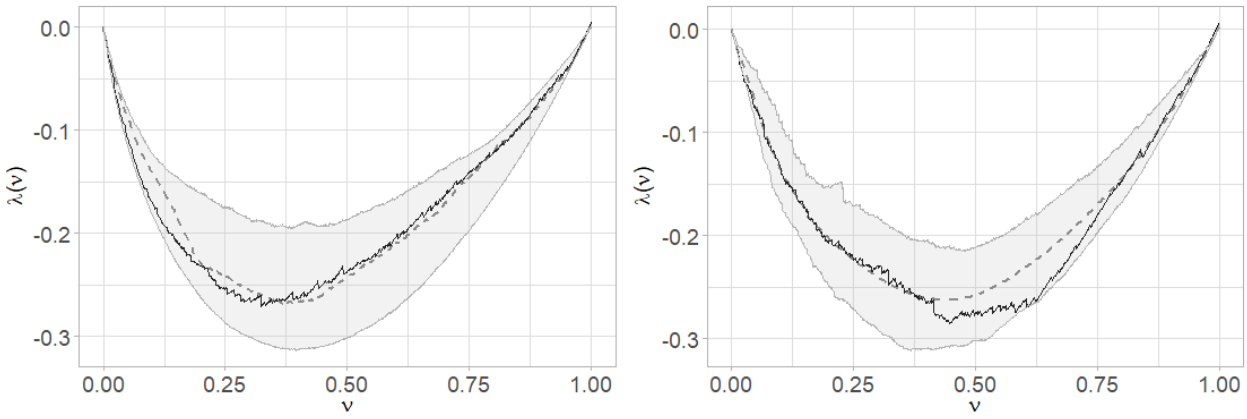

Figure 3.4 – $\lambda(.)$ for the samples simulated using the generator functions for the whole population (left) and for the patients who were first diagnosed with diabetes before the age of twenty (right).

Table 3.4 shows which copula models among the Clayton, Frank, Gumbel and Joe were selected with the validation procedure for each simulation. Note that the results are presented here in terms of (1-p-values). The best fitted model is then the one with the smallest value displayed in the table. For the subset of individuals whose diabetes started before the age of twenty, 460 out of 1000 simulations select the Clayton

copula as the most appropriate model, 305 simulations prefer the Frank copula and 235 opt for the Gumbel model. None of the simulations result in the Joe copula being selected as best fitted to this subset of the data. For the data as a whole, only 138 out of the 1000 simulations select the Clayton copula as best fitted. Most simulations, that is 464 out of the 1000, opt for the Frank model, 354 select the Gumbel copula and 41 choose the Joe model. The Frank copula thus appears to be best fitted for the whole data, followed by the Gumbel copula while for the subset of patients who were diagnosed with diabetes before turning twenty, the Clayton copula is the preferred model, followed by the Frank copula in second place. The age at onset of diabetes thereby seems to have an impact not only on the strength of dependence as discussed earlier, but also on the shape of dependence. When modeling the correlated times to blindness in both eyes, one might want to use different copula models for different groups of patients.

Note, however, that the pseudo p-values in Table 3.4 do not lead to a clear decision, particularly for the Clayton, Frank and Gumbel that present relatively similar values. More specifically, with the high levels of right-censoring observed in the data, copulas such as the Clayton and Frank may more closely resemble each other, leading to uncertainty in the choice of the most appropriate model. In their paper, although (Geerdens *et al.*, 2018) only use the covariate for modeling the dependence parameter, they implement a generalized likelihood ratio test to determine whether the assumed copula model is adequate. The test leads them to similar conclusions as observed here. However, since our approach leads to an estimator of the copula generator, we do not need to specify a known copula family, contrarily to other approaches found in the literature so far. Our method allows us to conclude that the conditioning variable impacts both the level and shape of dependence between the times to blindness. Even if we can not define specific copula families with a sufficient level of confidence for different levels of the covariate, we can directly work from the estimated generator functions obtained in the different cases and as such, acknowledge that the copula model varies with the conditioning variable.

Table 3.2 – Results from the copula selection procedure using the L^2-norm.

Z	Clayton	Frank	Gumbel	Joe
≤ 20	0.540	0.695	0.765	1.000
All	0.862	0.536	0.646	0.959

3.5 Automobile insurance dataset application

In this section, we apply the non-parametric estimator for Archimedean copula generators presented in Section 3.3 to a loss reserving context, using a Canadian automobile insurance dataset.

The data includes over $600\ 000$ claims that occurred between 2015 and 2021. Each of these claims relate to a policy in force under which four insurance coverages are provided : the Accident Benefits, Bodily Injury, Vehicle Damage and Loss of Use coverages. Note that, to simplify the presentation, we will work in a bivariate framework and only consider the Accident Benefits and Bodily Injury coverages that are the most important cost-wise in the portfolio.

From its moment of declaration to the insurer, each new claim will impact at least one of these coverages. We define the *activation delay* of a claim for a coverage as the time elapsed between the moment of declaration of the claim and the first moment at which the insurer realizes that this claim triggers the coverage and registers this information in his claims management system. For illustrative purposes, consider a claim declared on June 1^{st}, 2016. On the same day, based on the information available at the time, the insurer labels it as an Accident Benefits claim. The activation delay for this claim under this coverage is then equal to one day. On June 25^{th}, the insurer receives new information leading to label the claim as a Bodily Injury claim as well. The activation delay for this claim under the Bodily Injury coverage is thereby 25 days.

Considering that not all claims impact all coverages, the activation delays are subject to right-censoring. This occurs when, for a given claim that has reached its settlement date, a coverage has not been triggered yet. The censoring variable is thus the settlement delay, defined here as the time elapsed between the declaration and settlement dates of a claim. In this portfolio, 21.63% of all claims impact both the Accident Benefits and Bodily Injury coverages. These are the claims for which neither of the activation delays are censored. For 23.46% of claims, only the Bodily Injury coverage is triggered. This corresponds to the percentage of claims for which we observe a censored value for the Accident Benefits coverage. The remaining 54.91% of claims only activate the Accident Benefits delay, i.e. these are the claims for which the activation delays of the Bodily Injury coverage are censored.

In addition to the coverages, the dataset provides for each claim information related to the policyholder, the vehicle driven or the claim itself. Our goal is to model the dependence between the coverages provided within a single policy by means of their activation delays and the impact of some risk factors on this depen-

dence, using the framework of Archimedean copulas. More specifically, we focus on the decade of birth of the policyholders, which is provided as a covariate in the dataset. Figure 3.5 shows more information about this risk factor. We choose to create a new categorical variable with three levels, grouping the different decades of birth in three groups of approximately equal sizes. Let Z_i be the value of this new covariate for claim i, such that

$$Z_i = \begin{cases} 1, & \text{if the policyholder was born in or after 1990} \\ 2, & \text{if the policyholder was born between 1970 (incl.) and 1989} \\ 3, & \text{if the policyholder was born in or before 1969.} \end{cases}$$

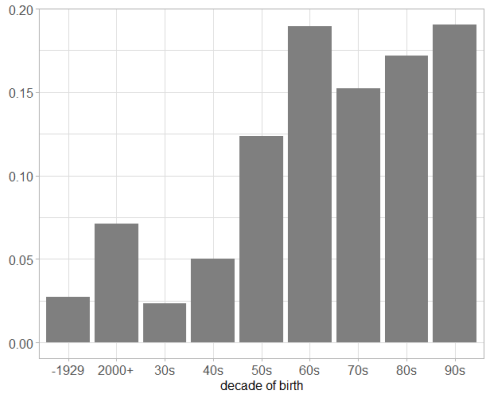

Figure 3.5 – Frequency of the different decades of birth of individuals in the dataset.

We will use this covariate as the conditioning variable in our copula model to investigate the possible effects of the age of policyholders on the strength and structure of dependence between the activation delays. We use the following notation :

- $\boldsymbol{T}_i = (Y_{i,1}, Y_{i,2})$ is the vector of activation delays for our two coverages, for claim i with $i = 1, ..., n$.
- $\boldsymbol{X}_i = (X_{i,1}, X_{i,2})$ is the vector of censoring times for both coverages and for claim i. This vector corresponds in this case to the settlement delay of the claim.
- $\boldsymbol{Y}_i = (Y_{i,1}, Y_{i,2}) = (\min(T_{i,1}, X_{i,1}), \min(T_{i,2}, X_{i,2}))$ is the vector of observed activation delays for claim i, taking censoring into account.
- $\delta_{i,j} = \mathbb{1}_{(T_{i,j} \leq X_{i,j})}$ for $j = 1, 2$ are the censoring indicators for claim i.
- Z_i is the age of the policyholder upon occurrence of claim i, belonging to one of three groups previously defined.

For illustration purposes, suppose that a policyholder born in the 1980s files a claim following a car accident. His claim only triggers the Accident Benefits coverage upon reporting, i.e. with a delay of one day, and settles 90 days later, without activating further coverages. The data entry related to this claim i is $(Y_{i,1}, Y_{i,2}, \delta_{i,1}, \delta_{i,2}, Z_i) = (1, 90, 1, 0, 2)$.

3.5.1 Dependence modeling without covariate

Like we did for the diabetic retinopathy study, we first model the dependence between the censored activation delays without conditioning on the covariate. We compare the results derived from a non-parametric model using Beran's estimators in the joint distribution to those derived from a parametric model where we replace Beran's estimators by censored parametric regressions. Similarly to what we did in Section 3.4, we choose Weibull distributions to model the marginal distributions of both activation delays.

Figure 3.6 shows the estimated Kendall distributions and corresponding $\lambda(.)$ functions with both approaches. Both give almost identical results in the shapes of dependence. The strenght of dependence, captured via the estimated Kendall's tau is also almost identical with both models. For the non-parametric model, with obtain $\hat{\tau} = 0.2534$ and for the parametric model, $\hat{\tau} = 0.2569$.

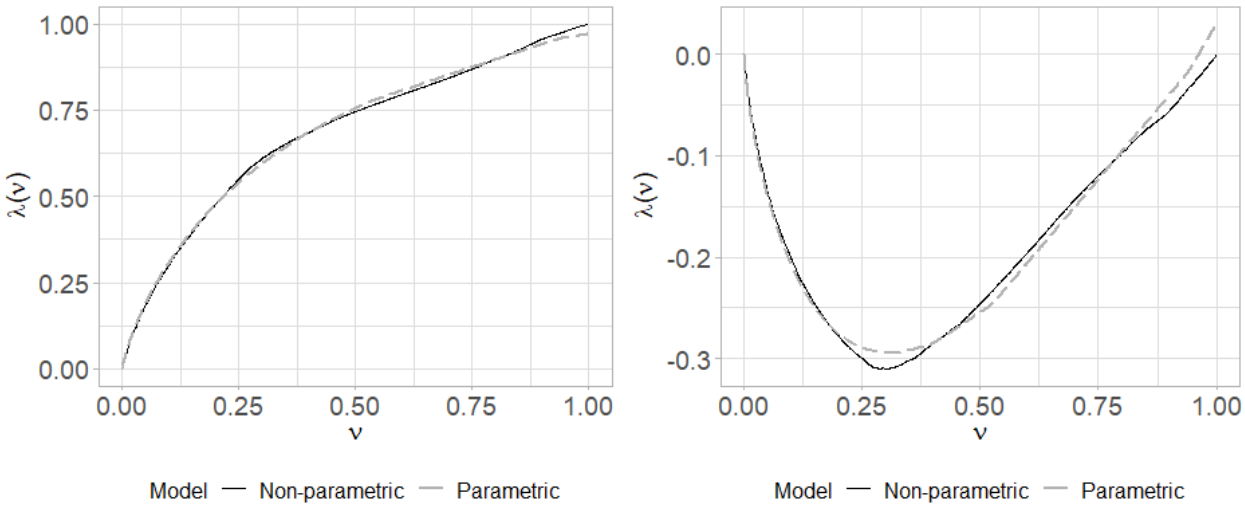

Figure 3.6 – $K(\nu)$ with Beran's estimator (black lines) and GAMLSS (blue dashed lines) for all data, with censored delays set as the settlement delays.

3.5.2 Dependence modeling with a covariate

We now focus on the parametric model only and use Z as the conditioning variable. We estimate four univariate Kendall distributions : one for the dataset as a whole, $\hat{K}(\nu)$, one for the subset of the data in which policyholders were born after 1990, $\hat{K}(\nu|Z = 1)$, one in which they were born between 1970 and 1989, $\hat{K}(\nu|Z = 2)$, and one in which they were born before 1969, $\hat{K}(\nu|Z = 3)$. We use again Weibull censored regressions for the marginal distributions of the activation delays.

Figure 3.7 shows the plot of these four Kendall distribution curves on the left, and the corresponding curves for the $\lambda(.)$ function on the right. The black continuous curves depict, respectively, $\hat{K}(\nu)$ and $\hat{\lambda}(\nu)$, the dark grey dashed curves stand for $\hat{K}(\nu|Z = 2)$ and $\hat{\lambda}(\nu|Z = 2)$, the dark grey dotted curves represent $\hat{K}(\nu|Z = 1)$ and $\hat{\lambda}(\nu|Z = 1)$, and the light grey curves show $\hat{K}(\nu|Z = 3)$ and $\hat{\lambda}(\nu|Z = 3)$. The simplifying assumption clearly does not hold here, as different values of the age of policyholders result in different copulas.

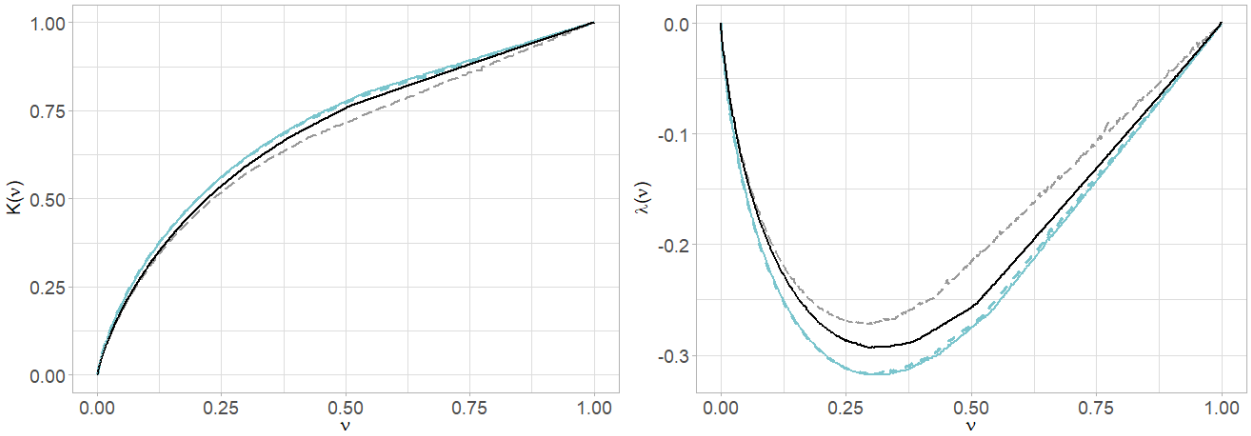

Figure 3.7 – $K(\nu)$ and $\lambda(\nu)$ when conditioning on Z.

Table 3.3 displays the estimated values of Kendall's tau for the data as a whole and for the different subsets. While conditioning on the age group of the policyholders does not appear to have a strong impact for the population as a whole, the results are quite different when we consider the different values of Z separately. While the dependence seems to decrease for the younger and older policyholders in the portfolio, it increases from 25.81% when considering all individuals to 33.8% for the individuals born between 1970 and 1989.

Table 3.3 – Kendall's tau for different values of the covariate

Model	$\hat{\tau}$
All	0.2581
$Z = 1$	0.1919
$Z = 2$	0.3379
$Z = 3$	0.1813

Next, we apply the same simulation procedure as the one described in Section 3.4 to investigate the best copula model in each scenario. For the complete population and for each of the three sub-groups of policyholders, we perform simulations from the estimators of the generator functions $\hat{\psi}(\nu)$, $\hat{\psi}(\nu|Z = 1)$, $\hat{\psi}(\nu|Z = 2)$ and $\hat{\psi}(\nu|Z = 3)$ using the `RLAPTRANS` algorithm. For each simulated sample, we select the most fitting model among the Clayton, Frank, Gumbel and Joe copulas by searching for the minimal L^2-norm. The graphical results of 1000 of these simulations are presented in Figure 3.8. In these plots, the black continuous curves are identical to those from Figure 3.7, i.e. they represent the original $\hat{\lambda}(.)$ functions estimated from the data, for the whole population (top left) and when conditioning on the three different levels of the covariate Z. The grey dashed curves show the average of the simulations and the grey areas represent the 95% confidence intervals. On average, using simulations from the generator functions in each case produces results that are very similar to the original data.

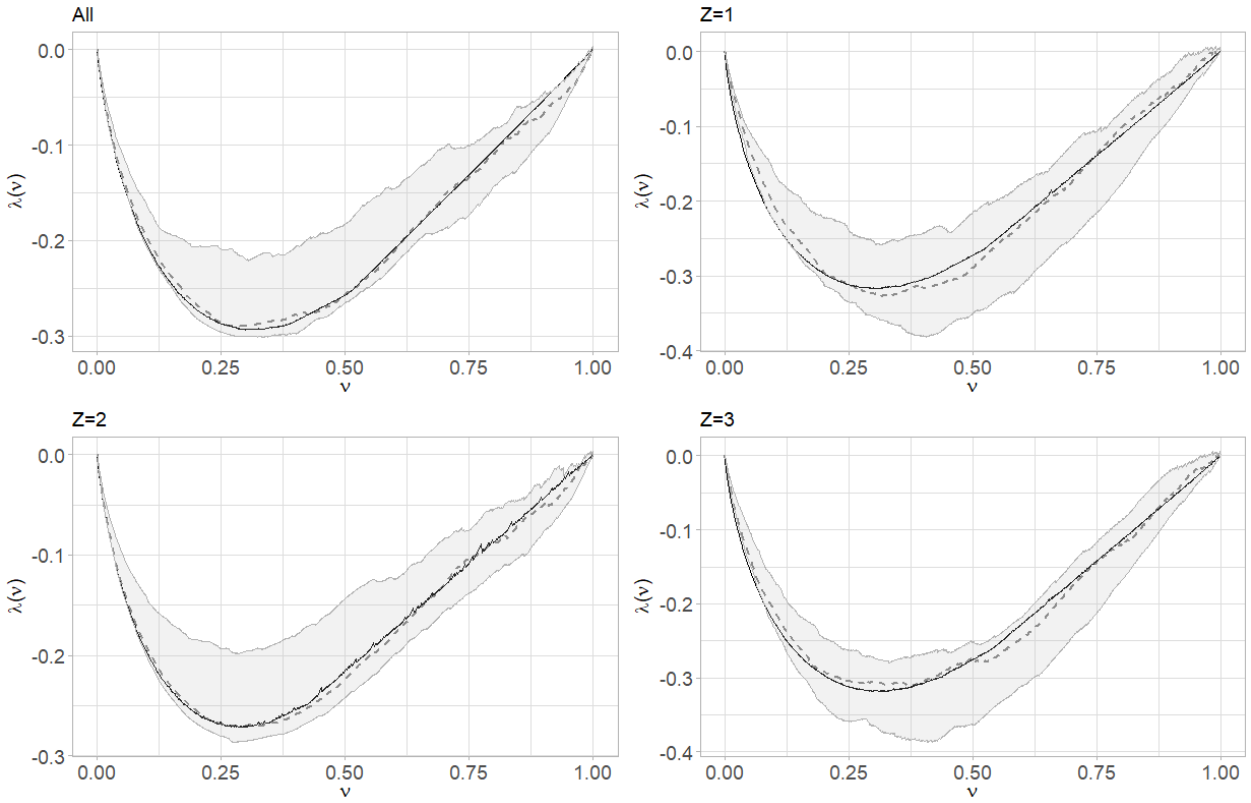

Figure 3.8 – Estimated generators (continuous curve) for the different levels of the covariates, and average results of new simulations performed from these generators (dashed curve), as well as 95% confidence intervals for these simulations (grey shaded areas).

The shapes of the curves in Figure 3.8 seem to resemble each other. This indicates that, although the different age groups of policyholders clearly have an impact on the strength of dependence, the shape seems to remain stable. We see that it is not entirely true when looking at the results displayed in Table 3.4. For each conditioning scenario, this table shows the 1-p-values obtained from performing the simulation procedure described in Section 3.4 1000 times. The copula model for which the displayed value is smallest is the one that was selected as best fitting by most simulations. We also present here the average estimate of Kendall's tau over all simulations for each conditioning scenario.

For the population as a whole and for each of the three age groups, the Joe copula appears to be the best fitting model. For all policyholders, 35 out of 1000 simulations reject and Joe copula as best fitting model and opt instead for the Gumbel copula. For the policyholders for which $Z = 2$, i.e. born between 1970 and 1989, only three of the thousand simulations reject the Joe model and favor the Gumbel copula. For the

whole population and for this sub-group, none of the simulations select the Frank or Clayton copulas. The Joe model comes out as the largely favoured choice in these two cases.

It is however not as evident for the younger and older policyholders. Although most simulations in both these cases still favor a Joe copula, there seems to be more variability in the choice. For the younger policyholders, that is those for whom $Z = 1$, 384 simulations out of a thousand reject the Joe copula, i.e. ten times more than for the population as a whole and a hundred times more than for the policyholders born between 1970 and 1989. Among these 384 simulations, 344 opt instead for a Gumbel copula, 27 for a Frank copula and the remaining 13 for a Clayton model. For the older policyholders, 266 simulations reject the Joe model. They opt for the Gumbel (233), Clayton (20) and Frank copulas (13). Although the Joe copula still appears best suited for both sub-groups, almost 40% and 27% of simulations reject it for, respectively, the younger and older policyholders. This clearly shows that different levels of the covariate have an impact on the dependence structure of the data.

Table 3.4 – Results from the copula selection procedure using the L^2-norm.

Z	Clayton	Frank	Gumbel	Joe	$\bar{\tau}$
All	1.000	1.000	0.965	0.035	0.2512
$Z = 1$	0.987	0.973	0.656	0.384	0.1925
$Z = 2$	1.000	1.000	0.997	0.003	0.3451
$Z = 3$	0.980	0.987	0.767	0.266	0.1882

Next, we illustrate the results of using the four different approaches described below to predict the activation delays for both coverages.

— Approach 1 : we do not condition on the different levels of Z and use the results displayed in Table 3.4, that is a Joe copula with Kendall's tau set to 0.2512, to perform predictions for the activation delays of the population as a whole.

— Approach 2 : we condition on the three levels of the covariate and use again the results from Table 3.4 to predict the activation delays for the different sub-groups of policyholders. We thereby use Joe copulas with Kendall's tau equal to 0.1925, 0.3451 and 0.1882 for, respectively, the younger, middle-aged and older policyholders.

— Approach 3 : we do not condition on the different levels of Z and obtain predictions by simulating directly from the generator $\hat{\psi}(\nu)$ using the `RLAPTRANS` algorithm.

— Approach 4 : we condition on the three levels of Z and obtain predictions by simulating directly from the three generators $\hat{\psi}(\nu|Z=1)$, $\hat{\psi}(\nu|Z=2)$ and $\hat{\psi}(\nu|Z=3)$, using again the `RLAPTRANS` algorithm.

While approaches 1 and 2 use known copulas, approaches 3 and 4 do not impose a specific model. Figure 3.9 shows the simulated densities for the activation delays of the Accident Benefits (left) and Bodily Injury (right) coverages. On each plot, the black curves represent the observed densities. The grey continuous (resp. dashed) curves show the densities resulting from approach 1 (resp. approach 2) and the blue continuous (resp. dashed) curves display the densities resulting from approach 3 (resp. approach 4). The vertical lines represent the observed average activation delays (in black) and average simulated delays for the corresponding approaches. These values are also displayed in Table 3.5.

We first observe on the plots of Figure 3.9 that the simulated densities of approaches 1 and 2 are quite similar and that the same goes for approaches 3 and 4. This is particularly true for the Accident Benefits coverage where the two grey (resp. blue) curves almost entirely overlap. For the Bodily Injury coverage, the blue curves, corresponding to the densities simulated directly from the generator functions, appear to be closer to the true density of the activation delays.

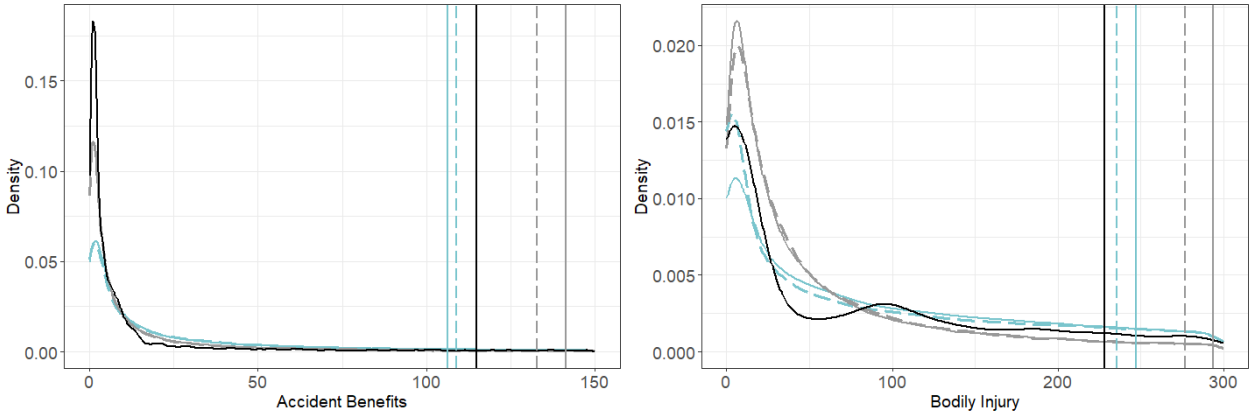

Figure 3.9 – Observed and simulated densities of the activation delays for both coverages, using the four different approaches. The vertical lines stand for the average activation delays.

We now look at the average values represented both by the vertical lines on the plots of Figure 3.9 and by the values in Table 3.5. For the approaches using Joe copulas, namely approaches 1 and 2, conditioning on

the age group of the policyholders appears to provide predictions that are closer to the observed activation delays. For the Accident Benefits coverage, approach 1 provides the highest prediction with an average delay of 141.47 days, while when conditioning on the covariate Z in approach 2, the average predicted delay decreases to 132.65 days, closer to the observed 114.81 days. We observe similar patterns for the Bodily Injury coverages. Approaches 3 and 4 show comparable results, although the difference between them is slightly less marked. For the Bodily Injury coverage, using the generator function for all policyholders provides an average estimated delay of 247.05 days while using the generator functions for the three different age groups results in an average delay of 235.59 days. For the Accident Benefits coverages, approach 3 results in an average estimated delay of 106.23 days, while it is 108.86 days for approach 4.

For both the Accident Benefits and Bodily Injury coverages, the observed activation delays are, respectively, 114.81 and 228.39 days. We can thereby conclude this section with the two following interesting observations. First, with both sets of approaches, namely approaches 1 and 2 based on the Joe copulas and approaches 3 and 4 based on the generator functions, the predictions closest to the true activation delays are obtained when conditioning on the age group of the policyholders. Second, using the generator functions rather than the Joe copulas and thereby not imposing a known model seems to lead to closer average predictions.

Table 3.5 – Average activation delays (observed and simulated).

Coverage	Observed	Approach 1	Approach 2	Approach 3	Approach 4
Accident Benefits	114.81	141.47	132.65	106.23	108.86
Bodily Injury	228.39	293.66	276.75	247.05	235.59

3.6 Conclusion

In this paper, we present a new parametric estimator for the generator function of Archimedean copulas that is suitable for censored data. By incorporating covariates in the estimator, we allow both the strength and shape of dependence to vary with different values of the endogenous variables, thereby bypassing the simplifying assumption often used in conditional copula models. Our model allows to capture the effects of some covariates of interest simply by conditioning the joint distribution on them. We demonstrate its performance in two different applications.

First, we study the diabetic retinopathy dataset and model the dependence between the times to blindness in both eyes for diabetic patients, as well as the effects of the age at onset of diabetes on this dependence. We do this for the patients population as a whole then investigate the model for different values of the covariate. In line with the results presented in (Geerdens *et al.*, 2018), we find that different ages at onset of diabetes lead to different estimates of Kendall's tau, the level of correlation being lower (resp. higher) for patients who were diagnosed with diabetes at a younger (resp. older) age. Even more interestingly, we show that in addition to impacting the strength of dependence, different values of the conditioning variable lead to different copula models. Although, because of high levels of censoring and a rather small sample of observations, we may not be able to define the copula families best suited to different levels of the covariate with high degrees of confidence, our model still allows us to use specific copulas for these different levels, thanks to the fact that we directly estimate the generator function rather than the copula itself.

Next, we provide an application to micro-level reserving. We use a Canadian automobile insurance dataset in which each policy in force provides multiple coverages to the insureds. We model the dependence between two of these coverages by means of their activation delays while conditioning on the age of the policyholders. We find again that different levels of the covariate provide different estimates of Kendall's tau. We also observe an impact of the covariate on the structure of dependence. Although Joe copulas are selected as best fitted to the data in both the unconditional and conditional models, this choice is not evident for some levels of the conditioning variable. We compare four different prediction approaches for the activation delays of the insurance coverages, using Joe copulas or the estimated generator functions, with and without conditioning on the age group of the policyholders. We show that using direct simulations from the generator function provides predictions that are closer to the observed activation delays. In addition, conditioning on the covariate also seems to result in better predictions, whether using Joe copulas or the generators.

These applications demonstrate the benefits that can be reaped from considering the impact of endogenous variables in dependence models. By doing so, we seem to be able to better capture the specific dependence structure present in a dataset. In a claims reserving setting as the one described in Section 3.5 where a lot of information is available to insurers, we show that making use this information can lead to predictions that more closely fit the data.

In this paper, we have selected the levels of the conditioning covariates for which to analyse the dependence

structure in a trivial way. For both the diabetic retinopathy study and the loss reserving application, we chose splitting points that provided subsets of similar size. It would be interesting, in future research, to explore a more robust method to select the levels of the conditioning variables at which the dependence model experiences a significant change.

CONCLUSION

Dans cette thèse, nous poursuivons deux objectifs. Premièrement, nous visons à contribuer à la littérature relative au calcul de micro-réserves en assurance non-vie. Nous proposons des approches permettant non seulement d'estimer le montant de la réserve au niveau individuel de chaque réclamation, mais également de modéliser différents événements prenant place durant le temps de vie de cette réclamation au sein du portefeuille de l'assureur. Nous cherchons ainsi à prédire le développement de chaque réclamation au plus près, de sa déclaration à sa fermeture, en passant par l'estimation des couvertures d'assurance que cette dernière impacte, les occurrences de paiements et les montants de ceux-ci.

Notre deuxième objectif est, au vu de la diversification et complexification des portefeuilles des assureurs, ainsi que des réglementations croissantes sur les marchés mondiaux de l'assurance, d'inclure la dépendance entre les risques dans nos modèles de provisionnement. Nous nous concentrons sur un portefeuille d'assurance automobile et proposons diverses approches afin d'incorporer la dépendance entre les différentes couvertures proposées au sein d'une même police dans l'estimation des réserves. Pour ce faire, les approches que nous employons visent à tirer parti de la grande quantité de données récoltées par les compagnies d'assurance, afin de représenter au mieux la structure de dépendance en leur sein.

Dans le premier chapitre, nous présentons un modèle de micro-réserves permettant une modélisation du développement complet de chaque réclamation. Ce modèle est basé sur le concept de *schémas d'activation* des couvertures d'assurance par lequel nous y incluons la dépendance entre ces dernières. Pour chaque réclamation, nous définissons le schéma d'activation comme étant un vecteur de longueur égale au nombre de couvertures offertes pour chaque police dans le portefeuille de l'assureur. Chaque entrée du vecteur est un élément binaire prenant la valeur 1 si la réclamation impacte, ou a *activé*, la couverture en question, et 0 sinon. Nous utilisons un modèle logistique multinomial qui nous permet d'estimer conjointement les entrées du vecteur et donc de prendre en compte la dépendance entre les différentes couvertures. Conditionnellement à l'activation des couvertures, nous modélisons par la suite l'occurrence de paiements pour chacune d'elles, et les sévérités en cas de paiement. Nous montrons comment utiliser cette stratégie pour plusieurs périodes de développement successives afin d'obtenir une estimation de la réserve totale pour notre portefeuille. Une comparaison entre ce modèle et des modèles de provisionnement plus classiques, ne prenant pas compte de la dépendance, permet d'illustrer la performance de notre approche. Par celle-ci, nous obtenons en effet des prédictions pour le montant des réserves plus proches des montants observés

que par des méthodes plus traditionnelles.

Dans le deuxième chapitre, nous passons en temps continu et troquons les schémas d'activation pour des *délais d'activation*. Ceux-ci sont définis, pour chaque couverture, comme le temps écoulé entre le moment de déclaration de la réclamation et le premier moment auquel l'assureur enregistre cette réclamation sous la couverture d'assurance en question. Si, pour une réclamation donnée, une couverture n'a pas été activée, le délai d'activation pour cette dernière est censuré et prend la valeur du délai de fermeture du sinistre. Nous utilisons la famille des copules archimédiennes pour modéliser conjointement les délais des différentes couvertures du portefeuille. Plus précisément, nous nous concentrons sur la fonction génératrice caractéristique des copules de cette famille et étendons un estimateur non-paramétrique proposé par (Genest et Rivest, 1993) au cas de données censurées. Pour ce faire, nous utilisons des techniques empruntées à l'analyse de survie. Une fois un estimateur non-paramétrique obtenu pour la fonction génératrice, nous proposons deux approches différentes pour réaliser des prédictions. Premièrement, nous illustrons une méthode graphique permettant de choisir la copule archimédienne la plus appropriée aux données, simplement en sélectionnant celle dont la fonction génératrice ressemble le plus à celle estimée. Nous discutons de plusieurs outils pour valider ce choix graphique. Deuxièmement, nous présentons un algorithme permettant de simuler des données directement à partir de l'estimateur non-paramétrique de la fonction génératrice. Nous montrons que ces données simulées présentent la même force et forme de dépendance que les données originales. Cette nouvelle méthode permet de ne pas imposer une copule pré-définie aux données, et est particulièrement utile lorsque la méthode graphique mentionnée précédemment ne mène pas à un choix évident de modèle. Dans une application au calcul de micro-réserves, nous montrons que l'approche par simulations à partir du générateur permet d'obtenir des prédictions pour les délais d'activation des couvertures plus proches des délais observés que l'approche utilisant une copule existante.

Le dernier chapitre de cette thèse s'inscrit dans la continuité des deux précédents. Nous y poursuivons la modélisation de la dépendance entre couvertures d'assurance au travers de leurs délais d'activation, à l'aide de la famille de copules archimédiennes. Dans cette dernière partie, nous poussons plus loin l'analyse de cette dépendance en y incorporant les effets de variables explicatives. Ce faisant, nous cherchons à nous soustraire à l'hypothèse simplificatrice qui accompagne très souvent la modélisation par copules. Cette hypothèse implique qu'une copule est invariante aux effets de variables explicatives. Pour nous en défaire, nous nous concentrons une fois de plus sur la fonction génératrice des copules archimédiennes pour laquelle nous proposons un nouvel estimateur paramétrique. En incorporant des variables explicatives dans

cet estimateur, nous sommes en mesure d'évaluer leur impact non seulement sur la force de dépendance via le tau de Kendall, mais également sur la forme de la copule. Nous démontrons l'utilité de notre modèle à travers deux applications : l'une sur une étude de la rétinopathie diabétique et une sur les données de notre portefeuille d'assurance pour lequel nous désirons estimer les réserves granulaires. Dans les deux cas, le fait de conditionner le modèle de dépendance sur une variable explicative fournit des résultats différents des modèles qui n'incluent pas ce conditionnement. Dans le cas du portefeuille d'assurance automobile, nous observons par exemple qu'en fonction de l'âge des détenteurs de police, la force et la forme de dépendance varient. Les prédictions obtenues par la suite avec les modèles qui prennent en compte ces variations sont plus proches des valeurs observées dans les données initiales que celles obtenues avec des modèles qui n'en tiennent pas compte.

Au travers de ces trois chapitres, nous présentons des méthodes novatrices pour inclure la dépendance au sein de modèles de calcul de micro-réserves en assurance non-vie. L'intérêt de ces méthodes dans un contexte d'assurance relève notamment du fait qu'elles tirent profit de la grande quantité de données disponibles aux compagnies d'assurance. Elles leur permettent également de représenter plus fidèlement les structures complexes et diversifiées de leurs portefeuilles de polices.

Cette thèse ouvre la porte à de possibles futurs travaux de recherche. Premièrement, les méthodes utilisées pour modéliser la dépendance entre couvertures d'assurance pourraient être étendues à d'autres types de dépendance. En particulier, nous pourrions ajouter à la dépendance entre les délais d'activation celle avec les délais de déclaration, avec la fréquence voire la sévérité des sinistres. Il pourrait également être intéressant d'investiguer d'autres familles de copules que la famille archimédienne. Comme dans le Chapitre 2, nous pourrions trouver une stratégie plus générale et s'appliquant à un plus grand nombre de familles de copules, pour estimer la force et la structure de dépendance d'un jeu de données sans devoir imposer une copule pré-définie. Enfin, dans le cadre du Chapitre 3, il serait intéressant d'élaborer une stratégie permettant de déterminer de façon non-triviale les niveaux d'une variable explicative pour lesquels le modèle de dépendance varie.

In this thesis, we pursue two main goals. First, we contribute to the actuarial literature on micro-level claims reserving for non-life insurance. We investigate approaches that not only estimate the reserve amount at the individual level of each claim, but also model various events occurring during the lifetime of each claim within the insurer's portfolio. As such, we seek to predict the development of each claim from reporting to settlement, by estimating the insurance coverages it impacts, the occurrences of payments, and their amounts.

Our second objective is, given the diversification and complexity of insurers' portfolios, as well as the increasing regulations in global insurance markets, to incorporate dependence between risks in our reserving models. We focus on an automobile insurance portfolio and propose various approaches to incorporate dependence between the different coverages offered within a single policy in our reserving models. To achieve this, our methods aim to leverage the large amount of data collected by insurance companies to best represent the dependency structure within them.

In the first chapter, we present a micro-level reserving model that allows to estimate the complete development of each claim. This model is based on the concept of *activation patterns* for the insurance coverages, through which we incorporate dependence. For each claim, we define the activation pattern as a vector of length equal to the number of coverages offered for each policy in the insurer's portfolio. Each entry of the vector is a binary element taking the value 1 if the claim impacts, or *activates*, the corresponding coverage, and 0 otherwise. We use a multinomial logit model to jointly estimate the vector entries, thus accounting for dependence between the different coverages. Conditioning on coverage activation, we subsequently model the occurrence of payments for each coverage and their severities in case of payment. We demonstrate how to use this strategy for successive development periods to obtain an estimation of the total reserve for our portfolio. A comparison between this model and more classical reserving models, which do not account for dependence, illustrates the performance of our approach. Using our model, we obtain predictions for the reserve amounts closer to the observed amounts than those obtained with more traditional methods.

In the second chapter, we move to continuous time and exchange activation patterns for *activation delays*. These are defined, for each coverage, as the time elapsed between the moment the claim is reported and

the first moment at which the insurer records this claim under the respective insurance coverage. If, for a given claim, a coverage has not been activated, its activation delay is censored and takes the value of the claim's settlement delay. We use the family of Archimedean copulas to jointly model the delays of the different coverages in the portfolio. Specifically, we focus on the generator function that characterizes copulas from this family and extend a non-parametric estimator proposed by (Genest et Rivest, 1993) to the case of censored data. To do this, we use techniques borrowed from survival analysis. Once a non-parametric estimator is obtained for the generator function, we propose two different approaches for making predictions. First, we illustrate a graphical method allowing the selection of the most appropriate Archimedean copula for the data, simply by selecting the one whose generator function most closely resembles the estimated generator. We discuss several methods to validate this graphical choice. Second, we present an algorithm to simulate new data directly from the non-parametric estimator of the generator function. We show that these simulated data exhibit the same strength and form of dependence as the original data. This new method allows us to avoid imposing a predefined copula on the data and is particularly useful when the graphical method mentioned above does not lead to an obvious model choice. In an application to loss reserving, we show that the simulation approach from the generator allows for predictions of coverage activation delays closer to observed delays than the approach using an existing copula.

The last chapter of this thesis builds on the previous two. We continue the modeling of dependence between insurance coverages through their activation delays, using the family of Archimedean copulas. In this final part, we further analyze this dependence by incorporating the effects of explanatory variables. In doing so, we aim to move away from the simplifying assumption that often accompanies copula modeling. This assumption implies that a copula is invariant to the effects of covariates. To overcome this, we once again focus on the generator function of Archimedean copulas, for which we propose a new parametric estimator. By incorporating endogenous variables into this estimator, we are able to evaluate their impact not only on the strength of dependence via Kendall's tau but also on the shape of the copula. We demonstrate the usefulness of our model through two applications : one on a diabetic retinopathy study and one on our insurance portfolio data for which we want to estimate the reserves. In both cases, conditioning the dependence model on a covariate provides results different from models that do not include this conditioning. In the case of the automobile insurance portfolio, for example, we observe that depending on the age of policyholders, the strength and form of dependence vary. Predictions subsequently obtained with models that take these variations into account are closer to the values observed in the initial data than those obtained with models that do not.

Through these three chapters, we present innovative methods for including dependence in non-life micro-level claims reserving. The interest of these methods in an insurance context lies in the fact that they leverage the large amount of data available to insurance companies. They also allow them to more faithfully represent the complex and diversified structures of their portfolios of claims.

This thesis opens the door to possible future research. First, the methods used to model dependence between insurance coverages could be extended to other types of dependence. In particular, we could add to the dependence between activation delays that between reporting delays, with frequency or severity of claims. It could also be interesting to investigate other families of copulas than the Archimedean family. As in Chapter 2, we could find a more general strategy applicable to a larger number of copula families to estimate the strength and structure of dependence in a dataset without having to impose a predefined copula. Finally, in the context of Chapter 3, it would be interesting to develop a strategy to determine non-trivially the levels of an explanatory variable for which the dependence model varies.

APPENDIX A

CHAPTER 1

A.1 Observed activation patterns

Table A.1 presents the activation patterns $\boldsymbol{A}_{i,1}$ observed in the dataset for the first development period of the claims. The value 1 stands for the activation of the coverage. 85.67% of all claims activate the Vehicle Damage coverage alone or simultaneously with the Loss of Use coverage. The Accident Benefits coverage is most often triggered with the Vehicle Damage coverage or on its own. We also observe that around 1% of all claims simultaneously activate all four coverages upon their reporting.

Table A.1 – Frequency of the activation patterns observed in the first development period and some descriptive statistics (all delays are in periods of 6 months).

AB	BI	VD	LU	$\%$ of claims	Average payment	Average reporting delay	Average settlement delay
0	0	1	0	42.95	2924.37	1.08	1.31
0	0	1	1	42.72	3181.20	1.03	1.45
1	0	1	1	4.98	6983.12	1.01	2.94
0	1	1	1	1.99	10117.42	1.02	3.15
1	0	0	0	1.44	15641.02	1.13	2.97
0	1	1	0	1.24	16496.29	1.07	3.28
1	0	1	0	1.07	10643.04	1.02	2.65
0	0	0	1	1.04	918.83	1.05	1.38
1	1	1	1	1.01	14802.81	1.01	3.82
0	1	0	0	0.61	27863.38	1.30	3.59
1	1	1	0	0.43	27505.77	1.04	4.04
1	1	0	0	0.38	44262.49	1.14	4.41
1	0	0	1	0.10	6833.90	1.01	3.18
0	1	0	1	0.04	9856.29	1.01	3.20
1	1	0	1	0.01	15799.64	1.02	4.86

A.2 Risk factors

Figure A.1 completes Section 1.2.2 by providing further insights on the different risk factors used in the activation patterns model.

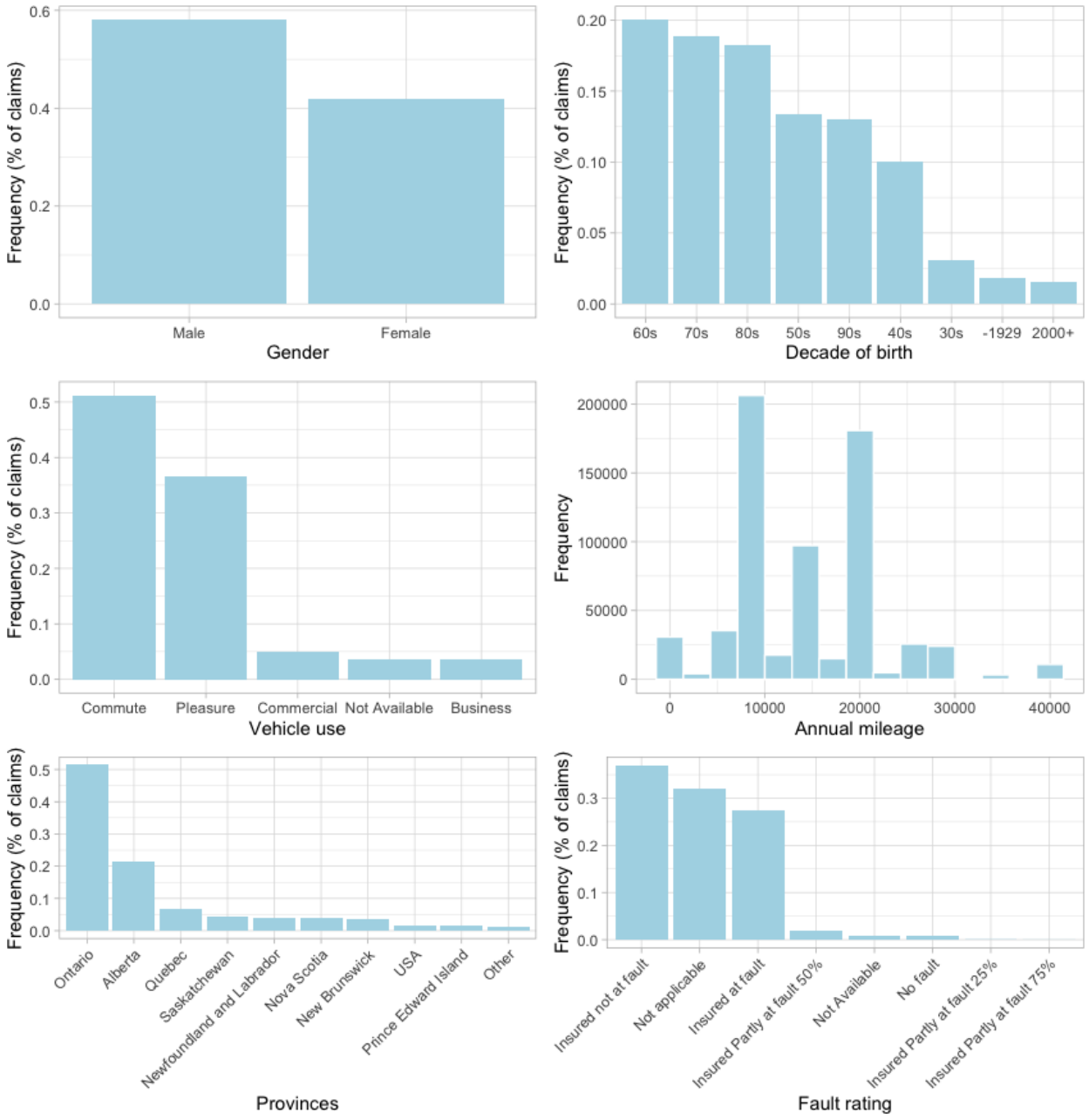

Figure A.1 – Risk factors

Our dataset includes around 60% of male insureds against 40% of female insureds, more than half born between 1960 and 1989. Approximately half of the insureds use their vehicle for commuting, while an addi-

tional 37% use it for pleasure. The remaining 13% of the insureds use their car for commercial or business reasons or did not disclose that information to the insurer. 76% of the insureds drive between 10,000 and 20,000 kilometers per year according to the annual mileage risk factor, which is the only continuous one in our dataset. More than 50% of the claims occur in Ontario, 20% in Alberta, and the remaining 30% in other Canadian provinces or in the United States of America. Finally, the insurance company assesses the insured's level of responsibility in an accident and issues a fault rating. In 37% of the claims, the insurance company considered the insured not to be at fault. The company could not apply the rating for 32% of the claims, and in 27% of the cases, the insured bore the entire fault. For the remaining 4% of the claims, there was either no fault at all or the fault was divided between the insured and any other parties involved in the claim.

A.3 Training and valuation sets

To fit the activation patterns model, we split the data into a training set for estimation, and a valuation set for simulation, as illustrated in Figure A.2. In this Figure, each black continuous (resp. dashed) line symbolizes the observed (resp. future) development of a claim.

The red vertical line in Figure A.2 represents the chosen valuation date of the reserves. All the information displayed on the left of this valuation date (continuous lines) is the observed claims data at the valuation date. Since we only focus on RBNS claims in this paper, it contains all the claims that were reported before the valuation date. For those already settled, it includes the complete claim development. For those still open at the valuation date, it contains all the information (i.e., activation of coverages, payments made, and corresponding amounts) observed up to that date. All the information on the right of the red line, represented by the dashed lines, makes up the valuation dataset we will use to predict the reserves. Our dataset also contains around 1% of open claims, i.e., claims not yet settled on the 31^{st} of June 2021. We include these claims by taking their last observed payment as the settlement date.

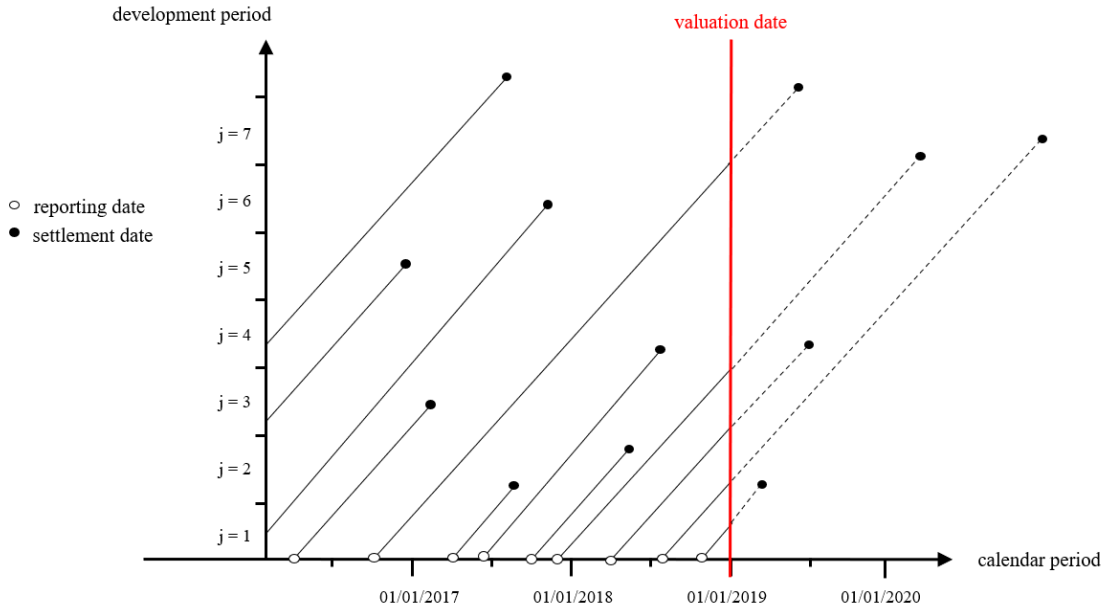

Figure A.2 – Illustration of the separation of the dataset into the training and valuation sets.

A.4 Multinomial logit model

Table A.2 – Parameter estimates for the multinomial logit model

	Risk factors	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Intercept		3.45	3.44	-0.26	-12.78	1.31	1.93	0.30	-1.04	1.14	1.67	0.28	-1.68	0.56	1.69
PROV	New Brunswick	0.21	-0.16	1.11	0.43	0.05	-0.49	-0.93	0.29	-0.18	0.35	0.48	1.07	0.40	0.35
	Newfoundland and Labrador	0.55	0.02	0.58	0.63	0.09	0.27	-0.06	-0.16	-0.09	0.26	0.79	0.83	0.61	0.65
	Nova Scotia	0.36	-0.05	1.09	0.18	-0.08	-0.72	-1.08	0.23	-0.73	0.30	0.48	0.01	-0.01	0.29
	Ontario	-0.14	-0.12	0.93	0.25	-0.35	-0.32	-0.50	-0.15	-1.26	-0.34	0.15	0.41	-0.65	-0.31
	Other	0.57	-0.14	0.61	-0.62	-0.06	-0.08	0.25	0.10	0.34	-0.28	0.76	1.87	0.71	0.29
	Prince Edward Island	0.93	-0.13	0.93	0.98	0.17	-0.75	-0.57	0.31	0.01	0.31	0.58	-3.21	1.11	0.61
	Quebec	1.64	-0.11	-2.10	-19.74	-2.77	-3.31	-2.34	-2.10	-1.97	-2.01	-1.66	-0.22	-2.46	-2.34
	Saskatchewan	1.53	-0.40	-1.21	-23.02	-1.04	-2.84	-1.05	-0.78	-0.41	-1.27	-2.73	-390.59	-0.60	-2.37
	USA	-5.52	-1.50	-3.83	11.02	-5.16	-2.24	-7.22	0.32	-5.19	-0.99	-3.53	0.10	-3.45	-0.86
YOB	2000+	0.71	0.70	3.71	14.79	3.87	3.80	3.90	2.92	3.61	3.48	3.94	3.49	3.84	3.62
	1930s	-0.34	-0.21	-1.13	9.95	-1.09	-0.84	-0.96	-1.10	-0.72	-1.0	-1.47	-1.95	-1.07	-1.04
	1940s	-0.41	-0.26	-0.72	10.46	-0.49	-0.56	-1.39	-1.82	-1.16	-1.08	-1.90	-2.24	-1.68	-1.64
	1950s	-0.55	-0.27	-0.82	10.36	-0.56	-0.73	-0.84	-1.57	-0.69	-0.82	-1.54	-1.81	-1.36	-1.29
	1960s	-0.55	-0.29	-0.58	10.44	-0.33	-0.44	-0.81	-1.31	-0.82	-0.88	-1.79	-1.36	-1.57	-1.60
	1970s	-0.60	-0.26	-0.54	10.37	-0.30	-0.37	-0.75	-1.32	-0.74	-0.81	-1.88	-1.43	-1.52	-1.51
	1980s	-0.73	-0.33	-0.48	10.18	-0.29	-0.40	-0.69	-1.14	-0.78	-0.71	-1.62	-1.18	-1.27	-1.26
	1990s	-0.44	-0.30	-0.30	11.04	-0.08	-0.12	0.15	-1.20	-0.07	-0.48	-0.50	-0.50	-0.39	-0.63
GENDER	Male	0.24	0.08	-0.22	-0.44	-0.23	-0.28	-0.26	-0.89	-0.35	-0.89	-0.28	-0.69	-0.45	-0.92
AM		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VU	Commercial	0.68	-0.11	0.83	0.25	0.71	-0.19	0.28	-0.299	0.89	-0.19	0.71	-0.55	0.84	-0.27
	Commute	-0.13	-0.00	-0.08	-0.20	-0.05	-0.14	0.03	-0.11	0.15	0.09	0.07	-0.34	0.17	-0.02
	Not Available	2.37	0.13	2.73	-0.07	2.59	0.28	2.25	-1.88	1.44	-1.13	0.11	-3.26	1.28	-1.50
	Pleasure	0.15	-0.03	-0.00	-0.15	0.05	-0.21	0.25	-0.22	0.57	0.04	0.34	-0.46	0.57	-0.06
FR	Insured not at fault	-0.94	-0.06	-2.53	-2.32	-3.38	-2.70	0.55	0.84	-0.23	0.72	-1.28	-1.39	-1.82	-1.35
	Insured Partly at fault 25%	1.10	1.37	2.96	-12.13	1.62	2.13	0.89	-9.41	2.23	2.10	3.51	-2.41	3.43	2.98
	Insured Partly at fault 50%	0.34	0.56	-0.01	-0.42	-0.12	-0.01	-0.92	-0.38	0.18	0.71	-0.07	0.16	0.17	0.50
	Insured Partly at fault 75%	-0.55	-0.27	0.61	-8.55	0.88	0.34	0.93	-6.36	0.61	0.67	2.64	-3.71	0.80	1.49
	No fault	4.30	3.23	-1.19	-5.06	-1.25	0.05	6.37	1.61	2.30	4.04	1.29	-2.13	-0.83	0.87
	Not applicable	1.46	-0.07	-1.88	-28.36	-3.83	-4.53	1.05	-1.28	-1.12	-1.36	-0.77	-2.98	-2.35	-3.16
	Not Available	5.81	1.34	3.74	-12.24	4.27	1.26	7.88	0.74	4.41	0.99	4.16	-0.43	3.32	1.06

A.5 Choice of the severity models

Table A.3 – Choice of distributions for the severity

Coverage	Model	$\forall j$	
		AIC	BIC
Accident Benefits	Log-Normal	1,101,322	1,101,790
	Gamma	1,099,669	1,100,137
	Pareto	1,093,130	1,093,598
	Generalized Beta II	**1,092,513**	**1,092,998**
	Weibull	1,096,566	1,097,034
Bodily Injury	Log-Normal	519,469	519,896
	Gamma	519,451	519,878
	Pareto	520,152	520,580
	Generalized Beta II	518,800	519,236
	Weibull	**517,012**	**517,439**
Vehicle Damage	Log-Normal	6,818,688	6,819,250
	Gamma	6,811,968	6,812,530
	Pareto	6,767,629	6,768,192
	Generalized Beta II	**6,760,099**	**6,760,684**
	Weibull	6,797,965	6,798,528
Loss of Use	Log-Normal	2,318,483	2,319,003
	Gamma	2,315,044	2,315,564
	Pareto	2,345,577	2,346,097
	Generalized Beta II	**2,305,159**	**2,305,699**
	Weibull	2,338,368	2,338,888

A.6 Results stability

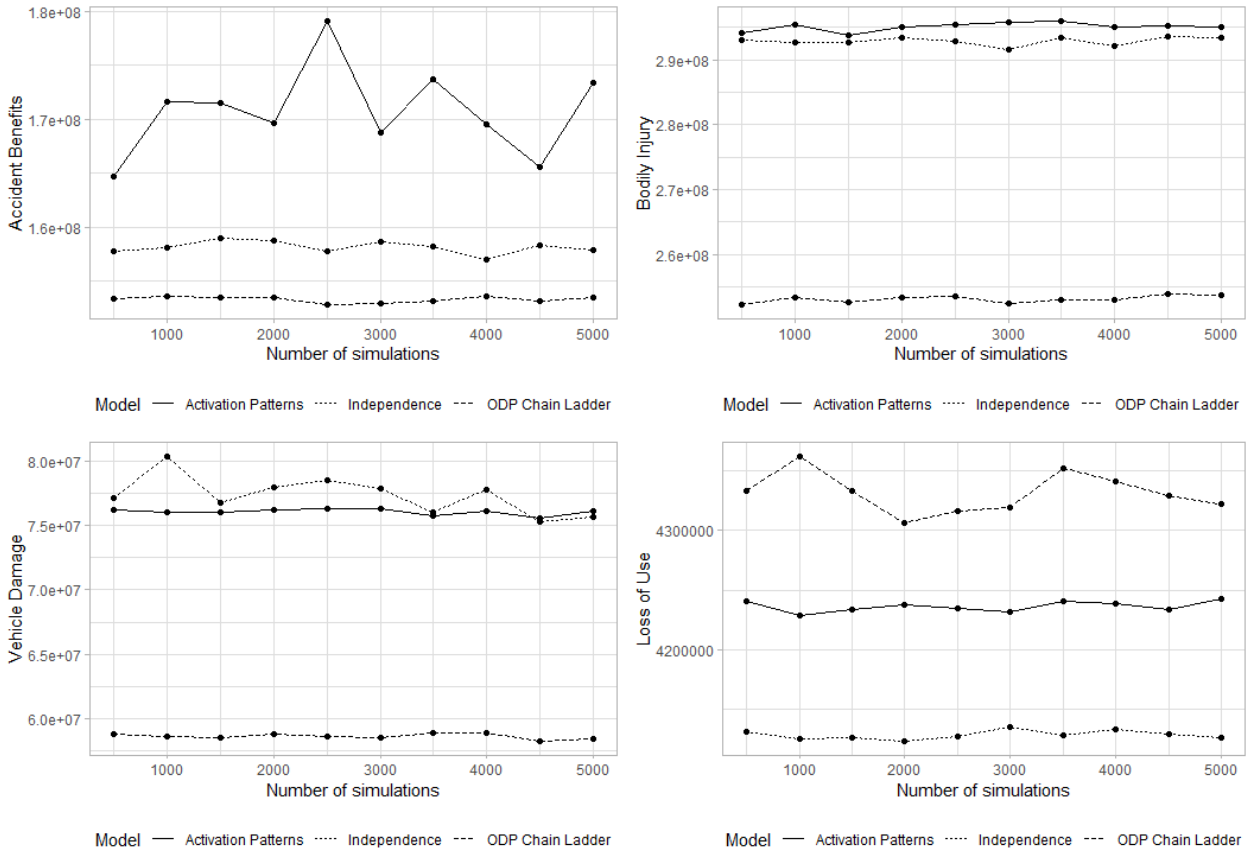

Figure A.3 – Results of the RBNS reserves (95% VaR) based on the number of simulations performed

APPENDIX B

CHAPTER 2

B.1 Useful functions and relationships for some popular Archimedean copulas

Table B.1 – Expressions of the copula, generator function and relation between α, Kendall's tau and $\lambda(.)$ for some commonly used Archimedean copulas.

Copula	$C_\alpha(u_1, u_2)$	$\psi(\nu)$	τ	$\lambda_\alpha(\nu)$
Clayton	$(u_1^{-\alpha} + u_2^{-\alpha} - 1)^{-1/\alpha}$	$(1+t)^{-1/\alpha}$	$\alpha/(\alpha+2)$	$\nu(\nu^\alpha - 1)/\alpha$
Frank	$-\frac{1}{\alpha}\ln\left(1 + \frac{(e^{-\alpha u_1}-1)(e^{-\alpha u_2}-1)}{e^{-\alpha}-1}\right)$	$-(\ln(e^{-t}(e^{-\alpha}-1)+1))/\alpha$	$1 + \frac{4}{\alpha}\left(\int_0^\alpha \frac{\xi}{\alpha(e^\xi - 1)} d\xi - 1\right)$	$-\ln\left\{(e^{-\alpha\nu}-1)/(e^{-\alpha}-1)\right\}/(-\alpha/(e^{\alpha\nu}-1))$
Gumbel-Hougaard	$\exp\left(-[(\tilde{u}_1)^\alpha + (\tilde{u}_2)^\alpha]^{1/\alpha}\right)$	$\exp(-t^{1/\alpha})$	$1 - 1/\alpha$	$\nu\ln(\nu)/\alpha$
Joe	$1 - (\bar{u}_1^\alpha + \bar{u}_2^\alpha - \bar{u}_1^\alpha \bar{u}_2{}^\alpha)^{1/\alpha}$	$1 - (1-e^{-t})^{1/\alpha}$	$-$	$-\ln\left\{1-(1-\nu)^\alpha\right\} / -\left((\alpha(1-\nu)^{(\alpha-1)})/(1-(1-\nu)^\alpha)\right)$

B.2 Distributions of U and V proposed in (Wang, 2010) for diverse censoring patterns

The following results are based on Corollary 1., Corollary 2. and Theorem 4. from (Wang, 2010).

Let (T_1, T_2) be a bivariate vector submitted to censoring by the bivariate continuous vector (X_1, X_2), as described in Section 2.3.1. Let $S_1(t_1)$ and $S_2(t_2)$ be the marginal survival functions of, respectively, T_1 and T_2, whose dependence function can be modelled by an absolutely continuous Archimedean copula such that

$$C(S_1(t_1), S_2(t_2)) = S(t_1, t_2) = \phi^{-1}\Big\{\phi(S_1(t_1)) + \phi(S_2(t_2))\Big\}.$$

Then, we have that :

1. if both observations are censored, the distribution function of $(V|T_1 > x_1, T_2 > x_2)$ is

$$F_1(v, x_1, x_2) = \frac{1}{S(x_1, x_2)}\left[v - \frac{\phi(v) - \phi\{S(x_1, x_2)\}}{\phi'(v)}\right], \quad 0 \leq v \leq S(x_1, x_2)$$

   and the distribution function of $(U|T_1 > x_1, T_2 > x_2)$ is uniformly distributed on the interval

$$\left[\frac{\phi\{S_1(x_1)\}}{\phi(v)}, 1 - \frac{\phi\{S_2(x_2)\}}{\phi(v)}\right];$$

2. if only the second observation is censored, the distribution function of $(V|T_1 = t_1, T_2 > x_2)$ is

$$F_2(v, t_1, c_2) = \frac{p'(\phi(v))}{p'(\phi(S(t_1, x_2)))}, \quad 0 \leq v \leq S(t_1, x_2)$$

   with $p(.) = \phi^{-1}(.)$ and the distribution function of $(U|T_1 = t_1, T_2 > x_2)$ is

$$G_2(y, t_1, x_2) = \frac{p'\{\phi(S_1(t_1))/u\}}{p'\{\phi(S(t_1, x_2))\}}, \quad 0 \leq u \leq \frac{\phi\{S_1(t_1)\}}{\phi\{S(t_1, x_2)\}};$$

3. if only the first observation if censored, the distribution function of $(V|T_1 > x_1, T_2 = x_2)$ is

$$F_3(v, x_1, t_2) = \frac{p'(\phi(v))}{p'(\phi(S(x_1, t_2)))}, \quad 0 \leq v \leq S(x_1, t_2)$$

   and the distribution function of $(U|T_1 > x_1, T_2 = t_2)$ is

$$G_3(y, x_1, t_2) = \frac{p'\{\phi(S_2(t_2))/(1-u)\}}{p'\{\phi(S(x_1, t_2))\}}, \quad \frac{\phi\{S_1(x_1)\}}{\phi\{S(x_1, t_2)\}} \leq u \leq 1;$$

4. if both observations are uncensored, the distribution function of $(V|T_1 = t_1, T_2 = t_2)$ is the same as in Proposition 1.1. from (Genest et Rivest, 1993) and the conditional distribution of $(U|V = v)$ is just a uniform distribution on $[0, 1]$ independent of V.

B.3 Model validation approaches for the parametric copula selected via the graphical procedure

B.3.1 Omnibus estimation procedure

The omnibus estimation procedure is a semi-parametric optimization procedure that substitutes empirical versions of the marginal distributions in the (parametric) likelihood function of the copula model, $L(.)$.

We use rescaled Kaplan-Meier estimators for T_1 and T_2, multiplying the original Kaplan-Meier versions by $n/(n+1)$, where n is the total number of observations in the sample. As mentioned in (Genest *et al.*, 1995) as well as (Denuit *et al.*, 2006), this rescaling prevents issues related to the potential unboundedness of the copula log-density as u_1 and u_2 tend to one.

Once we have the marginal distribution functions, we use them in the estimation step, i.e., the optimization of the model likelihood to obtain the pseudo-likelihood estimator $\hat{\alpha}^*$ of the dependence parameter :

$$\hat{\alpha}^* = \arg\max L(u_1, u_2,\delta_1, \delta_2; \alpha),$$

where the likelihood function under flexible censoring scenarios is given by :

$$\begin{aligned} L(u_1, u_2;\delta_1, \delta_2; \alpha) = \prod_{i=1}^{n} c_\alpha(u_{1i}, u_{2i}; \alpha)^{\delta_{1i}\delta_{2i}} + \left(\frac{\partial C_\alpha(u_{1i}, u_{2i}; \alpha)}{\partial u_1}\right)^{\delta_{1i}(1-\delta_{2i})} \\ + \left(\frac{\partial C_\alpha(u_{1i}, u_{2i}; \alpha)}{\partial u_2}\right)^{(1-\delta_{1i})\delta_{2i}} + C_\alpha(u_{1i}, u_{2i}; \alpha)^{(1-\delta_{1i})(1-\delta_{2i})}. \end{aligned} \tag{B.1}$$

C_α is the candidate Archimedean copula under consideration, c_α its density and $\delta_{ji} = 1$ if $T_j < X_j$ for $j = 1, 2$ for observation i and equals zero otherwise (i.e., for a censored observation). The partial derivatives with respect to either u_1 or u_2 can be found in Table B.2.

In order to use the omnibus procedure to validate the model choice resulting from the graphical comparison, we retrieve the pseudo-maximum likelihood estimator $\hat{\alpha}^*$ for the chosen candidate copula models and compare them to the estimators $\hat{\alpha}$ found using Equation (2.7). The copula with the smallest difference between these two estimators is the most appropriate for the data.

B.3.2 L^2-norm

The second approach to validate the choice of copula model is based on the L^2-norm. We follow the methodology laid out in (Wang et Wells, 2000) that introduce a goodness-of-fit statistic specifically designed

Table B.2 – Partial derivatives for some popular Archimedean copulas with $\tilde{u} = -\ln u$ and $\bar{u} = 1 - u$.

Copula	$\partial C_\alpha(u_1.u_2)/\partial u_1$
Clayton	$[1 + u_1^\alpha(u_2^{-\alpha} - 1)]^{-1-1/\alpha}$
Frank	$[e^{-\alpha u_1} - e^{-\alpha(u_1+u_2)}] \times [(1 - e^{-\alpha}) - (1 - e^{-\alpha u_2})(1 - e^{-\alpha u_1})]^{-1}$
Gumbel	$u_1^{-1} e^{\{-(\tilde{u}_2^\alpha + \tilde{u}_1^\alpha)^{1/\alpha}\}} [1 + (\tilde{u}_2/\tilde{u}_1)^\alpha]^{-1+1/\alpha}$
Joe	$(1 - \bar{u}_2^\alpha)(1 - \bar{u}_2^\alpha + \bar{u}_2^\alpha \bar{u}_1^{-\alpha})^{-1+1/\alpha}$

for Archimedean copulas with censored data under the assumption that a consistent non-parametric estimator of the bivariate joint distribution function is available. Note that with estimated parameters present in the specification of the null hypothesis for this test, the approach we show here is not an actual statistical test but rather a way to validate our model. However, we will use the word *test* in this section for simplicity. As mentioned in the introduction, more formal and general methods have been proposed in the literature. We discuss one in the next paragraph.

(Wang et Wells, 2000)'s test relies on the L^2-norm distance between the empirical estimator $\hat{K}_n(\nu)$ and the corresponding $K_{\hat{\alpha}}(\nu)$ of the hypothesized parametric copula model,

$$D(\hat{\alpha}) = \int_\xi^1 \left(\hat{K}_n(\nu) - K_{\hat{\alpha}}(\nu)\right)^2 d\nu.$$

We use Riemann sum approximations to simplify numerical analysis :

$$D(\hat{\alpha}) = \sum_{i=1}^{n} \left(\hat{K}_n(\nu_{(i)}) - K_{\hat{\alpha}}(\nu_{(i)})\right)^2 (\nu_{(i)} - \nu_{(i-1)}),$$

where $\nu_{(i)}$ is the i^{th} ordered value of $\{\nu_1, \nu_2, ..., \nu_n\}$. However, the asymptotic distribution of $D(\hat{\alpha})$ is difficult to retrieve analytically.

B.3.3 Goodness-of-fit test for censored dependent data

The omnibus procedure and L^2-norm validation approach are easy to implement and give a good idea of whether a copula model is appropriate for the data at hand. However, neither are formal statistical tests. To further increase the confidence in our choice of copula model, we present a more formal approach for Archimedean copulas, proposed by (Wang, 2010).

This goodness-of-fit test applies to uncensored and right-censored data thanks to a multiple imputation procedure. It bases its premise on Proposition 1.1. from (Genest et Rivest, 1993).

For a bivariate vector of dependent observations (T_1, T_2) whose marginal survival functions $S_1(.)$ and $S_2(.)$ have a dependence function $C(.)$ of the form

$$C(S_1(t_1), S_2(t_2)) = \psi\Big\{\psi^{-1}(S_1(t_1)) + \psi^{-1}(S_2(t_2))\Big\},$$

then, the random variables

$$U = \frac{\psi^{-1}(S_1(t_1))}{\psi^{-1}(S_1(t_1)) + \psi^{-1}(S_2(t_2))} \quad \text{and} \quad V = \psi\Big\{\psi^{-1}(S_1(t_1)) + \psi^{-1}(S_2(t_2))\Big\} = C(S_1(t_1), S_2(t_2)). \tag{B.2}$$

are independent. In other words, under the null hypothesis that an Archimedean copula can model the dependence between (T_1, T_2) with generator $\psi(.)$, the correlation coefficient between U and V is null :

$$H_0 : \rho = 0 \quad \text{vs} \quad H_1 : \rho \neq 0.$$

Let

$$r_n = \frac{\sum_{i=1}^{n}(\hat{U}_i - \bar{\hat{U}})(\hat{V}_i - \bar{\hat{V}})}{\sqrt{\sum_{i=1}^{n}(\hat{U}_i - \bar{\hat{U}})^2 \sum_{i=1}^{n}(\hat{V}_i - \bar{\hat{V}})^2}}$$

where $\hat{U}$ and $\hat{V}$ are consistent estimators of U and V and where $\bar{\hat{U}}$ and $\bar{\hat{V}}$ are the sample means of $\hat{U}_i$ and $\hat{V}_i$. The test statistic is defined as

$$Z_n = \frac{1}{2}\log\Big[\frac{1 + r_n}{1 - r_n}\Big]$$

and we have that $\sqrt{n}Z_n \to N(0, 1)$ in distribution.

For censored data, we can not directly retrieve $\hat{U}_i$ and $\hat{V}_i$ from the estimators proposed by (Genest et Rivest, 1993) based on Equation (B.2). Instead, (Wang, 2010) shows that, depending on the censoring pattern, the data can be simulated using a multiple imputation procedure from one of the distributions presented below. All proofs and further details are in (Wang, 2010).

Let (T_1, T_2) be a bivariate vector submitted to censoring by the bivariate continuous vector (X_1, X_2), as described in Section 2.3.1. Let $S_1(t_1)$ and $S_2(t_2)$ be the marginal survival functions of, respectively, T_1 and

T_2, whose dependence function can be modelled by an absolutely continuous Archimedean copula such that

$$C(S_1(t_1), S_2(t_2)) = S(t_1, t_2) = \psi\Big\{\psi^{-1}(S_1(t_1)) + \psi^{-1}(S_2(t_2))\Big\}.$$

Then, we have that :

1. if both observations are censored, the distribution function of $(V|T_1 > x_1, T_2 > x_2)$ is

$$F_1(v, x_1, x_2) = \frac{1}{S(x_1, x_2)}\left[v - \frac{\psi^{-1}(v) - \psi^{-1}\{S(x_1, x_2)\}}{\psi^{-1'}(v)}\right], \quad 0 \leq v \leq S(x_1, x_2)$$

and the distribution function of $(U|T_1 > x_1, T_2 > x_2)$ is uniformly distributed on the interval

$$\left[\frac{\psi^{-1}\{S_1(x_1)\}}{\psi^{-1}(v)}, 1 - \frac{\psi^{-1}\{S_2(x_2)\}}{\psi^{-1}(v)}\right];$$

2. if only the second observation is censored, the distribution function of $(V|T_1 = t_1, T_2 > x_2)$ is

$$F_2(v, t_1, c_2) = \frac{p'(\psi^{-1}(v))}{p'(\psi^{-1}(S(t_1, x_2)))}, \quad 0 \leq v \leq S(t_1, x_2)$$

with $p(.) = \psi(.)$ and the distribution function of $(U|T_1 = t_1, T_2 > x_2)$ is

$$G_2(y, t_1, x_2) = \frac{p'\{\psi^{-1}(S_1(t_1))/u\}}{p'\{\psi^{-1}(S(t_1, x_2))\}}, \quad 0 \leq u \leq \frac{\psi^{-1}\{S_1(t_1)\}}{\psi^{-1}\{S(t_1, x_2)\}};$$

3. if only the first observation if censored, the distribution function of $(V|T_1 > x_1, T_2 = x_2)$ is

$$F_3(v, x_1, t_2) = \frac{p'(\psi^{-1}(v))}{p'(\psi^{-1}(S(x_1, t_2)))}, \quad 0 \leq v \leq S(x_1, t_2)$$

and the distribution function of $(U|T_1 > x_1, T_2 = t_2)$ is

$$G_3(y, x_1, t_2) = \frac{p'\{\psi^{-1}(S_2(t_2))/(1-u)\}}{p'\{\psi^{-1}(S(x_1, t_2))\}}, \quad \frac{\psi^{-1}\{S_1(x_1)\}}{\psi^{-1}\{S(x_1, t_2)\}} \leq u \leq 1;$$

4. if both observations are uncensored, the distribution function of $(V|T_1 = t_1, T_2 = t_2)$ is the same as in Proposition 1.1. from (Genest et Rivest, 1993) and the conditional distribution of $(U|V = v)$ is just a uniform distribution on $[0, 1]$ independent of V.

B.4 Simulation study : Results of the Omnibus procedure

For each of the three scenarios, Table B.3 displays the results of the omnibus procedure. For each simulated sample, we compare the dependence parameters $\hat{\alpha}$ retrieved from Equation (2.7) to the maximum likelihood estimator $\hat{\alpha}^*$. We observe that in most cases, the distance between $\hat{\alpha}$ and $\hat{\alpha}^*$ is minimal for the correct copula, i.e., the omnibus procedure leads to selecting the copula model from which the data was originally simulated. We count three occurrences where the procedure selects an alternative model. In the single-censoring scenario, the Gumbel model is chosen for the Clayton sample, and the Joe copula is selected for the Gumbel sample. The Gumbel copula is selected for the Frank sample in the double-censoring scenario. However, we observe that the correct copulas present in each case are the second smallest distance between $\hat{\alpha}$ and $\hat{\alpha}^*$ and that the difference with the third and fourth distances is quite marked. For the Gumbel sample in the single censoring scenario for example, the smallest distance between the parameters is equal to 0.0428 for the Joe copula, 0.0992 for the Gumbel copula and then, 0.3988 and 0.4971 for, respectively, the Clayton and Frank copulas.

Table B.3 – Omnibus procedure for different censoring scenarios.

True copula	Candidate model	No censoring		single-censoring		double-censoring	
		$\hat{\alpha}$	$\hat{\alpha}^*$	$\hat{\alpha}$	$\hat{\alpha}^*$	$\hat{\alpha}$	$\hat{\alpha}^*$
Clayton ($\alpha = 1.3332$)	Clayton	**1.7694**	**1.8284**	1.1789	1.3263	**0.9240**	**0.8541**
	Frank	5.2063	5.7575	3.7724	4.1154	2.1323	2.7655
	Gumbel	1.8846	1.7294	**1.5895**	**1.4926**	1.4621	1.3070
	Joe	2.6347	1.7373	2.0732	1.4713	1.5231	1.2792
Frank ($\alpha = 4.1611$)	Clayton	1.3202	0.8049	1.2055	0.7428	1.0034	0.5326
	Frank	**4.1284**	**4.2527**	**3.8403**	**3.9827**	3.3149	2.9126
	Gumbel	1.6602	1.5274	1.6029	1.4533	**1.5018**	**1.1694**
	Joe	2.2066	1.6870	2.0983	1.6260	1.9084	1.1759
Gumbel ($\alpha = 1.6667$)	Clayton	1.1377	0.6557	1.1380	0.7392	0.9464	0.6493
	Frank	4.1641	4.3241	3.6674	4.1645	2.8121	3.6805
	Gumbel	**1.5690**	**1.5934**	1.5691	1.6683	**1.4734**	**1.5283**
	Joe	2.0345	1.8735	**2.0347**	**1.9919**	1.8671	1.7554
Joe ($\alpha = 2.2191$)	Clayton	1.2609	0.6067	1.1494	0.5514	0.9344	0.5170
	Frank	3.9802	4.2203	3.6968	3.8131	3.1298	3.4508
	Gumbel	1.6305	1.7284	1.5748	1.6356	1.4674	1.5638
	Joe	**2.151**	**2.1997**	**2.0454**	**2.0431**	**1.8441**	**1.9360**

APPENDIX C

CHAPTER 3

C.1 Simulating bivariate samples from a univariate distribution

In this appendix, we provide more details on the Marshall, Olkin ((Marshall et Olkin, 1988)) and `RLAPTRANS` ((Ridout, 2009)) algorithms that can be used to simulate a d-variate vector of dependent random variables from the generator function of Archimedean copulas. This strategy relies on Bernstein's theorem.

Theorem C.1 (Bernstein) *A function $\psi(.)$ is strictly monotonic and $\psi(0) = 1$ if and only if it can be written as*

$$\psi(\nu) = \mathcal{L}_{\Theta}(\nu), \tag{C.1}$$

where $\mathcal{L}(.)$ is the Laplace-Stieltjes transform of the strictly positive random variable Θ.

Knowing that Archimedean generators can be expressed as the Laplace-Stieltjes transform of a random variable Θ, we can generate samples from this random variable and use an algorihtm such as the Marshall, Olkin algorithm described below to obtain a d-variate vector from an Archimedean copula characterized by its generator.

Algorithm 4: Marshall, Olkin

1: Generate a random observation θ from the distribution with Laplace transform ψ.

2: For $i = 1, ..., d$, generate i.i.d. $X_i \sim \mathsf{U}(0, 1)$.

3: Return $(U_1, ..., U_d)$ where $U_i = \psi(-\log(X_i)/\theta)$, for $i = 1, ..., d$.

Algorithm 5, the so-called `RLAPTRANS` algorithm, provides a simple strategy based on a standard modification of the Newton-Rapjsom method for the first step of Algorihtm 4. Knowing the generator function $\psi(.)$, we can sample from its inverse Laplace-Stieltjes transform and generate observations from the random variable Θ.

Algorithm 5: RLAPTRANS

1: Generate n independent $U(0,1)$ observations and sort them : $u_{(1)} < ... < u_{(n)}$.

2: Find a value θ_{max} to serve as upper bound, i.e. $\psi^{-1}(\theta_{max}) \geq u_{(n)}$.

3: Set the lower bound $\theta_L = \theta_{(i-1)}$ and upper bound $\theta_U = \theta_{max}$. Repeat the modified Newton-Raphson procedure for $i = 1, ..., n$ to obtain the ordered sample $\theta_{(1)}, ..., \theta_{(n)}$.

4: Permute the ordered sample randomly to obtain the unordered sample $\theta_1, ..., \theta_n$.

BIBLIOGRAPHIE

Acar, E., Genest, C. et Nešlehová, J. (2012). Beyond simplified pair-copula constructions. *Journal of Multivariate Analysis*, *110*, 74–90.

Acar, E., Radu, C. et Yao, F. (2011). Dependence calibration on conditional copulas : A nonparametric approach. *Biometrics*, *67*(2), 445–453.

Akritas, M. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistics*, *22*, 1299–1327.

Akritas, M. et Van Keilegom, I. (2003). Estimation of bivariate and marginal distributions with censored data. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, *65*(2), 457–471.

Antonio, K. et Plat, R. (2014). Micro-level stochastic loss reserving for general insurance. *Scandinavian Actuarial Journal*, *2014*(7), 649–669.

Beran, R. (1981). *Nonparametric Regression with Randomly Censored Survival Data*. Tech. Rep, University of California, Berkeley.

Bornhuetter, R. et Ferguson, R. (1972). The actuary and ibnr. *Proc. CAS*, *59*, 181–195.

Bouezmarni, T., Lemyre, F. et El Ghouch, A. (2019). Estimation of a bivariate conditional copula when a variable is subject to random right censoring. *Electronic Journal of Statistics*, *13*, 5044–5087.

Breiman, L., Friedman, J., Olshen, R. et Stone, C. (1984). *Classification and Regression Trees*. New York : Wadsworth Statistics/Probability Series.

Buhlmann, H., Schnieper, R. et Straub, E. (1980). Claims reserves in casualty insurance based on a probabilistic model. *Bulletin of the Association of Swiss Actuaries*, *80*, 21–45.

Clayton, D. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, *65*, 141–151.

Crèvecoeur, J., Antonio, K., Desmedt, S. et Masquelein, A. (2023). Bridging the gap between pricing and reserving with an occurrence and development model for non-life insurance claims. *ASTIN Bulletin*, *53*(2), 185–212.

Crèvecoeur, J., Antonio, K. et Robben, J. (2022). A hierarchical reserving model for reported non-life insurance claims. *Insurance : Mathematics and Economics*, *104*, 158–184.

Côté, M.-P., Genest, C. et Stephens, D. (2022). A bayesian approach to modeling multivariate multilevel insurance claims in the presence of unsettled claims. *Bayesian Anal.*, *17*(1), 67–93.

Deheuvels, P. (1997). La fonction de dépendance empirique et ses propriétés : un test non paramétrique d'indépendance. *Bulletins de l'Académie Royale de Belgique*, *65*(6), 274–292.

Delong, L., Lindholm, M. et Wüthrich, M. (2021). Collective reserving using individual claims data. *Scandinavian Actuarial Journal*, *2022*(1), 1–28.

Delong, L. et Wüthrich, M. (2020). Neural networks for the joint development of individual payments and claim incurred. *Risks*, *7*, 102.

Denuit, M., Purcaru, O. et Van Keilegom, I. (2006). Bivariate archimedean copula models for censored data in non-life insurance. *Journal of Actuarial Practice*, *5*.

Deresa, N., Van Keilegom, I. et Antonio, K. (2022). Copula-based inference for bivariate survival data with left truncation and dependent censoring. *Insurance : Mathematics and Economics*, *107*, 1–21.

Derumigny, A. et Fermanian, J.-D. (2017). About tests of the 'simplifying' assumption for conditional copulas. *Dependence Modeling*, *5*, 154–197.

Derumigny, A. et Fermanian, J.-D. (2019). On kernel-based estimation of conditional kendall's tau : finite-distance bounds and asymptotic behavior. *Dependence Modeling*, *7*, 292–321.

Derumigny, A. et Fermanian, J.-D. (2022). Conditional empirical copula processes and generalized measures of association. *Electronic Journal of Statistics*, *16*, 5692–5719.

Duval, F. et Pigeon, M. (2019). Individual loss reserving using a gradient boosting-based approach. *Risks*, *7*, 79.

Emura, T., Lin, C.-W. et Wang, W. (2010). A goodness-of-fit test for archimedean copula models in the presence of right-censoring. *Computational statistics and data analysis*, *54*(12), 3033–3043.

Fermanian, J.-D. et Wegkamp, M. (2012). Time-dependent copulas. *Journal of Multivariate Analysis*, *110*, 19–29.

Frank, M. (1979). On the simultaneous associativity of f(x,y) and x+y-f(x,y). *Aequationes Mathematicae*, *19*, 194–226.

Frees, E., Shi, P. et Valdez, E. (2009). Actuarial applications of a hierarchical insurance claims model. *ASTIN Bulletin*, *39*(1), 165–197.

Frees, E. et Valdez, E. (2008). Hierarchical insurance claims modeling. *Journal of the American Statistical Association*, *103*(484), 1457–1469.

Frees, E. et Valdez, E. A. (1998). Understanding relationships using copulas. *North American Actuarial Journal*, *2*(1).

Frees, E. W., Jin, X. et Lin, X. (2013). Actuarial applications of multivariate two-part regression models. *Annals of Actuarial Science*, *7*(2), 258–287.

Frees, E. W. J., Meyers, G. et Cummings, A. (2010). Dependent multi-peril ratemaking models. *ASTIN Bulletin*, *39*(1), 699–726.

Geerdens, C., Acar, E. et Janssen, P. (2018). Conditional copula models for right-censored clusered event time data. *Biostatistics*, *19*(2), 247–262.

Genest, C., Ghoudi, K. et Rivest, L. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, *82*, 543–552.

Genest, C. et Rivest, L. (1993). Statistical inference procedures for bivariate archimedean copulas. *Journal of the American Statistical Association*, *88*, 1034–1043.

Genest, C., Rémillard, B. et Beaudoin, D. (2009). Goodness-of-fit tests for copulas : A review and a power study. *Insurance : Mathematics and Economics*, *4*(2), 199–213.

Gijbels, I., Omelka, M., Pešta, M. et Veraverbeke, N. (2017). Score tests for covariate effects in conditional copulas. *Journal of Multivariate Analysis*, *159*, 111–133.

Gribkova, S. et Lopez, O. (2015). Non-parametric copula estimation under bivariate censoring. *Scandinavian Journal of Statistics*, *42*(4).

Größer, J. et Okhrin, O. (2021). Copulae : An overview and recent developments. *Wiley Interdisciplinary Reviews : Computational Statistics*.

Gumbel, E. (1960). Distributions des valeurs extrêmes en plusieurs dimensions. *Publ. Inst. Statist. Univ. Paris*, *9*, 171–173.

Haastrup, S. et Arjas, E. (1996). Claims reserving in continuous time : a non-parametric bayesian approach. *ASTIN Bulletin*, *26*(2), 139–164.

Hachemeister, C. (1980). A stochastic model for loss reserving. *Transactions of the 21st International Congress of Actuaries*, *1*, 185–194.

Hans, N., Klein, N., Faschingbauer, F. et Schneider, M. (2022). Boosting distributional copula regression. *Biometric Methodology*, *79*, 2298–2310.

Hastie, T. et Tibshirani, R. (1990). *Generalized Additive Models*. CRC Press.

Henckaerts, R., Antonio, K., Clijsters, K. et Verbelen, R. (2018). A data driven binning strategy for the construction of insurance tariff classes. *Scandinavian Actuarial Journal*, *2018*(1), 1–25.

Hobæk Haff, I., Aas, K. et Frigessi, A. (2010). On the simplified pair-copula construction - simply useful or too simplistic ? *Journal of Multivariate Analysis*, *101*(5), 1296–1310.

Hofert, M. (2008). Sampling archimedean copulas. *Computational Statistics and Data Analysis*, *52*(12), 5163–5174.

Huster, W., Brookmeyer, R. et Self, S. (1989). Modeling paired survival data with covariates. *Biometrics*, *45*, 145–156.

Joe, H. (1993). Parametric families of multivariate distributions with given margins. *Journal of Multivariate Analysis*, *46*(2), 262–282.

Kendall, M. (1938). A new measure of rank correlation. *Biometrika*, *20*(1), 81–93.

Klein, N. et Smith, M. (2021). Bayesian inference for regression copulas. *Journal of Business and Economic Statistics*, *39*(3), 712–728.

Kraus, D., Killiches, M. et Czado, C. (2017). Examination and visualisation of the simplifying assumption for

vine copulas in three dimensions. *Australia and New Zealand Journal of Statistics*, *59*(1), 95–117.

Lakhal-Chaieb, M. (2010). Copula inference under censoring. *Biometrika*, *97*(2), 505–512.

Levi, E. et Craiu, R. (2018). Bayesian inference for conditional copulas using gaussian process single index models. *Computational Statistics and Data Analysis*, *122*, 115–134.

Liu, G., Long, W., Yang, B. et Cai, Z. (2021). Semiparametric estimation and model selection for conditional mixture copula models. *Scandinavian Journal of Statistics*, *49*, 287–330.

Lopez, O. (2019). A censored copula model for micro-level claim reserving. *Insurance : Mathematics and Economics*, *87*(4), 1–14.

Lopez, O., Milhaud, X. et Thérond, P. (2016). Tree-based censored regression with applications in insurance. *Electronic Journal of Statistics*, *10*(2), 2685–2716.

M., S. et Stahl, G. (2021). The standard formula of solvency ii : a critical discussion. *European Actuarial Journal*, *11*, 3–20.

Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin*, *23*(2), 213–225.

Mack, T. (1994). Which stochastic model is underlying the chain ladder model ? *Insurance : Mathematics and Economics*, *15*, 133–138.

Mack, T. (1999). The standard error of chain ladder reserve estimates : Recursive calculation and inclusion of a tail factor. *ASTIN Bulletin*, *29*(3), 361–366.

Mack, T. et Venter, G. (2000). A comparison of stochastic models that reproduce chain ladder reserve estimates. *Insurance : Mathematics and Economics*, *26*(1), 101–107.

Marshall, A. et Olkin, I. (1988). Families of multivariate distributions. *Journal of the American Statistical Association*, *83*, 834–841.

Michaelides, M., Pigeon, M. et Cossette, H. (2023). Individual claims reserving using activation patterns. *European Actuarial Journal*, *13*(2), 837–869.

Nelder, J. et Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, *135*(3), 370–384.

Nelsen, R. (2006). *An Introduction to Copulas* (2 éd.). New York : Springer.

Norberg, R. (1986). A contribution to modeling of ibnr claims. *Scandinavian Actuarial Journal*, *1986*(3-4), 155–203.

Norberg, R. (1993). Prediction of outstanding liabilities in non-life insurance. *ASTIN Bulletin*, *23*(1), 95–115.

Norberg, R. (1999). Prediction of outstanding liabilities ii. model variations and extensions. *ASTIN Bulletin*, *29*(1), 5–25.

Patton, A. (2006). Modeling asymmetric exchange rate dependence. *Int. Econ. Rev. (Philadelphia)*, *47*, 527–556.

Pešta, M. et Okhrin, O. (2014). Conditional least squares and copulae in claims reserving for a single line of business. *Insurance : Mathematics and Economics*, *56*, 28–37.

Pigeon, M., Antonio, K. et Denuit, M. (2013). Individual loss reserving with the multivariate skew normal framework. *ASTIN Bulletin*, *43*(3), 399–428.

Pigeon, M., Antonio, K. et Denuit, M. (2014). Individual loss reserving using paid-incurred data. *Insurance : Mathematics and Economics*, *58*, 121–131.

Pitt, M., Chan, D. et Kohn, R. (2006). Efficient bayesian inference for gaussian copula regression models. *Biometrika*, *93*(3), 537–554.

Ridout, M. (2009). Generating random numbers from a distribution specified by its laplace transform. *Statistics and Computing*, *19*(439).

Schweizer, B. et Sklar, A. (1983). *Probabilistic Metric Spaces. North-Holland Series in Probability and Applied Mathematics*. New York : North-Holland Publishing Co.

Segers, J. (2012). Asymptotics of empirical copula processes under non-restrictive smoothness assumptions. *Bernoulli*, *18*(3), 764–782.

Shi, P., Feng, X. et Boucher, J.-P. (2016). Multilevel modeling of insurance claims using copulas. *The Annals of Applied Statistics*, *10*(2), 834–863.

Shi, P. et Lee, G. (2022). Copula regression for compound distributions with endogenous covariates with application in insurance deductible pricing. *Journal of the American Statistical Association*, *117*(539), 1094–1109.

Shi, P. et Shi, K. (2022). Non-life insurance risk classification using categorical embedding. *North American Actuarial Journal*, 1–23.

Shi, P. et Yang, L. (2018). Pair copula constructions for insurance experience rating. *Journal of the American Statistical Association*, *113*, 122–133.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris 8*, 229–231.

Spanhel, F. et Kurz, M. (2019). Simplified vine copula models : Approximations based on the simplifying assumption. *Electronic Journal of Statistics*, *13*(1), 1254–1291.

Spanhel, F. et Kurz, M. (2022). Testing the simplifying assumption in high-dimensional vine copulas. *Electronic Journal of Statistics*, *16*, 5226–5276.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101.

Stasinopoulos, M., Ridgby, R., Heller, G., Voudouris, V. et De Bastinani, D. (2017). *Flexible regression and*

smoothing : using gamlss in R. New York : Chapman and Hall/CRC.

Stöber, J., Joe, H. et Czado, C. (2013). Simplified pair copula constructions - limitations and extensions. *Journal of Multivariate Analysis*, *119*, 101–118.

Sun, T., Cheng, Y. et Ding, Y. (2023). An information ratio-based goodness-of-fit test for copula models on censored data. *Biometrics*, *79*(3), 1713–1725.

Valle, L., Leisen, F. et Rossini, L. (2017). Bayesian nonparametric conditional copula estimation of twin data. *SSRN Electronic Journal*.

Van Buuren, S. et Groothuis-Oudshoorn, K. (2011). mice : Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*(3), 1–67.

Wang, A. (2010). Goodness-of-fit tests for archimedean copula models. *Statistica Sinica*, *20*, 441–453.

Wang, W. et Wells, M. (2000). Model selection and semiparametric inference for bivariate failure-time data. *Journal of the American Statistical Association*, *95*(449), 62–72.

Wang, Y., Oka, T. et Zhu, D. (2023). Bivariate distribution regression with application to insurance data. *Insurance : Mathematics and Economics*, *113*, 215–232.

Wei, Y., Wojtys, M., Sorrell, L. et Rowe, P. (2023). Bivariate copula regression models for semi-competing risks. *Statistical Methods in Medical Research*, *32*(10), 1902–1918.

Wüthrich, M. V. (2018). Machine learning in individual claims reserving. *Scandinavian Actuarial Journal*, *2018*(6), 465–480.

Yang, L. (2022). Nonparametric copula estimation for mixed insurance claim data. *Journal of Business and Economic Statistics*, *40*(2), 537–546.

Yang, L., Frees, E. et Zhang, Z. (2020). Nonparametric estimation of copula regression models with discrete outcomes. *Journal of the American Statistical Association*, *115*(530), 707–720.

Yang, L. et Shi, P. (2019). Multiperil rate making for property insurance using longitudinal data. *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, *182*(2), 647–668.

Yilmaz, Y. et Lawless, J. (2011). Likelihood ratio procedures and tests of fit in parametric and semiparametric copula models with censored data. *Lifetime Data Analysis*, *17*(3), 386–408.

Zhou, Q. (2021). Information matrix equivalence in the presence of censoring : A goodness-of-fit test for semiparametric copula models with multivariate survival data. *arXiv :2109.09782*.

Zhou, X. et Zhao, X. (2010). Applying copula models to individual claim loss reserving methods. *Insurance : Mathematics and Economics*, *46*(2), 10.