DAVID BEAUDOIN

ESTIMATION DE LA DÉPENDANCE ET CHOIX DE MODÈLES POUR DES DONNÉES BIVARIÉES SUJETTES À CENSURE ET À TRONCATION

Thèse présentée
à la Faculté des études supérieures de l'Université Laval
dans le cadre du programme de doctorat mathématiques, spécialisation statistique
pour l'obtention du grade de Philosophiae Doctor (Ph.D.)

Faculté des sciences et de génie
UNIVERSITÉ LAVAL
QUÉBEC

Juillet 2007

Résumé

Cette thèse contribue à l'analyse de durées de vie bivariées. Elle s'appuie sur quatre articles, rédigés en collaboration avec Christian Genest (directeur), Thierry Duchesne (co-directeur) ou d'autres collaborateurs (Lajmi Lakhal-Chaieb, Bruno Rémillard, Louis-Paul Rivest).

Le premier article, soumis à *Insurance : Mathematics and Economics* en novembre dernier, propose deux nouveaux tests d'adéquation d'un modèle de copules pour une paire (X, Y) de durées de vie. Leur performance est comparée à celle de six tests omnibus par voie de simulation.

Le second article, à paraître dans *Computational Statistics and Data Analysis*, propose de nouveaux estimateurs du tau de Kendall entre les variables X et Y, lorsque seule la seconde est sujette à censure. Une étude de Monte-Carlo montre que parce qu'ils exploitent l'information conditionnelle entre les variables, ces nouveaux estimateurs sont plus performants que ceux couramment utilisés dans ce contexte.

Le troisième article, en cours de révision pour *Lifetime Data Analysis*, propose un estimateur de type Horvitz–Thompson pour $\tau(X, Y)$ lorsque les deux variables sont sujettes à censure. On démontre qu'au contraire des estimateurs existants, celui-ci demeure convergent même quand $\tau \neq 0$.

Le quatrième article, soumis à *Statistics in Medicine*, présente un critère de sélection de modèle lorsque la paire (X, Y) n'est observable que dans la région $Y > X$ et que Y est sujette à censure à droite. La procédure compare une estimation paramétrique à une estimation non paramétrique de la version tronquée du tau de Kendall.

Abstract

This thesis contributes to bivariate survival data analysis. It is based on four papers, written jointly with Christian Genest (supervisor), Thierry Duchesne (co-supervisor) or other collaborators (Lajmi Lakhal-Chaieb, Bruno Rémillard, Louis-Paul Rivest).

The first paper, submitted to *Insurance : Mathematics and Economics* last November, proposes two new statistics for goodness-of-fit testing of a copula model for a pair (X, Y) of lifetimes. Their performance is compared to that of six omnibus tests through simulation.

The second paper, which is due to appear in *Computational Statistics and Data Analysis*, proposes new estimators of Kendall's tau between variables X and Y when only the second is subject to censoring. A Monte Carlo study shows that because they take into account the conditional information between the variables, these new estimators perform better than those currently used in this context.

The third paper, currently under revision for *Lifetime Data Analysis*, proposes a Horvitz–Thompson type estimator for $\tau(X, Y)$ when both variables are subject to censoring. It is shown that by opposition to existing estimators, this one remains consistent even when $\tau \neq 0$.

The fourth paper, submitted to *Statistics in Medicine*, presents a model selection criterion when the pair (X, Y) can only be observed in the region $Y > X$ and Y is subject to right censoring. The procedure compares a parametric estimate to a nonparametric estimate of the truncated version of Kendall's tau.

Avant-propos

Cette thèse est constituée de quatre articles. Le premier (présenté au chapitre 3) est le fruit de travaux réalisés en collaboration avec Christian Genest et Bruno Rémillard, respectivement professeurs à l'Université Laval et à HEC Montréal. Christian et un autre professeur de Laval, Thierry Duchesne, co-signent le deuxième article (chapitre 4). Le troisième article (chapitre 5) résulte d'une collaboration avec Lajmi Lakhal-Chaieb et Louis-Paul Rivest, également professeurs à Laval. Enfin, un autre projet mené avec Lajmi a conduit au quatrième article (chapitre 6).

Je souhaite exprimer ici ma reconnaissance envers mes co-directeurs de thèse, Christian Genest et Thierry Duchesne, qui ont su me guider d'une main de maître au fil de mes recherches. Christian, je te sais gré de m'avoir enseigné la rigueur et le souci du détail, de ton ouverture envers mes idées, sans oublier ta bonne humeur lors de nos nombreuses rencontres empreintes d'humour ! Thierry, merci de ta disponibilité, de ta grande patience et de tes conseils fort judicieux !

Je tiens également à exprimer ma gratitude envers Lajmi Lakhal-Chaieb avec lequel le partage d'idées a été florissant, sans oublier l'aide mutuelle que nous nous sommes accordée au cours des trois dernières années. Je remercie en outre Bruno Rémillard et Louis-Paul Rivest de leur bonne collaboration.

Ma première expérience d'enseignement a été rendue possible grâce à Jean-Pierre Carmichael, lui qui a aussi participé à ma formation au Service de consultation statistique de l'Université Laval en compagnie de Gaétan Daigle, Hélène Crépeau et Sophie Baillargeon. À vous quatre, un grand merci !

Sur le plan personnel, je désire également remercier chaleureusement mes parents, Annie et Pierre, pour leur appui et leur amour sans limite. Ils ont su inculquer en moi de belles valeurs, dont la persévérance, le désir de se dépasser et le perfectionnisme. Sans eux, je ne me serais jamais rendu aussi loin. Je désire également exprimer ma reconnaissance envers mon frère Patrick qui, malgré la distance qui nous sépare, a su

m'encourager et me conseiller au fil du temps. Je remercie aussi P. "Le Mammouth" Malenfant, Jvaozex, Mister S.A.M., Doudou/Blondy, ainsi que tous mes coéquipiers de balle molle (Go Meunier ! Champions 2006 !) qui m'ont permis de sortir du cadre professionnel !

Au cours de mes études supérieures, j'ai bénéficié de financements octroyés par le Conseil de recherches en sciences naturelles et en génie du Canada, ainsi que par le Fonds québécois de la recherche sur la nature et les technologies. Qu'ils trouvent ici l'expression de ma reconnaissance.

Table des matières

Liste des tableaux

Table des figures

Chapitre 1

Introduction

On observe des durées de vie dans une vaste gamme de champs d'application. Il est donc très important de disposer de bonnes méthodes d'analyse pour ce type de données. La tâche est d'autant plus complexe lorsque certaines de ces observations sont incomplètes, soit en raison du phénomène de censure ou de troncation.

Sur le plan mathématique, une durée de vie n'est rien d'autre qu'une variable aléatoire non négative. Ce type de variable est fréquent, notamment en médecine, en biologie, en épidémiologie, en finance, en actuariat et en fiabilité. Plusieurs auteurs se sont penchés sur des problèmes liés à l'analyse des durées de vie dans le cas bivarié. Brown et al. (1974) ont investigué les variables qui peuvent être liées à la survie de patients suite à une transplantation cardiaque. Isobe et al. (1986) ont exposé aux astronomes les techniques statistiques d'analyse de survie. Notons également les travaux de Hougaard et al. (1992) qui ont étudié l'âge au décès chez des jumeaux.

On désire fréquemment modéliser la loi jointe de deux variables aléatoires positives X et Y. À cette fin, les copules s'avèrent très utiles et leur popularité n'a cessé de croître depuis leur introduction dans le monde de la statistique, à partir des travaux de Genest & MacKay (1986). Les copules permettent en effet de saisir les éléments essentiels de la dépendance entre deux variables, et ce, indépendamment de l'effet de leurs lois marginales. À l'opposé, un chercheur choisissant, par exemple, la loi normale bivariée comme modèle de la loi jointe des variables X et Y serait forcé de prendre la loi normale comme lois marginales pour ces deux variables. Ce problème est évité par le biais des copules puisque les choix de loi univariée pour X, de loi univariée pour Y et de copule peuvent se faire en trois étapes totalement distinctes et indépendantes !

En présence de deux variables aléatoires X et Y, on peut décrire une copule comme étant une fonction de répartition bivariée qui lie la loi jointe de X et de Y à ses lois marginales. En termes mathématiques, si H est la fonction de répartition conjointe de X et de Y dont les lois univariées sont notées F et G, le théorème de représentation de Sklar (1959) dit qu'il existe une copule C telle que pour tous couples $(x, y) \in \mathbb{R}^2$,

$$H(x, y) = C\left\{F(x), G(y)\right\}.$$

À la suite de l'étape de modélisation par une copule, il semble naturel de vouloir tester l'adéquation du modèle choisi en se basant sur le jeu de données à notre disposition. Ce problème a grandement suscité l'intérêt des chercheurs. Notons entre autres les travaux de Genest & Rivest (1993), Wang & Wells (2000b), Breymann et al. (2003), Fermanian (2005) et Genest et al. (2006). Si on s'intéresse à l'adéquation de la copule normale, ajoutons à cette liste l'article de Malevergne & Sornette (2003).

Le premier article de cette thèse compare le seuil et la puissance de huit tests d'adéquation de modèles de copule. Sept modèles sont considérés dans le cadre d'une vaste étude de Monte-Carlo. Les deux premiers tests qui nous intéressent se basent sur la copule empirique et on peut retrouver les détails de ces tests dans Genest & Rémillard (2005). Le premier d'entre eux utilise la statistique de Cramér–von Mises, tandis que le second se fonde sur la statistique de Kolmogorov–Smirnov. Ces tests mènent à des seuils observés qui permettent d'accepter ou de rejeter l'adéquation d'une copule donnée via une procédure de bootstrap paramétrique. Les troisième et quatrième tests étudiés emploient une stratégie similaire, mais cette fois à l'aide du processus de Kendall (Genest et al., 2006). Le cinquième test applique diverses transformations aux données de façon à ce que la statistique de test soit asymptotiquement khi-deux, du moins approximativement. Cette procédure, détaillée dans Breymann et al. (2003), est légèrement modifiée par Dobrić & Schmid (2007) ce qui constitue notre sixième test. Quant aux nouvelles méthodes d'adéquation, elles font appel à la transformation de Rosenblatt afin de rendre les points d'un même couple indépendants. On vérifie ensuite la proximité entre la copule empirique et la copule d'indépendance en chacun des points transformés. Une fois de plus, le bootstrap paramétrique fournit un seuil observé en vue de tester la validité d'un modèle donné.

Les méthodes d'inférence se compliquent grandement en présence de données de survie. Ces difficultés peuvent être causées par deux sources : la censure et la troncation. Plus précisément, lorsqu'une observation subit une censure, sa valeur exacte nous est alors inconnue mais nous pouvons tout de même en tirer une certaine information. En effet, un intervalle de temps possible sur la vraie valeur de l'observation nous est fourni. Quant à la troncation, elle fait en sorte que seuls les individus dont la durée de vie tombe dans un certain intervalle de temps sont observés et donc inclus dans le jeu de données.

Voici l'information partielle dont nous disposons dans les trois types majeurs de censure :

(i) Censure à droite : La vraie valeur de l'observation est supérieure à un certain nombre c (on sait que $X > c$).
(ii) Censure à gauche : La vraie valeur de l'observation est inférieure à un certain nombre c (on sait que $X < c$).
(iii) Censure par intervalle : La vraie valeur de l'observation est située entre deux nombres c_1 et c_2 (on sait que $X \in [c_1, c_2]$).

Par ailleurs, en pratique on rencontre généralement deux sortes de troncation :

(i) Troncation à gauche : Seules les observations dont la valeur est supérieure à un nombre c peuvent être incluses dans le jeu de données.
(ii) Troncation à droite : Seules les observations dont la valeur est inférieure à un nombre c peuvent être incluses dans le jeu de données.

Les deuxième et troisième articles de cette thèse traitent de l'estimation du tau de Kendall, selon que l'une des deux variables est sujette à la censure à droite ou que les deux variables peuvent être censurées. L'intérêt du tau de Kendall vient du fait que sa définition et son estimation ne supposent aucune forme paramétrique pour les lois marginales de X et de Y, ni pour leur association. Il donne une première piste au chercheur sur la présence ou l'absence d'association entre deux variables avant que celui-ci ne se lance dans une étude plus poussée concernant ces variables. Le problème majeur lié à l'estimation de τ dans un contexte de censure provient de la nature concordante/discordante inconnue de certaines paires d'observations à cause de la censure.

Le deuxième article se restreint au cas où seule la variable Y peut subir le phénomène de censure à droite. À l'exception de l'estimateur de Wang & Wells (2000a), tous les estimateurs déjà existants n'utilisent que l'information contenue dans les lois univariées. Ceci engendre un biais important dans l'estimation de τ par ces méthodes lorsque la dépendance entre X et Y est non nulle. Nous proposons de nouvelles méthodes qui font usage d'un estimateur de la loi conditionnelle de Y sachant X. Cela fait en sorte que l'information jointe est prise en compte, permettant ainsi d'améliorer considérablement la précision des estimateurs (tel que montré au moyen de simulations).

Le troisième article réussit à éliminer complètement le biais dans l'estimation de τ lorsque les deux variables sont sujettes à la censure à droite. Afin d'y parvenir, un estimateur du type Horvitz–Thompson est défini en attribuant à chaque paire d'observations un poids qui est égal à l'inverse de la probabilité que cette paire soit ordorable.

On retrouve également dans cet article une démonstration de la normalité asymptotique de notre méthode. Une comparaison avec certaines procédures existantes est conduite à l'aide de simulations. Ce travail couvre quatre schémas de censure : une seule variable censurée, une censure bivariée unique pour les deux variables, une censure bivariée indépendante et une censure bivariée dépendante.

Le quatrième article vise principalement à introduire un critère de sélection de modèle lorsqu'une variable d'intérêt Y est soumise à la troncation dépendante à gauche et à la censure indépendante à droite. Plus spécifiquement, on n'observe que les points tels que $Y > X$ (troncation dépendante à gauche) ; de plus, la variable Y peut être censurée par une variable C qui ne dépend nullement de la valeur de Y (censure indépendante à droite). Les données observées prennent donc la forme (X_i, T_i, δ_i) avec $T_i = \min(Y_i, C_i)$, $\delta_i = 1(Y_i \leq C_i)$ et $T_i > X_i$ pour $i = 1, \ldots, n$. Un exemple concret d'une telle situation est décrit à la section 1.16 de Klein & Moeschberger (1997). On y définit Y comme étant l'âge au décès de résidents de la maison de retraite Channing House en Californie. Cette variable est tronquée à gauche par la variable X qui représente l'âge de l'individu lors de son entrée dans le centre. En effet, une personne ayant fait son entrée à Channing House à l'âge X devait survivre au moins jusqu'à cet âge afin de faire partie de la base de données ($Y > X$ obligatoirement). Le phénomène de censure à droite peut intervenir pour plusieurs raisons, notamment la fin de l'étude ou le départ d'un résident (c'est-à-dire son retrait de l'étude).

Dans cet article, une copule de semi-survie est employée afin de modéliser le lien qui existe entre X et Y :

$$\Pr(X \leq x, Y > y | \, Y > X) = \frac{C_\alpha \left\{F_X(x), S_Y(y)\right\}}{c}, \qquad y \geq x$$

où C_α est une copule archimédienne (voir la définition de cette classe de copules dans la section 2.1.3) et c est une constante de normalisation. Notons également que $F_X(x) = \Pr(X \leq x)$ et $S_Y(y) = \Pr(Y > y)$.

Le paramètre α d'association de la copule ainsi que la constante c sont estimés via la méthode décrite dans Lakhal-Chaieb et al. (2006). Puis, on tire profit de ces estimations afin de calculer une statistique dont la valeur sera faible si le modèle de copule choisi est adéquat pour le jeu de données à l'étude. Un chercheur en quête du meilleur modèle archimédien possible en lien avec ses données peut donc calculer cette statistique sous une flopée de structures de dépendance et ainsi retenir celle qui mène à la valeur minimale de la statistique. Cette dernière mesure en fait une distance entre le tau tronqué paramétrique de Manatunga & Oakes (1996) et un estimateur non paramétrique de ce même tau. Une étude de simulation démontre l'efficacité de ce critère de sélection.

Chapitre 2

Préliminaires

2.1 Copules

Beaucoup de situations pratiques nécessitent l'étude de la nature du lien qui existe entre plusieurs variables aléatoires. Le problème qui consiste à vérifier l'adéquation d'une distribution multivariée est fort complexe. Une façon de faciliter les choses est de considérer les variables par paires, ce qui réduit le problème au cas bivarié. Et même dans le cas bivarié, tout n'est pas trivial !

Les copules s'avèrent un outil intéressant en ce qui a trait à la modélisation multivariée, particulièrement dans l'éventualité où les lois marginales ne suivent pas la loi gaussienne. Nous concentrerons notre énergie sur l'analyse simultanée de deux facteurs, c'est-à-dire le cas à deux dimensions. Plusieurs modèles de copules bivariées ont été développés au cours des années. Cette section a pour but de mettre en valeur les faits saillants concernant ce type précis de modèle mathématique.

2.1.1 Définitions et propriétés

Supposons que l'on s'intéresse à deux variables aléatoires continues X et Y. Définissons les lois H, F et G de la manière suivante :

- $H(x, y) = \Pr(X \leq x, Y \leq y) =$ fonction de répartition jointe de X et de Y
- $F(x) = \Pr(X \leq x) =$ fonction de répartition marginale de X
- $G(y) = \Pr(Y \leq y) =$ fonction de répartition marginale de Y

Selon le théorème de Sklar (1959), il existe alors une copule unique C telle que

$$H(x,y) = C_\alpha \left\{F(x), G(y)\right\}. \tag{2.1}$$

En d'autres mots, la copule, qui dépend d'un paramètre d'association α, fait office de pont qui relie la loi jointe à ses marges. Par ailleurs, ce même théorème affirme que si l'on dispose d'une copule $C = C_\alpha$ et de fonctions de répartition F et G, alors H telle que définie à l'équation (2.1) est nécessairement une fonction de répartition bivariée.

On peut également voir une copule comme étant une fonction de répartition bivariée ayant des marges uniformes sur $[0,1]$. Elle possède les propriétés suivantes :

(i) $C_\alpha(w,0) = C_\alpha(0,w) = 0$ pour tout $w \in [0,1]$;
(ii) $C_\alpha(w,1) = C_\alpha(1,w) = w$ pour tout $w \in [0,1]$;
(iii) Si $(u_1,u_2,v_1,v_2) \in [0,1]^4$ avec $u_1 \leq u_2$ et $v_1 \leq v_2$ alors

$$C_\alpha(u_2,v_2) - C_\alpha(u_1,v_2) - C_\alpha(u_2,v_1) + C_\alpha(u_1,v_1) \geq 0.$$

(iv) Toute copule C_α est bornée par les copules W et M de telle sorte que

$$W(u,v) \leq C_\alpha(u,v) \leq M(u,v),$$

où pour tous $u,v \in [0,1]$,

$$W(u,v) = \max(u+v-1,0) \quad \text{et} \quad M(u,v) = \min(u,v).$$

Ces deux dernières copules portent respectivement le nom de borne inférieure et borne supérieure de Fréchet–Hoeffding.
(v) Les variables aléatoires X et Y régies par la copule C_α sont indépendantes si et seulement si $C_\alpha(u,v) = uv$ pour tous $u,v \in [0,1]$.
(vi) Si f et g sont des fonctions strictement croissantes, alors la copule liée au couple de variables aléatoires (X,Y) est la même que celle du couple $(f(X), g(Y))$.

En lien à la propriété (iv), notons que si la copule régissant les variables aléatoires X et Y est la borne inférieure W, alors cela signifie que Y est presque sûrement une fonction monotone décroissante de X. À l'inverse, si $C_\alpha = M$ alors Y est presque sûrement une fonction monotone croissante de X.

Notons que tous les éléments clés en regard de la dépendance entre deux variables aléatoires continues se retrouvent dans leur copule. La dépendance jointe de ces variables est caractérisée de façon complète et unique par la copule.

2.1.2 Copules de semi-survie

Le quatrième article de cette thèse fait appel aux copules de semi-survie. Dans ce travail, on se place dans un contexte où seules les paires satisfaisant la condition $Y > X$ sont observées (données tronquées). L'ensemble des individus à risque au temps t est donc formé des points tels que $X \leq t$ et $Y > t$. Nous utilisons un système à deux équations développé par Lakhal-Chaieb et al. (2006) qui considère la probabilité, pour un individu donné, d'être à risque de subir l'événement d'intérêt au temps t, soit $\Pr(X \leq t, Y > t)$. De façon plus générale, on s'intéresse à $\Pr(X \leq x, Y > y)$ que l'on peut développer ainsi :

$$\begin{aligned}\Pr(X \leq x, Y > y) &= \Pr(X \leq x) - \Pr(X \leq x, Y \leq y)\\ &= F_X(x) - C_\alpha\left\{\Pr(X \leq x), 1 - \Pr(Y > y)\right\}\\ &= \tilde{C}_\alpha\left\{F_X(x), \bar{G}_Y(y)\right\}\end{aligned}$$

où $\bar{G}_Y(y) = \Pr(Y > y)$ et $\tilde{C}_\alpha$ est définie comme étant la copule de semi-survie. En posant $u = F_X(x)$ et $v = \bar{G}_Y(y)$, on obtient $\tilde{C}_\alpha(u, v) = u - C_\alpha(u, 1 - v)$.

2.1.3 Copules archimédiennes

Un grand nombre de copules appartiennent à la classe dite *archimédienne*. Les membres de cette classe possèdent de belles propriétés. La modélisation statistique à l'aide de copules archimédiennes a été popularisée par Genest & MacKay (1986). En fait, une copule C_α appartient à cette classe si elle prend la forme

$$C_\alpha(u, v) = \phi_\alpha^{-1}\left\{\phi_\alpha(u) + \phi_\alpha(v)\right\},$$

où $\phi : [0, 1] \to [0, \infty]$ est une fonction décroissante convexe telle que $\phi(1) = 0$. La fonction ϕ est appelée le *générateur* de la copule.

La section 2.2 définit le tau de Kendall et nous verrons au paragraphe 2.2.2 qu'une formule simple existe afin de le calculer pour une copule archimédienne. Par ailleurs, le premier article de cette thèse fait référence à la fonction de répartition de Kendall, K, qui est définie comme étant la fonction de répartition de la variable aléatoire $C_\alpha(U, V)$. Mathématiquement parlant,

$$K_\alpha(t) = \Pr\{C_\alpha(U, V) \leq t\}, \quad t \in (0, 1).$$

Or, il existe une formule simple pour le calcul de cette fonction dans le cadre des copules archimédiennes. En effet, pour les copules appartenant à cette classe on a

$$K_\alpha(t) = t - \frac{\phi_\alpha(t)}{\phi'_\alpha(t)}, \quad t \in (0, 1).$$

2.1.4 Exemples de copules

Nous listons ici les dix copules utilisées dans le cadre de cette thèse. Pour chacune d'entre elles, nous fournissons en annexe les renseignements suivants :

- Appartenance à la classe archimédienne
- Générateur de la copule (si archimédienne)
- Valeur du tau de Kendall, τ
- Fonction de répartition de Kendall, K
- Algorithme de génération de données

Voici donc les dix copules en question :

1. Clayton : Pour $\alpha \geq -1$,
$$C_\alpha(u,v) = \max\left\{(u^{-\alpha} + v^{-\alpha} - 1)^{-1/\alpha}, 0\right\}.$$
2. Frank : Pour $\alpha \in \mathbb{R}$,
$$C_\alpha(u,v) = -\frac{1}{\alpha}\ \ln\left\{1 + \frac{(e^{-\alpha u} - 1)\,(e^{-\alpha v} - 1)}{(e^{-\alpha} - 1)}\right\}.$$
3. Gumbel–Hougaard : Pour $\alpha \geq 1$,
$$C_\alpha(u,v) = \exp[-\left\{(-\ln u)^\alpha + (-\ln v)^\alpha\right\}^{1/\alpha}].$$
4. Plackett : Pour $\alpha > 0$,
$$C_\alpha(u,v) = \frac{1 + (\alpha-1)(u+v) - \sqrt{\{1 + (\alpha-1)(u+v)\}^2 - 4uv\alpha(\alpha-1)}}{2(\alpha-1)}.$$
5. Normale : Pour $\rho \in [-1,1]$,
$$C_\rho(u,v) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \exp\left\{-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right\} dydx,$$
où Φ est la fonction de répartition de la loi normale centrée réduite.
6. Student : Pour $\rho \in [-1,1]$,
$$C_{\rho,\nu}(u,v) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{t_\nu^{-1}(u)} \int_{-\infty}^{t_\nu^{-1}(v)} \left\{1 + \frac{x^2 - 2\rho xy + y^2}{\nu(1-\rho^2)}\right\}^{-\frac{\nu+2}{2}} dydx,$$
où ν est le nombre de degrés de liberté et t_ν est la fonction de répartition de la loi de Student à ν degrés de liberté.

7. Pearson type II : Pour $\rho \in [-1,1]$,

$$C_{\rho,\nu}(u,v) = \int\int_{B} \frac{\nu}{\sqrt{1-\rho^2}} \frac{1}{\pi A^{2\nu}} \{A^2 - q(x,y)^2\}^{\nu-1} dydx$$

où

$$q(x,y) = \sqrt{\frac{x^2+y^2-2\rho xy}{1-\rho^2}},$$

$A > 0$ est un paramètre et

$$B = \left\{(x,y) : x \leq F^{-1}(u), y \leq F^{-1}(v), |q| < A\right\}.$$

Ici, F est la fonction de répartition des marges. Par exemple, si l'on fixe $A = 1$ on obtient :

$$F(x) = \frac{\Gamma(\nu+1.5)}{\sqrt{\pi}\Gamma(\nu+1)} \int_{-1}^{x} (1-u^2)^{\nu} du, \quad \text{pour } x \in [-1,1].$$

8. Copule 3 du quatrième article de cette thèse : Pour $\alpha \geq 1$,

$$C_\alpha(u,v) = \max([1 - \left\{(1-u^{1/\alpha})^\alpha + (1-v^{1/\alpha})^\alpha\right\}^{1/\alpha}]^\alpha, 0).$$

9. Copule 4 du quatrième article de cette thèse : Pour $\alpha \in [0,1]$,

$$C_\alpha(u,v) = \max\left\{uv - \alpha(1-u)(1-v), 0\right\}.$$

10. Copule 5 du quatrième article de cette thèse : Pour $\alpha \geq 1$,

$$C_\alpha(u,v) = \max\left\{\frac{\alpha^2 uv - (1-u)(1-v)}{\alpha^2 - (\alpha-1)^2(1-u)(1-v)}, 0\right\}.$$

2.2 Tau de Kendall

Le tau de Kendall, τ, mesure la force d'association entre deux variables aléatoires X et Y. Il s'agit en fait d'un nombre réel dont la valeur est comprise dans l'intervalle $[-1,1]$. Une valeur positive de τ suggère la présence d'une dépendance positive entre les variables, une valeur négative indique une dépendance négative, tandis qu'une valeur près de zéro suggère l'indépendance des variables à l'étude. En fait, $\tau = 1$ (respectivement -1) si et seulement si Y est une fonction monotone croissante (respectivement décroissante) de X.

2.2.1 Tau de Kendall empirique

La notion de concordance/discordance est essentielle à la compréhension de la définition de τ. De façon intuitive, une paire (X, Y) de variables aléatoires est concordante si une grande valeur de X est souvent associée à une grande valeur de Y, et si une petite valeur de X a de fortes chances d'être liée à une petite valeur de Y. En termes mathématiques, la nature des paires (x_1, y_1) et (x_2, y_2) est dite

1. *concordante* si l'un des deux cas suivants est rencontré :
   (i) $x_1 > x_2$ et $y_1 > y_2$,
   (ii) $x_1 < x_2$ et $y_1 < y_2$,
   c'est-à-dire si $(x_1 - x_2)(y_1 - y_2) > 0$.
2. *discordante* si l'un des deux cas suivants est rencontré :
   (i) $x_1 > x_2$ et $y_1 < y_2$
   (ii) $x_1 < x_2$ et $y_1 > y_2$
   c'est-à-dire si $(x_1 - x_2)(y_1 - y_2) < 0$.

Supposons que l'on dispose d'un échantillon de variables aléatoires continues

$$(x_1, y_1), \ldots, (x_n, y_n)$$

de taille n. Si l'on note c comme étant le nombre de paires d'observations concordantes et d le nombre de paires discordantes dans ce jeu de données, alors on définit le tau de Kendall empirique comme suit :

$$\hat{\tau}_n = \frac{c - d}{c + d} = \frac{c - d}{\binom{n}{2}} .$$

Notons que $\hat{\tau}_n$ ne dépend donc que des *rangs* des observations.

2.2.2 Tau de Kendall dans la population

Soient (X, Y), (X_1, Y_1), (X_2, Y_2) trois observations indépendantes de loi H. Supposons que la copule C associée à H soit unique. La valeur théorique du tau de Kendall est alors donnée par

$$\begin{aligned} \tau(X, Y) &= \Pr\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - \Pr\{(X_1 - X_2)(Y_1 - Y_2) < 0\} \\ &= 4 \int_0^1 \int_0^1 C(u, v) dC(u, v) - 1 = 4\mathrm{E}\{C(U, V)\} - 1, \end{aligned}$$

où la paire (U, V) est de loi C.

Nous sommes donc en mesure de conclure que τ ne dépend pas des lois marginales, mais uniquement de la copule. Dans le cadre archimédien, le tau de Kendall peut se calculer aisément à partir du générateur ϕ décrit au paragraphe 2.1.3. Il peut être démontré (Nelsen, 1999) que pour cette classe importante de copules, on a la relation suivante :

$$\tau = 1 + 4 \int_0^1 \frac{\phi_\alpha(t)}{\phi'_\alpha(t)}\, dt.$$

2.3 Estimateur de la fonction de survie

On désire fréquemment estimer la fonction de survie d'une variable d'intérêt Y qui est sujette à la censure à droite. Nous présenterons dans un premier temps le célèbre estimateur de Kaplan & Meier (1958). Puis, nous décrirons l'estimateur de Kaplan–Meier généralisé qui permet d'estimer la survie conditionnelle de Y sachant que $X = x$, où la variable X peut être vue comme une covariable (non censurée). Notons que dans les deux cas, il est impératif de supposer que Y et sa variable de censure sont indépendantes. Si cette hypothèse est violée, l'estimateur qui en résulte peut être biaisé.

2.3.1 Estimateur de Kaplan–Meier standard

Supposons que l'on dispose d'un jeu de données de la forme $(T_1, \delta_1), \ldots, (T_n, \delta_n)$, où $T_i = \min(Y_i, C_i)$ et $\delta_i = 1(Y_i \leq C_i)$ pour $i = 1, \ldots, n$. Ainsi, T peut être vue comme étant la variable observée, Y est la vraie valeur de la durée de vie et C joue le rôle de la variable de censure. On désire estimer la survie de Y à partir de ces observations incomplètes. En supposant que des égalités sont possibles, la première étape consiste à ordonner les temps de décès observés (les observations pour lesquelles $\delta_i = 1$) distincts :

$$t_1^* < \cdots < t_D^*.$$

Pour chacun d'entre eux, on définit les deux variables suivantes :

$$\begin{aligned} d_i &= \text{nombre de décès observés au temps } t_i^*, \\ R_i &= \text{nombre d'individus qui étaient à risque de décéder au temps } t_i^* \\ &= \sum_{j=1}^{n} 1(t_j \geq t_i^*). \end{aligned}$$

L'estimateur de Kaplan–Meier est alors obtenu de la façon suivante :

$$\hat{S}_Y(t) = \begin{cases} 1 & \text{si } t < t_1^*; \\ \prod_{i:t_i^* \leq t} \left(1 - \frac{d_i}{R_i}\right) & \text{si } t \geq t_1^*. \end{cases}$$

Cet estimateur est bien défini sur le domaine $t \leq t_D^*$. Cela ne cause aucun problème lorsque la plus grande observation est un décès, car $\hat{S}_Y$ vaut 0 à droite de t_D^*. Par contre, si $\max(t_1, \ldots, t_n)$ est censurée, alors $\hat{S}_Y(t)$ est indéterminée à droite de cette observation. Un autre point à noter est le fait qu'en l'absence de censure, cet estimateur se réduit à la fonction de survie empirique usuelle. Il a été démontré que l'estimateur de Kaplan–Meier est convergent, qu'il tend vers une loi normale sous certaines conditions de régularité et qu'il est en fait l'estimateur du maximum de vraisemblance non paramétrique.

2.3.2 Estimateur de Kaplan–Meier généralisé

Une covariable non sujette à la censure est ajoutée à nos observations, de sorte que les données prennent maintenant la forme (X_i, T_i, δ_i) pour $i = 1, \ldots, n$ (les définitions sont les mêmes qu'au paragraphe 2.3.1 pour T_i et δ_i). Notre but consiste maintenant en l'estimation de la survie conditionnelle de Y sachant $X = x$, c'est-à-dire l'estimation de $\Pr(Y > t | X = x)$. L'estimateur de Kaplan–Meier généralisé a été défini par Beran (1981) comme suit :

$$\hat{S}_{KMG}(t|x) = \begin{cases} \prod_{i:t_i^* \leq t} \left\{ 1 - \frac{B_i(x)}{\sum_{j=1}^{n} 1(t_j \geq t) B_j(x)} \right\} & \text{si } t < t_{(n)}; \\ 0 & \text{si } t \geq t_{(n)}. \end{cases}$$

Dans la définition ci-dessus, les termes $B_i(x)$ représentent des poids. Ceux-ci doivent être non négatifs, en plus de satisfaire la condition

$$\sum_{i=1}^{n} B_i(x) = 1.$$

En fait, chaque observation se voit attribuer un poids en fonction de sa distance par rapport à la valeur de x. Nous faisons appel à l'estimateur $\hat{S}_{KMG}(t|x)$ dans le deuxième article de cette thèse et nous avons opté pour des poids du type Nadaraya–Watson qui s'écrivent sous la forme suivante :

$$B_i(x) = \frac{K\left(\dfrac{x - x_i}{h_X}\right)}{\sum_{j=1}^{n} K\left(\dfrac{x - x_j}{h_X}\right)}.$$

Ici, K est un noyau et h_X est une fenêtre de lissage dont la valeur doit être choisie judicieusement. Voici les noyaux les plus connus :

- Uniforme :
$$K(x) = \frac{1}{2}1(-1 \leq x \leq 1)$$
- Epanechnikov :
$$K(x) = \frac{3}{4}(1 - x^2)1(-1 \leq x \leq 1)$$
- Bipoids :
$$K(x) = \frac{15}{16}(1 - x^2)^2 1(-1 \leq x \leq 1)$$
- Tripoids :
$$K(x) = \frac{35}{32}(1 - x^2)^3 1(-1 \leq x \leq 1).$$

On remarque que l'estimateur de Kaplan–Meier généralisé est continu en x, mais en forme d'escalier par rapport à t. En fait, les sauts de cette fonction par rapport à t surviennent aux points de décès t_i^* si $B_i(x) \neq 0$. Mentionnons que Dabrowska (1987) a étudié en détail cet estimateur. Elle en a notamment démontré la convergence forte uniforme et la normalité asymptotique.

Dans le cadre du deuxième article de cette thèse, nous considérons deux variations de cet estimateur : l'une proposée par Leconte et al. (2002), l'autre suggérée par Van Keilegom & Akritas (1999). En fait, Leconte et al. (2002) proposent simplement une version du Kaplan–Meier généralisé qui est continue non seulement en x, mais également en t :

$$\hat{S}_{LPT}(y|x) = \sum_{i=1}^{I+1} \left\{ \hat{S}_{KMG}\left(T_{(i-1)}^{+}|x\right) - \hat{S}_{KMG}\left(T_{(i)}^{+}|x\right) \right\} L\left(\frac{y - T_{(i)}^{+}}{\omega}\right),$$

où $I = \delta_1 + \cdots + \delta_n$ est le nombre d'observations non censurées, $T_{(0)}^{+} = 0$, $T_{(i)}^{+}$ est la i^e plus petite valeur non censurée en Y, $i \in \{1, \ldots, I\}$, et $T_{(I+1)}^{+}$ est la plus grande valeur de T, qu'elle soit censurée ou non. De plus, on note que ω est une autre fenêtre et que

$$L(t) = \int_{-\infty}^{t} K(u)du.$$

Quant à l'estimateur de Van Keilegom & Akritas (1999), il se présente sous la forme d'un escalier en y et il est défini par

$$\hat{S}_{VKA}(y|x) = \hat{S}_e \left\{ \frac{y - \hat{m}(x)}{\hat{\sigma}(x)} \right\},$$

où $\hat{S}_e(t)$ est l'estimateur de Kaplan–Meier original basé sur les versions standardisées des valeurs de T. Ces dernières sont obtenues comme suit :

$$E_i = \frac{T_i - \hat{m}(X_i)}{\hat{\sigma}(X_i)},$$

où $\hat{m}(x)$ et $\hat{\sigma}^2(x)$ estiment respectivement $\mathrm{E}(Y|X = x)$ et $\mathrm{var}(Y|X = x)$.

Chapitre 3

Omnibus goodness-of-fit tests for copulas : A review and a power study

Résumé

Plusieurs procédures permettant de tester l'adéquation de modèles de copules ont été proposées récemment. Nous les passons brièvement en revue et centrons ensuite notre attention sur les tests dits omnibus, c'est-à-dire ceux dont l'implantation ne nécessite ni catégorisation arbitraire des données, ni choix stratégique lié à la présence d'un paramètre de lissage, d'un noyau, d'une fenêtre, d'une pondération, etc. Nous présentons une critique de telles méthodes, en plus d'en proposer de nouvelles. Nous décrivons et interprétons les résultats d'une vaste étude de Monte-Carlo conçue pour mesurer l'effet de la taille d'échantillon et du degré d'association sur le seuil et la puissance des tests omnibus, et ce, pour diverses combinaisons de modèles de copules sous l'hypothèse nulle et la contre-hypothèse. Pour contourner les difficultés liées à la détermination de la loi asymptotique des tests sous l'hypothèse nulle, nous recommandons l'utilisation d'une procédure de bootstrap paramétrique dont l'implantation est détaillée. Nous formulons en conclusion plusieurs recommandations pratiques.

Abstract

Many proposals have been made recently for goodness-of-fit testing of copula models. After reviewing them briefly, the authors concentrate on omnibus procedures, i.e., those whose implementation requires neither an arbitrary categorization of the data nor any strategic choice of smoothing parameter, weight function, kernel, window, etc. The authors present a critical review of these omnibus tests and suggest new ones. They describe and interpret the results of a large Monte Carlo experiment designed to assess the effect of the sample size and the strength of dependence on the level and power of the omnibus tests for various combinations of copula models under the null hypothesis and the alternative. To circumvent problems with inaccurate asymptotic approximation of the tests' limiting null distributions, they recommend the use of a double parametric bootstrap procedure, whose implementation is detailed. They conclude with a number of practical recommendations.

3.1 Introduction

Consider a continuous random vector $\mathbf{X} = (X_1, \ldots, X_d)$ with joint cumulative distribution function H and marginals $F_1, \ldots, F_d$. Following Sklar (1959), the copula representation of H is given by $H(x_1, \ldots, x_d) = C\{F_1(x_1), \ldots, F_d(x_d)\}$, where C is a unique multivariate distribution having uniform margins on the interval $(0, 1)$. A copula model for X arises when C is assumed to belong to a parametric class

$$\mathcal{C} = \{C_\theta : \theta \in \mathcal{O}\},$$

where $\mathcal{O}$ is an open subset of $\mathbb{R}^p$ for some integer $p \geq 1$. The books of Joe (1997), Nelsen (1999) and Drouet-Mari & Kotz (2001) provide handy compendiums of the most common parametric families of copulas.

Copula modeling has found many successful applications of late, notably in actuarial science, survival analysis and hydrology ; see, e.g., Frees & Valdez (1998), Cui & Sun (2004), Genest & Favre (2007) and references therein. However, nowhere has the methodology been adopted and used with greater intensity than in finance. Ample illustrations are provided in the books of Cherubini et al. (2004) and McNeil et al. (2005), notably in the context of asset pricing and credit risk management.

Given independent copies $\mathbf{X}_1 = (X_{11}, \ldots, X_{1d}), \ldots, \mathbf{X}_n = (X_{n1}, \ldots, X_{nd})$ of $\mathbf{X}$, the problem of estimating θ under the assumption

$$H_0 : C \in \mathcal{C}_0$$

has already been the object of much work ; see, e.g., Genest et al. (1995), Shih & Louis (1995), Joe (1997, 2005), Tsukahara (2005) or Chen et al. (2006). However, the complementary issue of testing H_0 is only beginning to draw attention.

The situation is evolving rapidly but at this point in time, the literature on the subject can be divided broadly into three groups :

1. Procedures developed for testing specific dependence structures such as the normal copula (Malevergne & Sornette, 2003) or the equally popular Clayton family, also referred to as the gamma frailty model in survival analysis (Shih, 1998; Glidden, 1999; Cui & Sun, 2004).

2. Statistics that can be used to test the goodness-of-fit of any class of copulas but whose implementation involves :
   (a) an arbitrary parameter, as in the rank-based statistic due to Wang & Wells (2000b) ;
   (b) kernels, weight functions and associated smoothing parameters, as in Berg & Bakken (2005), Fermanian (2005), Panchenko (2005) and Scaillet (2007) ;
   (c) ad hoc categorization of the data into a multi-way contingency table in order to apply an analogue of the standard chi-squared test, along the lines of Genest & Rivest (1993), Klugman & Parsa (1999), Andersen et al. (2005), Dobrić & Schmid (2005) or Junker & May (2005).
3. Truly omnibus tests, i.e., those applicable to all copula structures and requiring no strategic choice for their use. Included in this category are variants of the Wang–Wells approach due to Genest et al. (2006), but also the procedures investigated or used by Breymann et al. (2003), Genest & Rémillard (2005) and Dobrić & Schmid (2007).

And then there are some authors who, in applied work, have used standard goodness-of-fit statistics as a tool for choosing between several copulas, but without attempting to formally test whether the selected model is appropriate, in the light of a P-value. See, e.g., the analysis of stock index returns due to Ané & Kharoubi (2003).

The purpose of this paper is to present a critical review of the *omnibus* goodness-of-fit tests proposed to date, to suggest variants or improvements, and to compare the relative power of these procedures through a Monte Carlo study involving a large number of copula alternatives and dependence conditions. General considerations are stated in Section 3.2. Existing tests are described in Section 3.3 and new statistics are proposed in Section 3.4. Listed in Section 3.5 are the factors considered in the study designed to assess the level and compare the power of the selected tests. Results are reported and discussed in Section 3.6. Finally, various observations and methodological recommendations are made in the Conclusion.

3.2 General considerations

There is a fundamental difference between the problem of estimating the dependence parameter of a copula model $\mathcal{C} = \{C_\theta : \theta \in \mathcal{O}\}$ and the complementary issue of testing the validity of the null hypothesis $H_0 : C \in \mathcal{C}_0$ for some class $\mathcal{C}_0$ of copulas. The distinction is spelled out below, as it helps to understand the technical challenges associated with goodness-of-fit testing in this context.

3.2.1 Estimation

Two broad approaches to the estimation of the dependence parameter θ have been developed. They mainly differ in the user's willingness to make parametric assumptions about the unknown margins.

Given specific choices of parametric families $\mathcal{F}_j = \{F_{\gamma_j} : \gamma_j \in \Gamma_j\}$ of univariate distributions, estimation can proceed via the standard maximum likelihood method under the assumption that

$$H_0' : F_1 \in \mathcal{F}_1, \ldots, F_d \in \mathcal{F}_d.$$

An alternative technique that is computationally more convenient has been advocated by Joe (1997). Known as the IFM or "Inference Functions for Margins" approach, this technique proceeds in two steps : estimates $F_{\hat{\gamma}_1}, \ldots, F_{\hat{\gamma}_d}$ of the margins are first obtained under H_0' ; they are then plugged into the estimating function

$$\sum_{i=1}^{n} \log[c_\theta\{F_{\hat{\gamma}_1}(X_{i1}), \ldots, F_{\hat{\gamma}_d}(X_{id})\}],$$

in which c_θ denotes the density of the copula C_θ. As shown by Joe (2005), however, the gain in computational convenience often comes at the expense of efficiency. In addition, an inappropriate choice of models for the margins could have detrimental effects on the estimation of the dependence parameter per se.

If one is unwilling to assume H_0', nonparametric estimation of the margins must be used. The most natural choice consists in replacing F_j by its empirical counterpart

$$\hat{F}_j(t) = \frac{1}{n} \sum_{i=1}^{n} 1(X_{ij} \leq t)$$

and then basing estimation on the pseudo log-likelihood

$$\ell(\theta) = \sum_{i=1}^{n} \log[c_\theta\{\hat{F}_1(X_{i1}), \ldots, \hat{F}_d(X_{id})\}].$$

The asymptotic normality of the resulting estimator $\hat{\theta}$ is established by Genest et al. (1995), and by Shih & Louis (1995) in the presence of censorship. As shown by Genest & Werker (2002), however, this method is not asymptotically semi-parametrically efficient in general. See Klaassen & Wellner (1997) for a notable exception and Tsukahara (2005) or Chen et al. (2006) for other estimating strategies.

3.2.2 Goodness-of-fit testing

When testing the hypothesis $H_0 : C \in \mathcal{C}_0$ that the dependence structure of a multivariate distribution is well represented by a specific parametric family $\mathcal{C}_0$ of copulas, the option of modeling the margins by parametric families is no longer viable. For, it would be tantamount to testing the much narrower null hypothesis $H_0 \cap H_0'$ corresponding to a full parametric model. In this context, the marginal distributions $F_1, \ldots, F_d$ are nuisance parameters.

Given that the underlying copula C of a random vector is invariant by increasing transformations of its components, it appears that the only reasonable option for testing H_0 consists of basing the inference procedure on the maximally invariant statistics with respect to this set of transformations, i.e., the ranks

$$R_{ij} = \sum_{k=1}^{n} 1(X_{kj} \leq X_{ij}), \quad i \in \{1, \ldots, n\}, \quad j \in \{1, \ldots, d\}.$$

Indeed, all formal goodness-of-fit tests mentioned in the introduction are rank-based. Alternatively, they can be viewed as functions of the collection $\mathbf{U}_1 = (U_{11}, \ldots, U_{1d})$, …, $\mathbf{U}_n = (U_{n1}$, …, $U_{nd})$ of pseudo-observations deduced from the ranks, viz.

$$U_{ij} = \frac{R_{ij}}{n+1} = \frac{n}{n+1}\, \hat{F}_j(X_{ij}),$$

where the scaling factor $n/(n+1)$ is only introduced to avoid potential problems of convergence of the associated empirical process at the boundary of $(0, 1)$.

The pseudo-observations $\mathbf{U}_1, \ldots, \mathbf{U}_n$ can be interpreted as a sample from the underlying copula C. It is plain, however, that they are *not* mutually independent and that their components are only *approximately* uniform on $(0, 1)$. Accordingly, any inference procedure based on these constructs should take these features into account. As will be seen, testing procedures that mistakenly ignore these considerations not only lack power but fail to hold their nominal level.

3.3 Existing omnibus tests

This section describes five rank-based procedures that have been recently proposed for testing the goodness-of-fit of any class of d-variate copulas. Of all the tests listed in Section 3.1, these are the only ones that qualify as “omnibus,” in the sense that they involve no parameter tuning or other strategic choices.

3.3.1 Two tests based on the empirical copula

As mentioned in Section 3.2, the pseudo-observations $\mathbf{U}_1, \ldots, \mathbf{U}_n$ constitute the maximally invariant statistics on which to test $H_0 : C \in \mathcal{C}_0$. The information they contain can be conveniently summarized by the associated empirical distribution, viz.

$$C_n(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^{n} 1\left(U_{i1} \le u_1, \ldots, U_{id} \le u_d\right), \quad \mathbf{u} = (u_1, \ldots, u_d) \in (0,1)^d. \tag{3.1}$$

The latter is asymptotically equivalent to the empirical copula introduced by Deheuvels (1979), which assigns a weight of $1/n$ to the rescaled vectors $(n+1)\mathbf{U}_j/n$.

Gänßler & Stute (1987), Fermanian et al. (2004) and Tsukahara (2005) show under various conditions that C_n is a consistent estimator of the true underlying copula C, i.e., whether H_0 is true or not. Given that it is entirely nonparametric, C_n is arguably the most objective benchmark for testing $H_0 : C \in \mathcal{C}_0$. Therefore, natural goodness-of-fit tests consist in comparing a "distance" between C_n and an estimation C_{θ_n} of C obtained under H_0. Here and in the sequel, $\theta_n = \mathcal{T}_n\left(\mathbf{U}_1, \ldots, \mathbf{U}_n\right)$ stands for an estimate of θ derived from the pseudo-observations.

The idea of basing goodness-of-fit tests on the process $\sqrt{n}(C_n - C_{\theta_n})$ is evoked in a special case by Fermanian (2005), who dismisses it as "unpractical." Its properties and implementation are considered in general by Genest & Rémillard (2005). The L_2 and L_∞ norms lead to the familiar Cramér–von Mises and Kolmogorov–Smirnov statistics, defined respectively by

$$S_n = n \int_{[0,1]^d} \left\{C_n(\mathbf{u}) - C_{\theta_n}(\mathbf{u})\right\}^2 dC_n(\mathbf{u}) = \sum_{i=1}^{n} \{C_n(\mathbf{U_i}) - C_{\theta_n}(\mathbf{U_i})\}^2, \tag{3.2}$$

and

$$T_n = \sup_{\mathbf{u} \in [0,1]^d} \sqrt{n} \left|C_n(\mathbf{u}) - C_{\theta_n}(\mathbf{u})\right|.$$

Large values of either one of these statistics lead to the rejection of H_0. Approximate P-values can be deduced from their limiting distributions, which depend on the asymptotic behavior of the empirical process $\sqrt{n}(C_n - C_{\theta_n})$. Genest & Rémillard (2005) establish the convergence of the latter under appropriate regularity conditions on the assumed parametric family $\mathcal{C}_0$ and the sequence (θ_n) of estimators. These authors also show that the tests based on S_n and T_n are consistent ; in other words, if $C \notin \mathcal{C}_0$, then as $n \to \infty$, the probability of rejecting H_0 tends to 1.

In practice, the limiting distributions of S_n and T_n depend on θ. As a result, approximate P-values for these tests can only be obtained via specially adapted Monte

Carlo methods. A specific parametric bootstrap procedure is described in Appendix A. Its validity is established by Genest & Rémillard (2005).

3.3.2 Two tests based on Kendall's transform

Another avenue successively explored by Genest & Rivest (1993), Wang & Wells (2000b) and Genest et al. (2006) consists in basing a test of H_0 on a probability integral transformation of the data. The specific mapping they consider is

$$\mathbf{X} \mapsto V = H(\mathbf{X}) = C(\mathcal{U}_1, \ldots, \mathcal{U}_d),$$

where the joint distribution of the $\mathcal{U}_i = F_i(X_i)$ is C. This may be called Kendall's transform, given that the expectation of V is an affine transformation of the multivariate version of Kendall's coefficient of concordance ; see Barbe et al. (1996) or Jouini & Clemen (1996).

Let K denote the (univariate) distribution function of V. Genest & Rivest (1993) show that K can be estimated nonparametrically by the empirical distribution function of a rescaled version of the pseudo-observations $V_1 = C_n(\mathbf{U}_1), \ldots, V_n = C_n(\mathbf{U}_n)$. Barbe et al. (1996) give weak regularity conditions under which they prove a central limit theorem for the slight variant

$$K_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1\left\{C_n(\mathbf{U_i}) \leq t\right\}, \quad t \in [0,1]. \tag{3.3}$$

In particular, the latter is a consistent estimator of the underlying distribution K.

Now under H_0, the vector $\mathbf{U}$ is distributed as C_θ for some $\theta \in \mathcal{O}$, and hence the Kendall transform $C_\theta(\mathbf{U})$ has distribution K_θ. Through a measure of distance between K_n and a parametric estimation K_{θ_n} of K, one can then test the hypothesis

$$H_0'' : K \in \mathcal{K}_0 = \{K_\theta : \theta \in \mathcal{O}\}.$$

Because $H_0 \subset H_0''$, of course, the non-rejection of H_0'' does not entail the acceptance of H_0. Consequently, tests based on the empirical process $\sqrt{n}(K_n - K_{\theta_n})$ are not generally consistent. Although they point out this limitation, Genest et al. (2006) investigate tests of H_0 based on this process. The idea had been put forward earlier (but not carried through) by Wang & Wells (2000b) in the case of bivariate Archimedean copulas, for which H_0'' and H_0 are equivalent.

The specific statistics considered by Genest et al. (2006) derive from the L_2 and L_∞ norms. These Cramér–von Mises and Kolmogorov–Smirnov statistics are given respectively by

$$\begin{aligned} S_n^{(K)} &= n \int_0^1 \{K_n(t) - K_{\theta_n}(t)\}^2 \, dK_{\theta_n}(t) \qquad (3.4) \\ &= \frac{n}{3} + n \sum_{j=1}^{n-1} K_n^2 \left(\frac{j}{n}\right) \left\{ K_{\theta_n}\left(\frac{j+1}{n}\right) - K_{\theta_n}\left(\frac{j}{n}\right) \right\} \\ &\quad - n \sum_{j=1}^{n-1} K_n \left(\frac{j}{n}\right) \left\{ K_{\theta_n}^2\left(\frac{j+1}{n}\right) - K_{\theta_n}^2\left(\frac{j}{n}\right) \right\} \end{aligned}$$

and

$$T_n^{(K)} = \sup_{t\in[0,1]} \sqrt{n}\, |K_n(t) - K_{\theta_n}(t)| = \sqrt{n} \max_{i=0,1;\, 0\le j\le n-1} \left\{ \left| K_n\left(\frac{j}{n}\right) - K\left(\theta_n, \frac{j+i}{n}\right) \right| \right\}.$$

Large values of either one of these statistics lead to the rejection of H_0''. Approximate P-values can be deduced from their limiting distributions, which depend on the asymptotic behavior of the empirical process $\sqrt{n}(K_n - K_{\theta_n})$. The convergence of the latter is established by Genest et al. (2006) under appropriate regularity conditions on the assumed parametric families $\mathcal{C}_0$, $\mathcal{K}_0$, and the sequence (θ_n) of estimators.

As the asymptotic distributions of $S_n^{(K)}$ and $T_n^{(K)}$ depend on the unknown value of θ, approximate P-values for these statistics must again be found via simulation. See Appendix B for a parametric bootstrap procedure, whose validity derives from Genest & Rémillard (2005).

3.3.3 A test based on Rosenblatt's transform

Another well known probability integral transformation on which goodness-of-fit tests could be based is due to Rosenblatt (1952). This mapping, which is commonly used for simulation, provides a simple way of decomposing a random vector with known distribution into mutually independent components that are uniformly distributed on the unit interval. Its standard definition is recalled below for convenience.

Definition *Rosenblatt's probability integral transform of a copula C is the mapping $\mathcal{R} : (0,1)^d \mapsto (0,1)^d$ which to every $\mathbf{u} = (u_1, \ldots, u_d) \in (0,1)^d$ assigns another vector $\mathcal{R}(\mathbf{u}) = (e_1, \ldots, e_d)$ with $e_1 = u_1$ and for each $i \in \{2, \ldots, d\}$,*

$$e_i = \frac{\partial^{i-1} C(u_1, \ldots, u_i, 1, \ldots, 1)}{\partial u_1 \cdots \partial u_{i-1}} \Bigg/ \frac{\partial^{i-1} C(u_1, \ldots, u_{i-1}, 1, \ldots, 1)}{\partial u_1 \cdots \partial u_{i-1}} . \qquad (3.5)$$

A critical property of Rosenblatt's transform is that $\mathbf{U}$ is distributed as C, denoted $\mathbf{U} \sim C$, if and only if the distribution of $\mathbf{E} = \mathcal{R}(\mathbf{U})$ is

$$C_\perp(e_1, \ldots, e_d) = e_1 \times \cdots \times e_d, \quad e_1, \ldots, e_d \in (0,1)$$

i.e., the d–variate independence copula. Thus $H_0 : \mathbf{U} \sim C \in \mathcal{C}_0$ is equivalent to $H_0^* : \mathcal{R}_\theta(\mathbf{U}) \sim C_\perp$ for some $\theta \in \mathcal{O}$.

To test this hypothesis, therefore, one can use the fact that under H_0, the pseudo-observations $\mathbf{E}_1 = \mathcal{R}_{\theta_n}(\mathbf{U}_1), \ldots, \mathbf{E}_n = \mathcal{R}_{\theta_n}(\mathbf{U}_n)$ can be interpreted as a sample from the independence copula $C_\perp$. Of course, these pseudos are *not* mutually independent and only *approximately* uniform on $(0,1)^d$. Any inference procedure involving these constructs should thus take these features into account. This point is raised though eventually ignored by Breymann et al. (2003), who propose a test based on a specific transformation of $\mathbf{E}_1, \ldots, \mathbf{E}_n$.

To describe the procedure of Breymann et al. (2003), let Φ denote the cumulative distribution function of a standard $\mathcal{N}(0,1)$ random variable and define

$$\chi_i = \sum_{j=1}^{d} \{\Phi^{-1}(E_{ij})\}^2, \quad i \in \{1, \ldots, n\}.$$

Exploiting the fact that $\mathbf{E}_1, \ldots, \mathbf{E}_n$ are "approximately" uniformly distributed over $(0,1)^d$, these authors argue that $\chi_1, \ldots, \chi_n$ can be interpreted as a sample from G, the distribution function of a chi-square random variable with d degrees of freedom. Now a natural estimate of G is the empirical distribution of the set $\chi_1, \ldots, \chi_n$, viz.

$$G_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1\left(\chi_i \leq t\right), \quad t \geq 0. \tag{3.6}$$

For this reason, Breymann et al. (2003) explicitly assume that the limiting behavior of the empirical process $\sqrt{n}\,(G_n - G)$ is not affected by estimation. They further suppose that the asymptotic distribution does not depend on the unknown value of θ, so that it could be represented as $\beta \circ G$, where β is the standard Brownian bridge.

Should these assumptions hold true, Breymann et al. (2003) argue that it would then be possible to test H_0 with the Anderson–Darling statistic

$$A_n = -n - \frac{1}{n} \sum_{i=1}^{n} (2i-1)[\log\{G(\chi_{(i)})\} + \log\{1 - G(\chi_{(n+1-i)})\}], \tag{3.7}$$

where $\chi_{(1)} \leq \cdots \leq \chi_{(n)}$ are the order statistics corresponding to $\chi_1, \ldots, \chi_n$. The P-value would be simply given by reference to the limiting distribution of the original Anderson–Darling statistic ; see e.g., Shorack & Wellner (1986).

However, the limiting distribution of A_n varies with the unknown value of θ ; see, e.g., Ghoudi & Rémillard (1998, 2004) for details. As confirmed by the simulations in Section 3.6, this invalidates the procedure proposed by Breymann et al. (2003) : it has essentially no power and does not even maintain its nominal level.

To fix this problem, Dobrić & Schmid (2007) explain how results of Genest & Rémillard (2005) can be exploited to compute reliable P-values for test statistics based on $\sqrt{n}(G_n - G)$. In their paper, the Anderson–Darling test statistic A_n is used, together with the parametric bootstrap procedure described in Appendix C.

3.4 New procedures based on Rosenblatt's transform

One avenue not covered by Breymann et al. (2003) or Dobrić & Schmid (2007) is to work directly with the process, using the full power of Rosenblatt's transform. The idea is not new, as it appeared in Klugman & Parsa (1999) for bivariate censored data. Those authors propose a Pearson chi-square statistic with $\mathbf{E}_1, \ldots, \mathbf{E}_n$. However, the calculation of the P-value that they describe is incorrect, because it assumes amiss that the limiting distribution is chi-square. The fact that the margins were estimated using parametric families is not taken into account in their work.

Under the null hypothesis H_0, the empirical distribution function

$$D_n(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^{n} 1\left(\mathbf{E}_i \leq \mathbf{u}\right), \quad \mathbf{u} \in (0,1)^d$$

associated with the pseudo-observations $\mathbf{E}_1, \ldots, \mathbf{E}_n$ should be "close" to $C_\perp$. Thus, any reasonable notion of distance between D_n and $C_\perp$ is a good candidate for testing goodness-of-fit. Here, two Cramér–von Mises statistics are considered, namely

$$S_n^{(C)} = n \int_{[0,1]^d} \left\{D_n(\mathbf{u}) - C_\perp(\mathbf{u})\right\}^2 dD_n(\mathbf{u}) = \sum_{i=1}^{n} \left\{D_n(\mathbf{E}_i) - C_\perp(\mathbf{E}_i)\right\}^2 \tag{3.8}$$

and

$$\begin{aligned} S_n^{(B)} &= n \int_{[0,1]^d} \{D_n(\mathbf{u}) - C_\perp(\mathbf{u})\}^2 \, d\mathbf{u} \qquad (3.9) \\ &= \frac{n}{3^d} - \frac{1}{2^{d-1}} \sum_{i=1}^n \prod_{k=1}^d \left(1 - E_{ik}^2\right) + \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \prod_{k=1}^d \left(1 - E_{ik} \vee E_{jk}\right), \end{aligned}$$

where $a \vee b = \max(a, b)$. These statistics only differ in their integration measure.

Using the tools described in the paper of Ghoudi & Rémillard (2004), one can easily determine the asymptotic null behavior of $\sqrt{n}(D_n - C_\perp)$ and, in turn, the convergence of $S_n^{(B)}$ and $S_n^{(C)}$. The limiting null distributions of these statistics are both unwieldy and, as in previous cases, functions of the unknown value of the dependence parameter θ. Nevertheless, goodness-of-fit testing is possible through the parametric bootstrap procedure described in Appendix D, whose validity stems from the work of Genest & Rémillard (2005).

3.5 Experimental design

A large-scale Monte Carlo experiment was conducted to assess the finite-sample properties of the proposed goodness-of-fit tests for various choices of dependence structures and degrees of association. Two characteristics of the tests were of interest : their ability to maintain their nominal level, arbitrarily fixed at 5% throughout the study, and their power under a variety of alternatives.

In order to curtail the computational effort, comparisons were limited to the bivariate case and to three degrees of dependence as measured by Kendall's tau, i.e.,

$$\tau = 0.25, \quad 0.50, \quad 0.75.$$

Seven one-parameter families of copulas spanning all possible degrees of positive dependence were also considered, both under the null hypothesis and under the alternative. They fall into three categories :

1. Three meta-elliptical copula families uniquely determined from the following classical bivariate distributions with correlation coefficient $\rho = \sin(\pi\tau/2)$:
   (a) the Gaussian distribution ;
   (b) the Student distribution with $\nu = 4$ degrees of freedom ;
   (c) the Pearson type II distribution with $\nu = 4$ degrees of freedom.

2. Three of the most common Archimedean copula models, namely
   (a) the Clayton family, also known in the survival analysis literature as the gamma frailty model (Clayton, 1978; Cook & Johnson, 1981) ;
   (b) the Frank family (Frank, 1979; Nelsen, 1986; Genest, 1987) ;
   (c) the Gumbel–Hougaard family originally considered by Gumbel (1960) in the context of extreme-value theory.
3. The Plackett family of copulas (Plackett, 1965; Mardia, 1970).

The class of meta-elliptical copulas was introduced by Fang et al. (2002, 2005) ; its properties were examined by Frahm et al. (2003) and Abdous et al. (2005). These dependence structures are popular in actuarial science and in finance, where data often (but not always) exhibit heavy-tail dependence ; see Malevergne & Sornette (2003), Cherubini et al. (2004), McNeil et al. (2005) and references therein.

The Archimedean models are also commonly used in practice, particularly in survival analysis, because of their interpretation as mixture models and the natural extension they provide for Cox's proportional hazards model ; see, e.g., Oakes (1989), Faraggi & Korn (1996) or Wang & Wells (2000b). Refer also to Frees & Valdez (1998) and Klugman & Parsa (1999) for actuarial applications.

Finally, the Plackett system of distributions, which is neither Archimedean nor meta-elliptical, has found applications in biostatistics because of its constant cross-ratio property ; see, e.g., Burzykowski et al. (2004). Among others, Dobrić & Schmid (2005) investigated the relevance of this specific copula model in a financial context.

For every possible choice of copula and fixed value of tau, 10,000 random samples of size $n = 50$ were generated. An equal number of samples of size $n = 150$ was also obtained. Each of these samples was then used to test the goodness-of-fit of the seven families of distributions. Each of the following eight tests was applied in turn :

1. The two tests derived by Genest & Rémillard (2005) from the empirical copula process, i.e., those based on the statistics S_n and T_n.
2. The two tests developed by Genest et al. (2006) using Kendall's transform, i.e., those involving statistics $S_n^{(K)}$ and $T_n^{(K)}$.
3. The test of Breymann et al. (2003) based on the statistic A_n and its corrected version developed by Dobrić & Schmid (2007), which both rely on Rosenblatt's transform.
4. The two new procedures suggested in Section 3.4, i.e., those based on the statistics $S_n^{(B)}$ and $S_n^{(C)}$.

In all cases, the number of (primary level) bootstrap samples was fixed at $N = 1000$. Whenever necessary, $m = 2500$ samples were drawn for the second-level bootstrap.

This occurred when a closed-form expression was unavailable for the copula C_θ or the associated Kendall distribution K_θ. Two of the meta-elliptical copula models fall into this category on both accounts ; for the normal and the Plackett distributions, only K_θ needed to be estimated via a two-level parametric bootstrap.

Finally, whenever the parameter of a copula model ought to be estimated, this was done by inversion of Kendall's tau. Given the sample version τ_n of τ, this involved solving for θ in the equation

$$4\int_{[0,1]^2} C_\theta(u_1,u_2)dC_\theta(u_1,u_2)-1=\tau_n.$$

In all families considered, the solution is unique. See Nelsen (1999) for appropriate formulas and Genest & Rémillard (2005) for arguments showing that this method meets all the conditions required for the validity of the parametric bootstrap.

To sum up, the simulations were run according to a balanced experimental design involving the following factors :

$\mathcal{C}_0$: hypothesized copula model under H_0 (7 choices) ;

$\mathcal{C}$: copula model from which the data were generated (7 choices) ;

τ : level of dependence in $\mathcal{C}$, as measured by Kendall's tau (3 choices) ;

n : size of each sample drawn from $\mathcal{C}$ (2 choices).

In each of these $7 \times 7 \times 3 \times 2 = 294$ cases, 10,000 repetitions were performed in order to estimate the level or power of each of the eight tests under consideration.

The simulations whose results are presented below required the nearly exclusive use of 140 CPUs over a one-month period in three locations. The GERAD (Montréal) lent 60 processors, each of which having a 64-bit, 2.2 GHz CPU. The other 80 machines are located in the Salle des marchés and in the Département de mathématiques et de statistique at Université Laval ; they are 32-bit, 1.4 GHz CPUs.

3.6 Results

Tables 3.1–3.6 report the level and power of the omnibus tests described in Sections 3.3 and 3.4. Each table corresponds to a specific combination of $\tau \in \{0.25, 0.50, 0.75\}$ and $n \in \{50, 150\}$. Each line of a table shows the percentage of rejection of $H_0 : C \in \mathcal{C}_0$ associated with the different tests, given a choice of $\mathcal{C}_0$ and a true underlying copula family $\mathcal{C}$.

As an example, Table 3.1 shows that when testing for the Frank copula from a random sample of size $n = 50$, there are approximately 16.3% of chances that the test based on the Cramér–von Mises statistic S_n will reject the null hypothesis when the data are from the Gumbel–Hougaard copula with $\tau = 0.25$.

Note that the results for the test of Breymann et al. (2003) are omitted from all tables, given that the percentage of rejection of H_0 observed in the simulations never exceeded 1.5%. This portrays vividly the difficulties associated with an improper identification of the limiting distribution of a test statistic.

Because of the sheer size of Tables 3.1–3.6, it is somewhat difficult to get a quick grasp of the relative performance of the tests in terms of level and power. To assist with the interpretation, various aspects of the question are examined and illustrated with the help of boxplots in each of the following subsections.

Tab. 3.1 – Percentage of rejection of H_0 by various tests for data sets of size $n = \mathbf{50}$ arising from different copula models with $\tau = \mathbf{0.25}$.

Copula	True	Test based on						
under H_0	copula	S_n	T_n	$S_n^{(K)}$	$T_n^{(K)}$	$S_n^{(B)}$	$S_n^{(C)}$	A_n
Clayton	Clayton	**5.1**	**5.2**	**3.0**	**3.2**	**4.3**	**4.8**	**5.1**
	Gumbel–Hougaard	47.0	26.3	9.2	5.0	37.9	29.6	9.4
	Frank	25.0	12.9	5.3	3.1	17.8	13.9	5.1
	Plackett	25.1	13.0	4.5	2.6	18.3	15.0	6.3
	Normal	23.3	12.5	3.7	2.2	18.2	12.8	5.7
	Student 4 dl	27.1	14.0	3.1	2.0	21.3	21.2	12.0
	Pearson 4 dl	23.0	12.4	3.8	2.3	17.3	10.9	5.5
Gumbel–Hougaard	Clayton	15.4	23.4	52.5	42.1	13.4	11.2	5.5
	Gumbel–Hougaard	**5.5**	**5.2**	**4.8**	**4.7**	**4.5**	**5.4**	**5.2**
	Frank	4.3	7.4	10.6	8.5	4.2	4.4	4.8
	Plackett	4.5	8.0	10.4	8.7	4.6	5.8	4.7
	Normal	3.8	6.6	13.2	10.8	3.8	3.7	4.6
	Student 4 dl	5.4	7.9	16.2	14.9	5.1	8.3	8.5
	Pearson 4 dl	3.6	6.6	12.0	9.9	3.8	3.1	6.6
Frank	Clayton	7.9	14.8	35.4	32.9	7.3	7.6	6.1
	Gumbel–Hougaard	16.3	10.2	3.3	3.2	13.0	13.0	6.9
	Frank	**5.0**	**4.9**	**4.2**	**4.6**	**4.3**	**4.5**	**5.0**
	Plackett	5.6	5.8	4.5	4.7	5.0	6.3	5.7
	Normal	5.9	6.0	6.5	6.9	4.9	5.5	5.1
	Student 4 dl	9.0	8.0	9.8	10.7	7.2	11.1	14.0
	Pearson 4 dl	5.2	5.6	5.3	5.6	4.8	4.3	5.1
Plackett	Clayton	7.1	14.0	32.7	29.1	6.7	6.5	5.0
	Gumbel–Hougaard	15.0	9.4	2.9	2.9	11.7	11.0	5.4
	Frank	4.8	4.9	4.1	4.2	4.1	3.7	5.1
	Plackett	**5.3**	**5.5**	**4.2**	**4.3**	**4.6**	**5.1**	**5.1**
	Normal	5.4	5.5	5.9	6.3	4.7	4.7	4.7
	Student 4 dl	7.9	7.2	8.4	8.8	6.1	8.7	10.4
	Pearson 4 dl	5.0	5.5	5.1	5.1	4.7	3.9	6.5
Normal	Clayton	6.7	12.7	26.3	21.5	8.1	8.3	5.3
	Gumbel–Hougaard	14.2	8.3	2.2	2.1	12.5	13.3	5.9
	Frank	5.6	5.1	3.4	3.6	5.1	5.6	5.0
	Plackett	6.2	5.8	3.4	3.1	6.0	7.6	5.5
	Normal	**5.2**	**4.9**	**3.9**	**4.6**	**4.7**	**5.3**	**4.9**
	Student 4 dl	7.8	6.2	5.2	5.4	8.0	13.1	11.5
	Pearson 4 dl	4.9	5.2	3.9	4.0	4.5	4.1	5.8
Student 4 dl	Clayton	5.0	10.9	23.5	18.2	5.6	3.8	6.5
	Gumbel–Hougaard	11.4	7.1	2.0	2.0	8.5	5.6	5.8
	Frank	4.8	5.3	3.3	3.4	4.0	2.4	9.0
	Plackett	5.0	5.5	3.3	3.3	4.3	3.5	7.0
	Normal	4.1	4.5	3.6	4.0	4.1	2.7	8.9
	Student 4 dl	**5.0**	**5.0**	**4.1**	**4.0**	**4.4**	**4.6**	**4.6**
	Pearson 4 dl	4.2	5.3	3.7	3.6	4.6	2.2	14.2
Pearson 4 dl	Clayton	7.9	13.8	27.0	23.3	9.3	11.9	8.2
	Gumbel–Hougaard	16.2	9.1	2.6	2.4	15.0	18.6	8.8
	Frank	6.4	5.5	3.4	3.8	6.1	8.4	5.9
	Plackett	7.3	6.5	3.6	3.6	7.7	11.1	7.5
	Normal	6.0	5.4	4.6	4.9	5.5	7.7	6.0
	Student 4 dl	9.2	7.1	6.3	6.6	10.3	19.3	18.0
	Pearson 4 dl	**5.7**	**5.3**	**4.1**	**4.1**	**5.0**	**5.6**	**5.0**

Tab. 3.2 – Percentage of rejection of H_0 by various tests for data sets of size $n = \mathbf{50}$ arising from different copula models with $\tau = \mathbf{0.50}$.

Copula	True	Test based on						
under H_0	copula	S_n	T_n	$S_n^{(K)}$	$T_n^{(K)}$	$S_n^{(B)}$	$S_n^{(C)}$	A_n
Clayton	Clayton	**5.6**	**5.1**	**4.3**	**4.5**	**4.8**	**4.8**	**4.6**
	Gumbel–Hougaard	86.0	63.9	42.8	24.3	71.1	59.2	15.4
	Frank	57.3	34.1	29.4	17.6	45.6	30.6	9.3
	Plackett	59.0	33.2	22.9	13.0	44.1	35.3	14.7
	Normal	54.1	32.3	18.3	10.1	37.6	25.8	7.0
	Student 4 dl	58.7	36.1	17.7	9.7	38.4	34.2	12.2
	Pearson 4 dl	53.3	32.2	19.1	10.6	39.5	24.9	6.0
Gumbel–Hougaard	Clayton	49.2	59.9	83.9	66.7	63.9	49.5	5.4
	Gumbel–Hougaard	**4.9**	**4.9**	**4.2**	**4.5**	**4.5**	**5.0**	**4.3**
	Frank	7.8	13.5	16.1	11.9	16.2	9.6	5.0
	Plackett	6.6	11.8	14.1	11.4	14.7	13.2	7.2
	Normal	4.1	10.5	16.9	13.6	10.7	6.3	3.6
	Student 4 dl	5.3	10.4	17.9	15.2	10.7	12.1	5.3
	Pearson 4 dl	4.1	10.7	16.9	13.8	12.1	6.3	3.5
Frank	Clayton	20.8	34.7	58.9	49.0	29.4	29.1	4.8
	Gumbel–Hougaard	26.6	14.4	6.4	4.3	14.4	19.6	4.9
	Frank	**4.6**	**4.7**	**4.1**	**4.2**	**4.9**	**4.9**	**4.8**
	Plackett	6.1	5.3	4.9	4.6	6.0	11.0	6.0
	Normal	7.5	7.2	8.1	7.1	4.6	6.2	3.6
	Student 4 dl	11.8	8.9	11.1	9.3	6.3	15.4	6.8
	Pearson 4 dl	6.5	7.4	7.7	6.6	4.1	4.6	4.3
Plackett	Clayton	17.2	31.3	53.8	40.3	26.5	19.0	5.5
	Gumbel–Hougaard	21.5	11.7	4.6	3.6	11.1	9.4	5.1
	Frank	4.7	5.5	4.9	5.4	5.0	2.8	5.7
	Plackett	**4.9**	**5.2**	**4.4**	**4.5**	**5.1**	**5.2**	**4.5**
	Normal	5.3	6.6	5.7	5.4	4.0	2.6	6.0
	Student 4 dl	7.3	6.6	7.2	6.1	4.2	6.0	5.5
	Pearson 4 dl	5.1	6.8	5.7	5.8	4.2	2.3	8.0
Normal	Clayton	18.3	28.0	47.5	33.5	37.2	30.7	4.6
	Gumbel–Hougaard	20.2	9.2	4.0	3.2	16.1	16.4	4.8
	Frank	7.9	6.7	5.9	6.2	7.5	6.2	5.7
	Plackett	7.6	5.8	4.1	4.5	9.6	13.8	7.6
	Normal	**4.9**	**5.0**	**4.0**	**4.5**	**4.9**	**5.0**	**4.4**
	Student 4 dl	6.7	4.8	4.5	4.1	8.6	15.5	6.2
	Pearson 4 dl	4.3	4.8	4.2	4.7	4.5	3.2	4.7
Student 4 dl	Clayton	15.9	28.8	48.3	33.4	30.5	17.4	4.0
	Gumbel–Hougaard	16.9	9.3	4.3	3.5	11.3	6.5	3.9
	Frank	8.5	8.8	8.3	8.4	7.0	3.0	4.7
	Plackett	6.7	6.3	5.1	5.5	6.3	5.5	5.3
	Normal	4.2	5.3	4.2	5.0	4.7	1.7	5.0
	Student 4 dl	**4.6**	**4.6**	**4.1**	**4.5**	**4.7**	**4.8**	**4.6**
	Pearson 4 dl	4.0	5.5	4.8	5.4	5.2	1.3	5.4
Pearson 4 dl	Clayton	20.2	28.2	47.5	33.7	39.4	37.1	5.0
	Gumbel–Hougaard	22.0	9.8	4.1	3.2	18.6	23.0	5.3
	Frank	8.2	6.7	5.6	6.1	8.3	8.6	6.5
	Plackett	8.1	5.6	3.9	4.6	12.1	19.0	9.1
	Normal	5.1	5.1	4.0	4.4	5.5	7.7	4.7
	Student 4 dl	7.8	5.7	4.7	4.4	11.4	22.5	7.5
	Pearson 4 dl	**4.9**	**4.7**	**4.5**	**4.8**	**4.8**	**5.0**	**4.5**

Tab. 3.3 – Percentage of rejection of H_0 by various tests for data sets of size $n = \mathbf{50}$ arising from different copula models with $\tau = \mathbf{0.75}$.

Copula	True	Test based on						
under H_0	copula	S_n	T_n	$S_n^{(K)}$	$T_n^{(K)}$	$S_n^{(B)}$	$S_n^{(C)}$	A_n
Clayton	Clayton	**5.1**	**4.6**	**3.6**	**4.1**	**5.1**	**4.9**	**4.9**
	Gumbel–Hougaard	96.0	74.2	66.1	44.0	78.7	64.8	3.5
	Frank	69.6	36.8	44.2	29.0	60.6	37.1	8.9
	Plackett	75.2	40.7	34.3	22.4	51.6	38.1	10.3
	Normal	76.3	45.6	33.0	22.4	49.1	31.2	2.7
	Student 4 dl	78.1	48.2	32.6	21.6	42.5	30.8	3.4
	Pearson 4 dl	75.7	43.7	33.2	21.8	51.4	32.3	2.7
Gumbel–Hougaard	Clayton	50.8	61.2	86.2	51.8	94.7	84.4	2.9
	Gumbel–Hougaard	**4.1**	**4.6**	**3.7**	**4.2**	**4.9**	**5.0**	**4.6**
	Frank	9.9	15.6	18.1	14.3	29.7	14.8	16.6
	Plackett	5.5	8.5	11.1	7.8	25.6	21.5	9.4
	Normal	2.4	8.4	10.4	8.3	19.7	10.4	3.7
	Student 4 dl	3.3	7.9	11.4	7.8	18.4	13.9	3.8
	Pearson 4 dl	2.6	8.4	10.8	9.0	20.8	9.9	4.2
Frank	Clayton	16.4	25.6	50.9	15.5	64.9	50.1	9.2
	Gumbel–Hougaard	28.0	8.6	10.0	4.6	12.8	15.7	15.2
	Frank	**3.9**	**3.9**	**3.6**	**2.4**	**5.1**	**4.8**	**4.7**
	Plackett	7.5	3.4	5.0	2.9	8.6	16.8	6.3
	Normal	9.1	6.2	8.3	4.0	4.6	5.2	19.7
	Student 4 dl	13.0	6.0	10.8	4.5	6.1	12.1	17.5
	Pearson 4 dl	8.4	6.9	8.2	3.9	4.3	4.1	21.3
Plackett	Clayton	12.6	33.4	50.5	23.0	62.6	38.1	6.8
	Gumbel–Hougaard	15.4	9.5	5.7	4.6	8.5	4.5	12.6
	Frank	5.2	8.3	6.9	8.5	4.7	1.8	5.4
	Plackett	**3.8**	**4.3**	**3.8**	**3.9**	**5.1**	**5.4**	**5.0**
	Normal	3.5	6.5	3.9	3.7	2.1	0.9	15.7
	Student 4 dl	4.3	5.5	4.5	3.5	2.7	2.0	14.8
	Pearson 4 dl	3.2	7.0	4.0	3.8	2.2	0.7	15.3
Normal	Clayton	20.2	32.7	51.6	23.4	77.0	62.4	3.5
	Gumbel–Hougaard	17.7	7.2	5.4	4.8	16.7	15.5	4.0
	Frank	12.9	9.6	11.5	11.1	10.3	7.7	17.6
	Plackett	8.1	4.2	4.6	4.5	13.4	19.2	9.4
	Normal	**4.1**	**4.5**	**3.2**	**3.9**	**4.9**	**4.6**	**4.8**
	Student 4 dl	5.0	3.6	3.3	3.3	7.2	11.3	4.6
	Pearson 4 dl	3.8	4.5	3.6	4.1	4.8	4.0	5.0
Student 4 dl	Clayton	18.5	36.0	53.0	26.4	72.8	50.8	4.0
	Gumbel–Hougaard	15.1	8.2	5.5	5.5	13.6	8.2	4.0
	Frank	13.4	12.2	14.5	13.3	9.3	4.4	21.5
	Plackett	7.4	5.2	5.6	5.7	9.8	11.0	11.3
	Normal	3.3	5.2	3.4	4.7	3.9	2.0	4.9
	Student 4 dl	**3.7**	**4.3**	**3.4**	**3.9**	**4.7**	**4.8**	**4.5**
	Pearson 4 dl	3.4	5.4	3.9	4.9	4.0	1.7	5.1
Pearson 4 dl	Clayton	21.2	31.2	51.2	22.5	77.2	66.0	3.5
	Gumbel–Hougaard	19.2	6.8	5.4	4.6	17.6	18.8	3.8
	Frank	13.1	8.7	10.8	10.4	10.4	9.2	16.1
	Plackett	8.8	4.1	4.7	4.0	14.7	22.9	8.8
	Normal	4.3	4.3	3.4	4.0	5.3	6.3	4.7
	Student 4 dl	5.4	3.6	3.6	3.2	8.4	14.8	4.6
	Pearson 4 dl	**4.4**	**4.4**	**3.5**	**4.0**	**5.0**	**4.9**	**4.8**

Tab. 3.4 – Percentage of rejection of H_0 by various tests for data sets of size $n = \mathbf{150}$ arising from different copula models with $\tau = \mathbf{0.25}$.

Copula	True	Test based on						
under H_0	copula	S_n	T_n	$S_n^{(K)}$	$T_n^{(K)}$	$S_n^{(B)}$	$S_n^{(C)}$	A_n
Clayton	Clayton	**4.6**	**4.8**	**4.1**	**4.5**	**4.7**	**4.8**	**4.9**
	Gumbel–Hougaard	86.1	62.4	57.9	42.7	80.9	76.7	22.4
	Frank	56.3	32.7	37.4	26.4	42.8	36.2	6.2
	Plackett	56.0	31.2	33.7	23.4	43.9	39.0	8.1
	Normal	50.2	27.5	24.5	16.8	41.8	34.6	6.4
	Student 4 dl	56.5	32.3	23.2	15.5	51.0	52.7	32.9
	Pearson 4 dl	49.9	28.7	26.1	17.3	43.3	32.9	6.4
Gumbel–Hougaard	Clayton	72.1	62.6	92.3	82.1	65.1	60.5	8.3
	Gumbel–Hougaard	**5.0**	**5.0**	**4.7**	**5.1**	**5.1**	**5.0**	**5.0**
	Frank	15.4	15.4	19.9	15.1	12.9	10.0	5.9
	Plackett	14.3	14.7	18.9	14.7	12.5	10.6	4.9
	Normal	10.1	11.7	24.4	18.9	10.2	7.5	5.9
	Student 4 dl	14.1	12.9	29.8	26.2	14.3	18.2	17.4
	Pearson 4 dl	10.2	12.6	23.6	18.5	12.8	7.8	11.3
Frank	Clayton	40.0	36.8	77.3	70.6	36.2	36.1	9.6
	Gumbel–Hougaard	33.4	18.5	9.1	6.1	27.8	29.5	12.4
	Frank	**5.3**	**5.1**	**5.1**	**5.0**	**4.9**	**4.9**	**5.1**
	Plackett	5.7	5.2	5.4	5.1	5.2	6.1	6.6
	Normal	7.8	7.3	10.5	9.9	6.2	6.3	5.3
	Student 4 dl	18.5	11.4	22.0	19.7	14.6	23.0	40.7
	Pearson 4 dl	6.5	7.3	7.7	7.6	6.5	5.2	7.0
Plackett	Clayton	37.6	34.2	69.8	60.5	33.3	31.9	6.2
	Gumbel–Hougaard	30.4	16.6	7.2	5.4	24.6	24.8	6.8
	Frank	5.0	5.2	4.8	5.1	5.0	4.2	6.5
	Plackett	**5.2**	**5.0**	**4.8**	**4.8**	**4.5**	**4.7**	**5.0**
	Normal	6.8	6.8	8.2	7.6	6.1	5.4	5.7
	Student 4 dl	14.1	9.8	15.6	14.4	10.1	15.6	26.2
	Pearson 4 dl	6.2	7.3	6.6	6.4	7.5	5.4	12.0
Normal	Clayton	31.6	26.6	56.9	45.8	33.3	33.0	7.2
	Gumbel–Hougaard	23.8	11.9	7.1	5.5	24.7	27.0	8.9
	Frank	7.9	7.2	5.6	5.3	7.2	7.0	5.5
	Plackett	7.9	6.8	4.4	4.4	8.2	9.4	6.0
	Normal	**5.1**	**5.0**	**4.7**	**5.2**	**4.7**	**5.0**	**4.8**
	Student 4 dl	10.5	6.8	7.4	7.4	16.6	27.8	29.9
	Pearson 4 dl	4.8	5.3	4.9	4.7	4.7	3.4	8.2
Student 4 dl	Clayton	27.7	26.2	52.1	39.0	25.1	17.4	11.2
	Gumbel–Hougaard	19.1	11.4	7.4	6.0	17.3	11.5	9.5
	Frank	9.1	8.2	9.5	7.6	8.9	4.5	23.3
	Plackett	7.7	7.7	7.3	6.2	6.6	3.6	13.9
	Normal	4.9	5.9	5.4	5.0	7.9	3.1	23.0
	Student 4 dl	**4.8**	**5.3**	**4.6**	**4.7**	**4.5**	**4.8**	**5.4**
	Pearson 4 dl	6.2	7.1	6.5	5.9	15.7	5.9	42.9
Pearson 4 dl	Clayton	35.7	29.3	60.0	50.9	40.5	44.8	16.9
	Gumbel–Hougaard	28.2	13.7	7.8	5.7	32.1	40.1	21.2
	Frank	9.0	7.0	4.9	4.7	10.4	12.4	7.6
	Plackett	9.2	7.2	4.2	4.1	13.2	17.8	12.4
	Normal	6.2	5.2	5.5	5.5	6.5	9.0	8.1
	Student 4 dl	16.4	8.6	10.2	9.3	28.0	47.3	52.6
	Pearson 4 dl	**5.2**	**5.1**	**5.0**	**4.8**	**4.9**	**4.7**	**5.3**

Tab. 3.5 – Percentage of rejection of H_0 by various tests for data sets of size $n = \mathbf{150}$ arising from different copula models with $\tau = \mathbf{0.50}$.

Copula	True	Test based on						
under H_0	copula	S_n	T_n	$S_n^{(K)}$	$T_n^{(K)}$	$S_n^{(B)}$	$S_n^{(C)}$	A_n
Clayton	Clayton	**5.3**	**5.0**	**4.5**	**4.5**	**5.1**	**5.0**	**5.0**
	Gumbel–Hougaard	99.9	98.3	98.5	91.4	99.7	99.5	78.3
	Frank	95.7	81.2	89.5	74.9	94.4	90.3	37.2
	Plackett	95.8	77.7	83.5	63.5	92.9	90.4	62.0
	Normal	93.7	74.1	75.1	53.7	89.0	85.5	35.2
	Student 4 dl	94.8	78.0	75.0	54.4	87.9	87.6	50.4
	Pearson 4 dl	94.0	74.3	75.8	55.0	91.9	88.0	31.9
Gumbel–Hougaard	Clayton	99.6	98.4	99.9	99.0	99.7	99.5	33.4
	Gumbel–Hougaard	**4.6**	**5.0**	**4.6**	**4.9**	**4.5**	**4.9**	**5.0**
	Frank	39.8	37.5	42.4	28.4	52.1	37.0	9.3
	Plackett	29.8	27.2	32.0	23.1	43.2	37.0	21.6
	Normal	18.3	21.1	37.7	27.4	33.7	25.2	4.9
	Student 4 dl	21.8	21.1	40.6	31.7	29.7	31.9	10.0
	Pearson 4 dl	18.1	21.7	36.6	26.2	41.2	28.9	4.0
Frank	Clayton	89.1	84.9	98.6	96.3	86.9	90.4	13.3
	Gumbel–Hougaard	63.0	39.6	28.3	15.8	44.1	57.6	9.2
	Frank	**4.8**	**5.1**	**4.8**	**5.2**	**4.8**	**4.8**	**5.1**
	Plackett	8.4	6.3	7.5	6.8	10.5	19.9	12.5
	Normal	19.9	15.0	22.6	17.3	8.9	14.4	4.8
	Student 4 dl	35.1	19.6	37.2	27.2	22.9	44.3	19.1
	Pearson 4 dl	15.0	13.0	17.3	13.1	7.1	8.9	5.8
Plackett	Clayton	83.9	78.4	95.5	86.4	79.6	78.0	12.5
	Gumbel–Hougaard	48.8	28.1	16.4	10.1	29.1	30.4	8.1
	Frank	6.8	7.8	8.2	8.0	10.2	3.9	10.5
	Plackett	**5.0**	**5.3**	**5.0**	**5.1**	**4.9**	**5.2**	**4.7**
	Normal	9.8	11.2	9.4	7.9	6.9	5.1	12.3
	Student 4 dl	15.1	11.4	15.1	10.6	7.4	11.7	7.4
	Pearson 4 dl	8.2	11.0	7.7	6.5	9.4	5.5	21.3
Normal	Clayton	80.0	68.8	90.3	75.2	90.8	88.2	7.8
	Gumbel–Hougaard	38.3	17.8	16.1	10.8	42.0	44.4	5.7
	Frank	20.2	14.3	17.4	14.1	13.4	8.5	8.7
	Plackett	13.2	9.7	6.8	6.6	18.0	22.7	18.1
	Normal	**4.9**	**5.0**	**4.9**	**5.2**	**5.0**	**5.3**	**4.8**
	Student 4 dl	8.2	5.3	5.9	5.2	20.4	32.1	8.8
	Pearson 4 dl	4.6	4.9	5.0	5.0	4.8	3.0	5.5
Student 4 dl	Clayton	77.3	70.5	90.6	73.2	84.9	74.9	6.0
	Gumbel–Hougaard	33.9	18.2	17.3	11.8	30.3	20.9	4.9
	Frank	26.9	18.9	29.3	20.7	24.2	8.1	6.0
	Plackett	13.8	11.0	11.6	9.5	10.2	6.9	10.4
	Normal	5.2	6.4	5.9	6.1	9.9	2.9	6.7
	Student 4 dl	**5.0**	**4.9**	**4.9**	**5.0**	**5.1**	**5.2**	**4.9**
	Pearson 4 dl	5.6	6.8	6.8	6.7	18.3	5.2	9.4
Pearson 4 dl	Clayton	81.8	69.6	91.2	76.3	92.9	92.8	11.2
	Gumbel–Hougaard	41.9	19.3	16.2	10.7	51.7	59.8	8.2
	Frank	18.9	13.5	13.9	11.7	14.5	12.8	11.4
	Plackett	13.9	9.7	5.6	5.7	28.8	37.5	23.7
	Normal	5.5	5.0	5.0	4.8	7.4	11.0	5.0
	Student 4 dl	10.9	6.2	6.7	5.7	34.3	52.4	14.1
	Pearson 4 dl	**4.7**	**4.7**	**4.7**	**4.8**	**4.7**	**4.9**	**4.8**

Tab. 3.6 – Percentage of rejection of H_0 by various tests for data sets of size $n = \mathbf{150}$ arising from different copula models with $\tau = \mathbf{0.75}$.

Copula under H_0	True copula	Test based on S_n	T_n	$S_n^{(K)}$	$T_n^{(K)}$	$S_n^{(B)}$	$S_n^{(C)}$	A_n
Clayton	Clayton	**5.4**	**5.0**	**4.9**	**5.1**	**5.1**	**5.2**	**5.0**
	Gumbel–Hougaard	99.9	99.9	99.9	98.7	99.9	99.9	49.1
	Frank	99.1	86.2	97.0	81.2	99.9	99.7	76.7
	Plackett	99.5	89.1	93.6	73.6	99.6	99.5	64.1
	Normal	99.8	91.7	94.9	77.7	99.5	99.6	23.8
	Student 4 dl	99.8	95.1	94.3	79.4	99.0	99.1	18.2
	Pearson 4 dl	99.7	90.7	95.1	77.1	99.7	99.7	29.8
Gumbel–Hougaard	Clayton	99.9	99.5	99.9	99.2	99.9	99.9	29.0
	Gumbel–Hougaard	**4.5**	**4.7**	**4.4**	**4.6**	**5.2**	**4.8**	**4.9**
	Frank	51.7	45.4	61.6	38.0	83.8	72.4	75.0
	Plackett	25.8	20.3	29.8	17.9	67.8	62.8	39.6
	Normal	12.3	17.0	29.4	18.6	60.7	53.6	5.9
	Student 4 dl	16.1	17.4	32.9	19.8	54.8	52.0	3.9
	Pearson 4 dl	11.8	18.6	30.1	19.6	66.9	58.7	6.5
Frank	Clayton	96.6	91.7	99.6	95.5	99.7	99.7	26.8
	Gumbel–Hougaard	81.9	43.6	53.2	27.1	59.9	74.2	40.0
	Frank	**4.7**	**4.7**	**4.5**	**4.7**	**5.0**	**5.1**	**5.2**
	Plackett	20.6	8.0	15.4	8.8	18.6	36.0	7.9
	Normal	40.9	21.2	40.2	20.5	18.4	30.1	49.8
	Student 4 dl	59.4	26.0	56.0	27.9	34.4	58.2	42.3
	Pearson 4 dl	34.2	21.0	34.5	18.0	15.0	22.3	54.1
Plackett	Clayton	89.8	86.8	97.7	78.6	99.5	99.1	18.8
	Gumbel–Hougaard	45.8	23.4	19.1	11.4	35.5	29.4	37.4
	Frank	14.9	15.4	18.5	15.3	9.7	3.6	10.9
	Plackett	**4.7**	**5.0**	**4.9**	**5.1**	**4.9**	**5.2**	**5.2**
	Normal	7.7	12.9	7.7	6.0	2.5	1.2	44.3
	Student 4 dl	11.0	12.3	11.4	6.7	4.3	3.6	45.2
	Pearson 4 dl	7.4	13.8	6.6	5.5	2.9	1.5	44.2
Normal	Clayton	91.8	82.4	97.3	75.4	99.9	99.9	8.2
	Gumbel–Hougaard	38.5	13.2	17.9	10.6	55.5	54.0	4.7
	Frank	42.2	22.9	41.4	24.6	32.8	20.1	70.2
	Plackett	16.5	7.6	7.0	7.0	23.0	30.6	30.0
	Normal	**4.9**	**4.4**	**4.4**	**4.8**	**4.9**	**4.6**	**5.1**
	Student 4 dl	6.6	4.3	4.9	4.5	12.3	18.3	4.9
	Pearson 4 dl	4.4	5.3	4.6	4.8	4.8	3.7	5.1
Student 4 dl	Clayton	90.6	86.6	97.7	78.6	99.9	99.7	10.9
	Gumbel–Hougaard	33.9	15.1	19.2	11.5	48.4	39.3	4.6
	Frank	48.2	30.5	53.9	32.4	39.3	20.3	81.8
	Plackett	15.7	8.9	11.0	9.7	16.4	17.2	43.5
	Normal	4.1	5.7	5.1	6.0	5.0	2.1	5.9
	Student 4 dl	**4.9**	**4.7**	**4.8**	**4.9**	**5.6**	**5.3**	**4.5**
	Pearson 4 dl	4.3	6.4	5.4	6.0	6.2	2.7	7.1
Pearson 4 dl	Clayton	93.0	81.4	97.4	75.1	99.9	99.9	7.3
	Gumbel–Hougaard	42.0	13.3	18.4	10.5	58.3	60.3	4.5
	Frank	41.2	21.0	37.4	22.8	28.4	20.2	63.4
	Plackett	17.5	7.3	6.5	6.2	26.7	37.3	25.8
	Normal	5.3	4.3	5.0	4.8	5.6	7.3	5.1
	Student 4 dl	8.3	4.3	5.4	4.3	18.3	28.9	5.2
	Pearson 4 dl	**4.5**	**4.8**	**4.6**	**4.7**	**4.8**	**5.0**	**4.7**

3.6.1 Level of the tests

Given that their finite-sample distribution is approximated by a parametric bootstrap procedure, the tests based on statistics

$$\begin{array}{llll} 1 : S_n, & 3 : S_n^{(K)}, & 5 : S_n^{(B)} & 7 : A_n \\ 2 : T_n, & 4 : T_n^{(K)}, & 6 : S_n^{(C)} & \end{array}$$

are expected to hold their nominal level. A cursory look at the figures highlighted in Tables 3.1–3.6 confirms that this happens in the vast majority of cases.

The same information is rendered graphically in Fig. 3.1, where two boxplots show the dispersion in the levels observed across the seven tests and the $21 = 7 \times 3$ combinations of null hypothesis $\mathcal{C}_0$ and level of dependence τ. The data for $n = 50$ and $n = 150$ are represented in the top and bottom panel, respectively.

It is obvious from the graphs in Fig. 3.1 that overall, the parametric bootstrap algorithm does a very good job of approximating the null distribution of all statistics. Except in a few cases, the performance is quite acceptable when $n = 50$. It is almost irreproachable when $n = 150$.

3.6.2 Effect of sample size

It is a stylized fact of statistics that the power of a test increases with sample size. As Fig. 3.2 clearly shows, the present case is no exception. The boxplot displayed there portrays the variation in the ratio power$(n = 150)$/power$(n = 50)$ for each of the seven

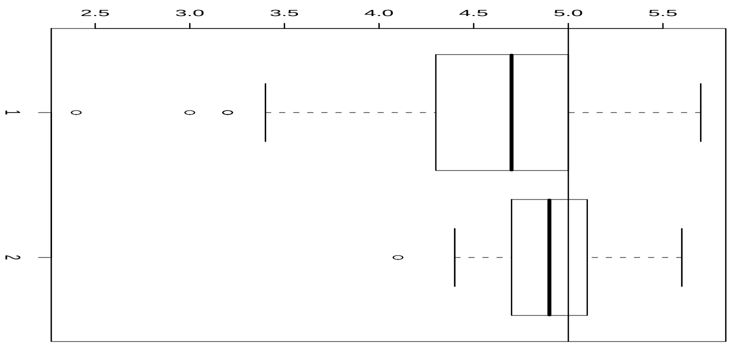


FIG. 3.1 – Level of seven goodness-of-fit tests, as observed across $21 = 7 \times 3$ choices of $\mathcal{C}_0$ and τ. Top panel : $n = 50$; bottom panel : $n = 150$.

tests, as observed across $126 = 7 \times 6 \times 3$ combinations of factors $\mathcal{C}_0$, $\mathcal{C}$ and τ, when the first two factors are different.

One can readily see from Fig. 3.2 that on average, the tests double their power as sample size goes from $n = 50$ to 150. In many instances, the improvement is more than four-fold, but a few cases can also be found where no gain in power occurs. It is instructive to examine more carefully what happens in those extreme cases.

1. What are the outliers identified in Fig. 3.2 and why is the increase in power so large in those cases ?

    (a) Most outliers occur either at $\tau = 0.25$ or 0.75.

    (b) The statistics S_n, T_n and $S_n^{(B)}$ have very few outliers, if any.

    (c) The outliers at $\tau = 0.25$ are for $S_n^{(K)}$ and $T_n^{(K)}$, which prove particularly apt at detecting that data are not of the Clayton type as n increases.

    (d) Most of the outliers at $\tau = 0.75$ are for the tests $T_n^{(K)}$ and A_n ; when $n = 150$, the first becomes much better at assessing the goodness-of-fit of a Frank copula, while the second can discriminate a Clayton dependence structure far more easily.

(2) In what cases does one observe an increase in power of 10% or less (as identified by the vertical line crossing the boxplots), and why ?

    (a) This phenomenon occurs mostly when $\tau = 0.25$ or 0.75, and twice as often in the former case than in the latter.

    (b) This problem spares $S_n^{(K)}$ and $T_n^{(K)}$ and affects all others equally.

    (c) In half of the cases, the problem occurs because of a failure to distinguish between the normal and the Pearson copulas ; most of the other instances of low increase in power occur when the null and the alternative are the Frank and Plackett copulas, or vice versa.

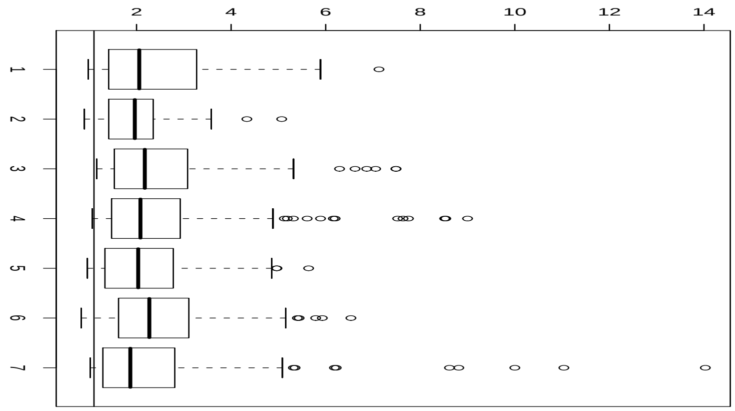


FIG. 3.2 – Ratio power($n = 150$)/power($n = 50$) for seven goodness-of-fit tests, as observed across $126 = 7 \times 6 \times 3$ combinations of factors $\mathcal{C}_0$, $\mathcal{C}$ and τ for which $\mathcal{C}_0 \neq \mathcal{C}$. The vertical line is at 1.1, to help identify the cases where the improvement in power is less than 10% when n goes from 50 to 150.

To illustrate the difficulties associated with the proper identification of a dependence structure from as small a sample as $n = 50$, Fig. 3.3 portrays typical scatter plots for the seven copula models considered in the study. For comparative purposes, $\tau = 0.5$ in all cases. The distinctive features of the models are hardly distinguishable and would be even fuzzier if one were to set $\tau = 0.25$.

In contrast, Fig. 3.4 displays scatter plots of random samples of size $n = 1000$ from the same copulas, again with $\tau = 0.5$. The dominating characteristics of the different models are then somewhat easier to pick up. The Clayton and the Gumbel–Hougaard, in particular, are fairly easy to spot, given their tail behavior. Their lower- and upper-tail dependence translate into greater densities of points in the lower-left and upper-right corners of the unit square, respectively.

While a trained eye could perhaps distinguish consistently between other pairs of copulas at $n = 1000$, some differences remain tenuous, e.g., between the Frank and the Plackett, or between the normal and the Pearson copulas. Thus the fact that many of the power figures in Tables 3.1–3.6 are low does not come as a total surprise.

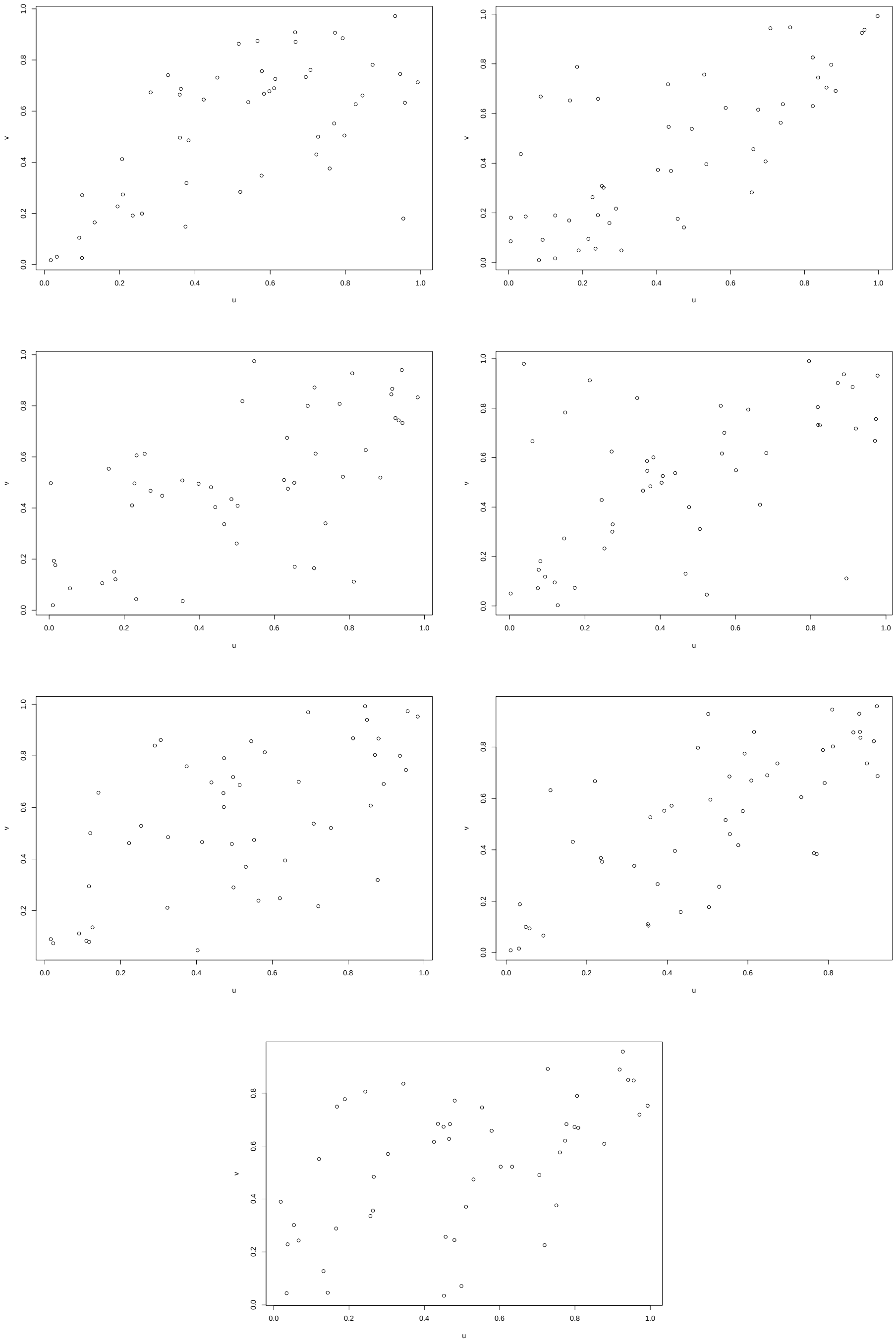


FIG. 3.3 – Samples of size $n = \mathbf{50}$ from seven different copulas with parameter $\tau = 0.50$. From left to right, and top to bottom : Clayton, Gumbel–Hougaard, Frank, Plackett, Normal, Student and Pearson with 4 degrees of freedom.

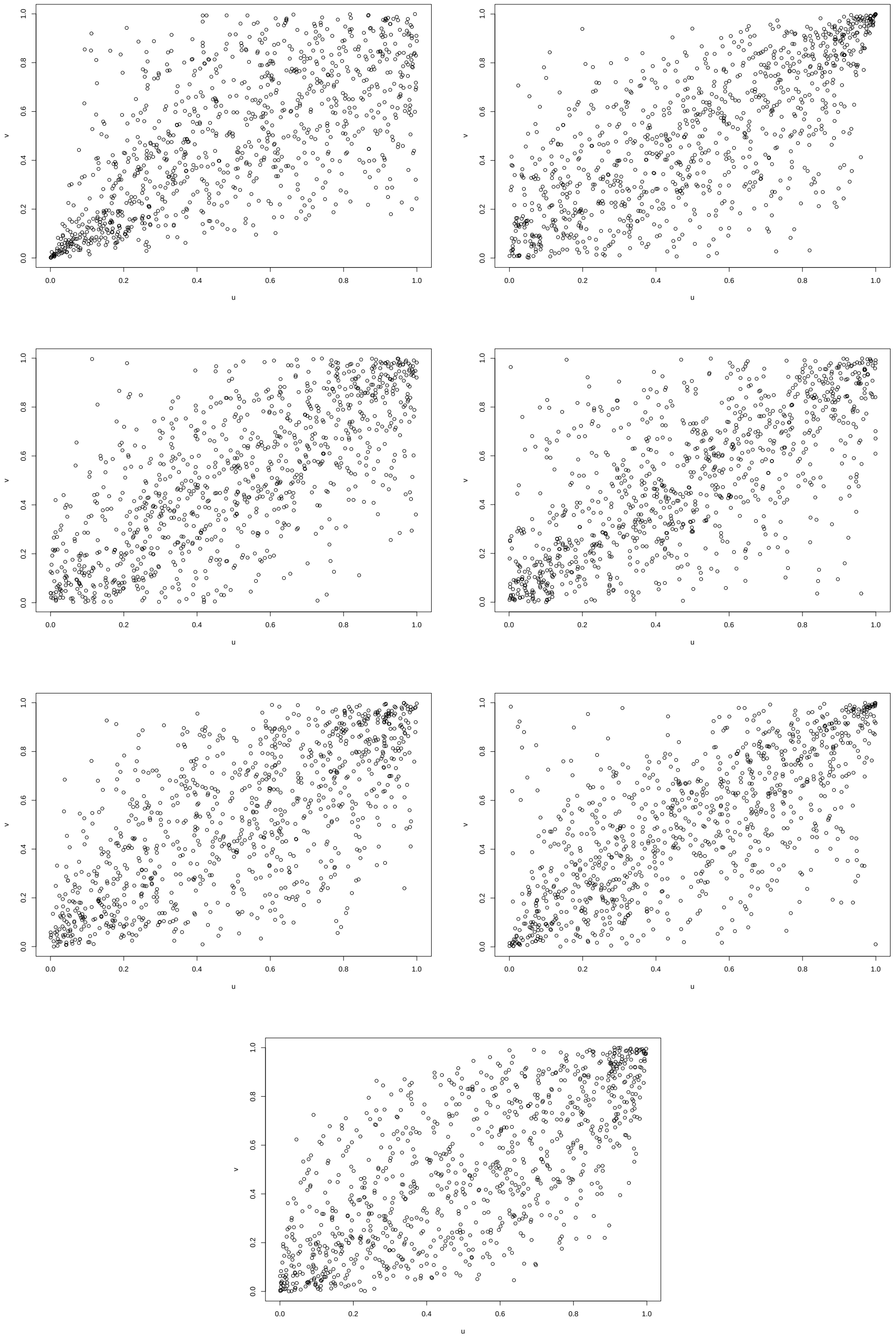


Fig. 3.4 – Samples of size $n = \mathbf{1000}$ from seven different copulas with parameter $\tau = 0.50$. From left to right, and top to bottom : Clayton, Gumbel–Hougaard, Frank, Plackett, Normal, Student and Pearson with 4 degrees of freedom.

To reinforce this view, Fig. 3.5 displays typical graphs of K_n based on samples of size $n = 50$ from the seven copula models, assuming $\tau = 0.25$. Given that the differences between them are small, and considering that goodness-of-fit tests must take into account sampling error, it is easy to understand why the power of the procedures based on $S_n^{(K)}$ and $T_n^{(K)}$ is slow to rise.

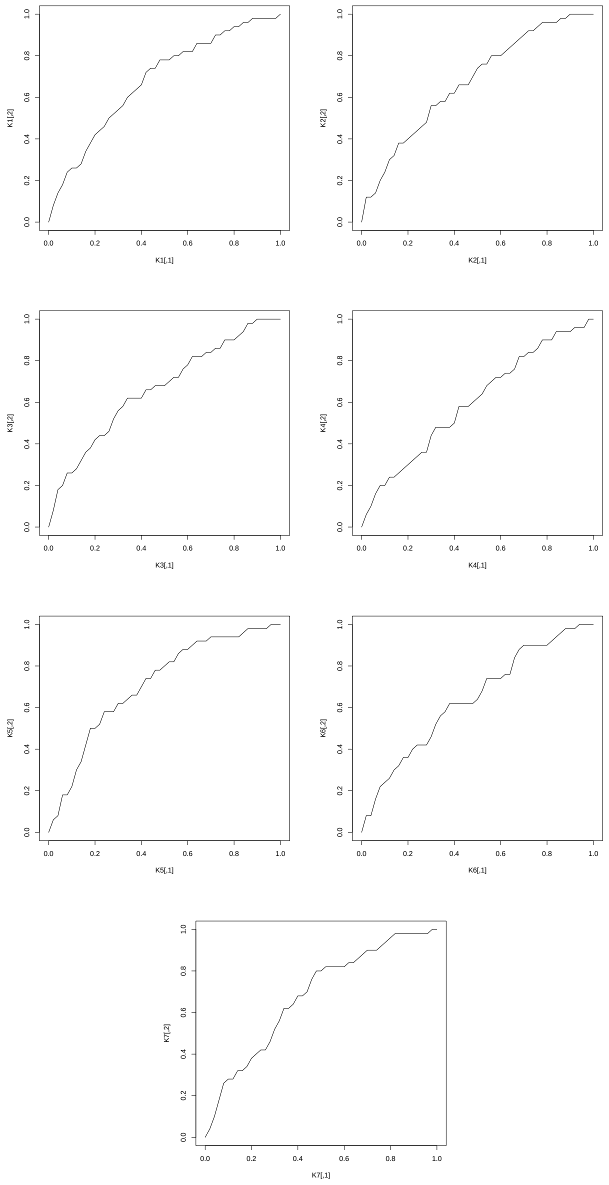


Fig. 3.5 – Examples of graphs of K_n based on samples of size $n = 50$ from seven different copulas with parameter $\tau = 0.25$. From left to right, and top to bottom : Clayton, Gumbel–Hougaard, Frank, Plackett, Normal, Student and Pearson with 4 degrees of freedom.

With increasing sample size, of course, sampling error gets smaller so that in the limit, one can distinguish between models, so long as their asymptotic Kendall distributions are distinct. Such is the case here, as illustrated in Fig. 3.6 when $\tau = 0.5$. Note in passing how Clayton's copula differs markedly from the other six. This helps to understand why power figures associated with this model were often among the largest throughout the study.

As shown by Genest et al. (2006), goodness-of-fit testing using $S_n^{(K)}$ and $T_n^{(K)}$ perform quite well when $n = 250$. As they point out, however, these tests are not generally consistent. In particular, they fail to discriminate bivariate extreme-value copulas having the same level of dependence, because for any such model, the limiting Kendall distribution is $K(w) = w - (1-\tau)w\log(w)$ for all $w \in (0,1)$.

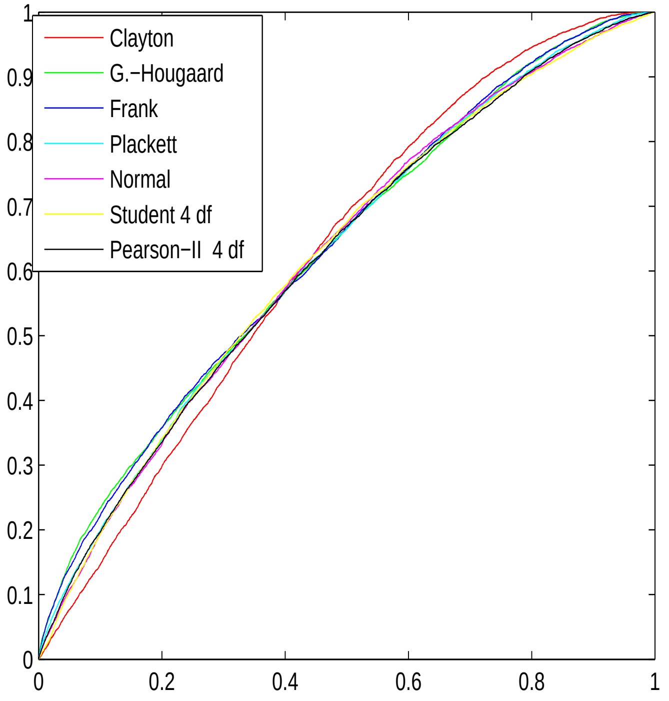


Fig. 3.6 – Asymptotic form of the Kendall distribution K for the seven different copulas with parameter $\tau = 0.50$.

3.6.3 Which test performs best ?

As might have been expected, no single test is preferable to all others, irrespective of the circumstances. Inasmuch as Tables 3.1–3.6 provide an accurate depiction of reality, the choice of the most powerful test would depend on the combination of factors $\mathcal{C}_0$, n, τ and $\mathcal{C}$.

In practice, of course, only the first two factors are known for sure, i.e., the null hypothesis under investigation and the sample size. At the expense of mild "data snooping," one can also get a fairly good idea of the level of dependence in the data, as measured by Kendall's tau. Prior knowledge of the exact nature of dependence in the data, however, would defeat the purpose of goodness-of-fit testing.

In order to extract methodological recommendations from the mass of data contained in Tables 3.1–3.6, it is convenient to rank the tests from 1 to 7 in each of the $252 = 7 \times 6 \times 3 \times 2$ experimental conditions corresponding to the seven possible choices of $\mathcal{C}_0$, the six alternatives $\mathcal{C}$, the three values of tau and the two sample sizes.

Table 3.7 displays average ranks computed over the alternatives, for given $\mathcal{C}_0$, τ and n. In the table, the best test is highlighted in each of the $42 = 7 \times 3 \times 2$ scenarios considered.

Tab. 3.7 – Average ranking **over factor** $\mathcal{C}$ of the seven goodness-of-fit tests in $42 = 7 \times 3 \times 2$ combinations of factors $\mathcal{C}_0$, τ and n.

			Test based on						
H_0	n	τ	S_n	T_n	$S_n^{(K)}$	$T_n^{(K)}$	$S_n^{(B)}$	$S_n^{(C)}$	A_n
Clayton	50	0.25	**7.0**	4.2	2.2	1.0	6.0	4.8	2.8
		0.50	**7.0**	4.8	3.0	1.7	6.0	4.2	1.3
		0.75	**7.0**	4.8	4.0	2.0	5.8	3.3	1.0
	150	0.25	**7.0**	3.5	3.3	1.8	5.8	5.0	1.5
		0.50	**7.0**	3.2	3.8	2.0	6.0	5.0	1.0
		0.75	**5.9**	3.5	4.0	2.0	5.8	5.8	1.0
Gumbel–Hougaard	50	0.25	2.3	4.6	**7.0**	6.0	2.1	2.5	3.6
		0.50	1.8	4.0	**6.7**	5.2	5.5	3.7	1.3
		0.75	1.2	3.7	5.6	2.8	**7.0**	5.1	2.7
	150	0.25	3.6	4.0	**7.0**	5.6	3.7	2.3	1.8
		0.50	3.5	2.8	**6.3**	3.5	6.2	4.7	1.0
		0.75	2.9	2.7	4.8	2.7	**6.8**	5.8	2.5
Frank	50	0.25	4.3	**4.8**	4.2	4.6	2.4	4.3	3.5
		0.50	**5.2**	4.4	**5.2**	3.5	2.8	4.8	2.1
		0.75	**5.5**	3.0	4.3	1.2	4.2	4.7	5.2
	150	0.25	4.6	3.4	**5.3**	4.0	2.8	4.0	3.8
		0.50	5.3	2.8	**5.5**	4.0	3.3	5.2	1.8
		0.75	**5.8**	2.3	4.8	2.2	3.6	5.6	3.7
Plackett	50	0.25	4.3	4.7	4.3	**4.8**	2.8	2.7	4.5
		0.50	4.0	**5.8**	4.5	4.2	3.0	2.2	4.3
		0.75	3.7	**5.5**	4.8	3.8	3.2	1.7	5.3
	150	0.25	4.2	4.1	**4.8**	4.2	3.8	2.8	4.3
		0.50	**4.9**	4.3	4.8	3.3	3.9	2.7	4.1
		0.75	4.8	5.0	4.6	2.8	3.3	2.3	**5.2**
Normal	50	0.25	5.1	4.4	2.3	2.3	4.4	**5.6**	3.8
		0.50	4.8	4.2	2.7	3.0	**5.5**	4.8	3.2
		0.75	4.3	3.2	3.1	2.8	**5.7**	4.8	4.2
	150	0.25	4.7	3.8	3.7	2.4	**5.0**	4.8	3.8
		0.50	4.3	3.3	4.3	2.8	**5.0**	4.7	3.7
		0.75	4.3	3.2	3.7	2.5	**5.5**	4.8	4.1
Student 4 dl	50	0.25	4.6	5.7	2.8	2.8	4.4	1.7	**6.0**
		0.50	4.4	**5.9**	3.4	4.4	4.6	2.3	3.0
		0.75	3.7	4.5	3.8	3.8	**4.7**	3.1	**4.7**
	150	0.25	4.6	4.4	4.5	2.6	4.7	1.9	**5.3**
		0.50	4.8	3.9	5.1	3.0	**5.7**	2.3	3.2
		0.75	3.7	3.3	4.2	3.3	**5.2**	3.5	4.8
Pearson 4 dl	50	0.25	4.4	3.5	2.3	2.4	4.7	**6.5**	4.3
		0.50	4.4	3.6	2.3	2.0	5.8	**6.7**	3.2
		0.75	4.7	2.8	3.3	2.1	5.6	**5.8**	3.8
	150	0.25	4.2	2.2	3.1	2.3	5.3	**6.5**	4.5
		0.50	4.8	3.0	3.3	1.8	**6.2**	**6.2**	2.7
		0.75	4.8	2.3	3.8	1.9	5.8	**5.9**	3.5
Average			4.60	3.88	4.19	3.02	4.75	4.20	3.36
Standard error			1.62	1.38	1.67	1.61	1.65	2.24	2.50

In Table 3.7, the tests are ranked from 1 to 7 in *increasing* order of power. Based on the number of times each test had the highest rank, it appears that :

1. the best procedures overall are those based on $S_n^{(B)}$ and S_n, with 10 and 9.5 "wins," respectively ;
2. the tests based on $S_n^{(K)}$ and $S_n^{(C)}$ are average with 7.5 and 6.5 wins, respectively ;
3. the performance of the tests involving T_n, A_n and $T_n^{(K)}$ is much less impressive, with 4, 3.5 and 1 wins, respectively.

These observations are consistent with the common wisdom of the goodness-of-fit literature, to the effect that test statistics based on the Cramér–von Mises functional of a process tend to be more powerful than those based on the Kolmogorov–Smirnov distance taken on the same process.

It is comforting to see that the preference ranking

$$T_n^{(K)} \prec A_n \prec T_n \prec S_n^{(C)} \prec S_n^{(K)} \prec S_n \prec S_n^{(B)}$$

suggested by the "number of wins" is nearly consistent with the average ranks reported at the bottom of Table 3.7 ; the only inversion is between $S_n^{(C)}$ and $S_n^{(K)}$ (which scored 4.20 and 4.19, respectively). Using the standard deviations given at the bottom of the table, one can also see that within the spectrum of experimental conditions considered, the tests based on A_n and $S_n^{(C)}$ performed much less consistently than the other five procedures.

Other salient features of Table 3.7 are as follows :

1. The two tests based on C_n (13.5 wins together, or 6.75 wins each on average) and the three tests deriving from Rosenblatt's transform (average of 6.67 wins each) outperformed the two tests involving K_n (average of 4.25 wins each) :
$$\{S_n^{(K)}, T_n^{(K)}\} \prec \{S_n^{(B)}, S_n^{(C)}, A_n\} \sim \{S_n, T_n\}.$$
2. The statistic S_n unequivocally yields the most powerful test of the Clayton hypothesis ; it also does quite well for goodness-of-fit testing of Frank's model.
3. The test based on $S_n^{(B)}$ seems particularly good at detecting lack of normal or Student types of dependence, while $S_n^{(C)}$ is most powerful for the Pearson hypothesis ; it would be interesting to see whether this conclusion extends to other meta-elliptical copula structures.
4. Among the tests constructed using the Kendall transform, the procedure based on $S_n^{(K)}$ was far superior and offered the best performance when testing the goodness-of-fit of Gumbel–Hougaard and Frank copula structures.
5. The statistics T_n, $T_n^{(K)}$ and A_n enjoyed mitigated success.
6. No clear recommendation emerges for goodness-of-fit testing of the Plackett.

Tab. 3.8 – Average ranking **over factor** $\mathcal{C}_0$ of the seven goodness-of-fit tests in $42 = 7 \times 3 \times 2$ combinations of factors $\mathcal{C}$, τ and n.

			Test based on						
True model	n	τ	S_n	T_n	$S_n^{(K)}$	$T_n^{(K)}$	$S_n^{(B)}$	$S_n^{(C)}$	A_n
Clayton	50	0.25	2.8	5.0	**7.0**	6.0	2.8	2.7	1.7
		0.50	2.0	4.0	**7.0**	5.5	5.0	3.5	1.0
		0.75	2.2	4.0	5.7	2.8	**7.0**	5.3	1.0
	150	0.25	4.3	3.2	**7.0**	6.0	3.7	2.8	1.0
		0.50	4.5	2.2	**6.5**	4.0	5.5	4.3	1.0
		0.75	4.3	2.8	5.1	2.2	**6.5**	6.2	1.0
Gumbel–Hougaard	50	0.25	**6.8**	4.2	1.8	1.2	5.6	5.3	3.2
		0.50	**6.8**	4.8	2.5	1.2	5.3	5.2	2.3
		0.75	**7.0**	4.1	3.1	1.9	5.2	4.3	2.5
	150	0.25	**6.3**	3.8	2.3	1.2	5.7	6.0	2.7
		0.50	**6.3**	3.8	3.2	2.0	5.7	6.0	1.0
		0.75	5.7	3.3	3.8	1.7	**5.8**	5.7	2.0
Frank	50	0.25	**5.3**	4.8	2.8	3.0	3.8	3.9	4.5
		0.50	4.8	**5.3**	3.2	3.6	**5.3**	3.3	2.5
		0.75	4.7	3.2	**5.3**	3.8	3.9	2.0	5.2
	150	0.25	**5.5**	4.3	4.2	2.5	4.7	3.2	3.7
		0.50	**5.7**	3.7	5.5	3.0	5.5	2.5	2.2
		0.75	4.8	2.8	**5.2**	2.8	4.7	2.5	**5.2**
Plackett	50	0.25	4.5	4.7	2.5	2.2	4.3	**5.5**	4.3
		0.50	4.9	3.8	2.3	2.1	**5.8**	**5.8**	3.3
		0.75	4.3	2.3	3.2	1.8	6.0	**6.2**	4.3
	150	0.25	**5.1**	3.9	3.8	2.2	4.4	4.7	4.0
		0.50	5.0	3.0	3.5	1.8	5.3	**5.5**	3.8
		0.75	4.4	2.3	3.1	1.8	5.8	**6.4**	4.2
Normal	50	0.25	4.6	**4.7**	4.0	4.3	3.3	3.2	3.9
		0.50	4.2	**5.4**	4.3	3.9	4.2	3.2	2.9
		0.75	3.8	**4.9**	3.9	2.8	4.5	3.4	4.7
	150	0.25	4.3	3.9	**5.1**	4.3	4.3	3.2	3.0
		0.50	4.5	4.0	4.7	3.3	**4.8**	3.5	3.2
		0.75	**4.4**	3.5	4.3	3.3	4.0	4.2	4.3
Student 4 dl	50	0.25	3.8	2.8	3.2	3.7	3.2	**5.7**	**5.7**
		0.50	5.3	3.7	4.2	2.8	4.0	**5.5**	2.6
		0.75	4.7	3.9	3.7	1.9	**4.8**	**4.8**	4.2
	150	0.25	3.8	1.3	4.2	3.3	3.7	5.8	**6.0**
		0.50	4.9	2.7	4.8	2.7	4.4	**6.2**	2.4
		0.75	5.0	2.8	4.3	2.4	4.8	**5.5**	3.3
Pearson 4 dl	50	0.25	4.2	**5.7**	3.8	3.7	3.7	1.8	5.3
		0.50	3.5	**5.8**	4.3	4.8	3.7	2.2	3.7
		0.75	3.5	**5.0**	4.0	3.3	4.7	2.5	**5.0**
	150	0.25	3.4	4.8	**5.0**	3.5	4.7	1.8	4.8
		0.50	3.8	4.1	**5.0**	3.6	**5.0**	2.7	3.8
		0.75	3.7	4.7	4.2	3.3	4.3	3.2	**4.8**
Average			4.60	3.88	4.19	3.02	4.75	4.20	3.36
Standard error			1.62	1.38	1.67	1.61	1.65	2.24	2.50

As a complement, the same 252 rankings of the data were averaged over the range of null hypotheses. The results are displayed in Table 3.8, in which the best test is highlighted in each of the $42 = 7 \times 3 \times 2$ combinations of factors $\mathcal{C}$, τ and n considered.

Although the averages and standard deviations appearing at the bottom of the table are obviously the same as those reported in Table 3.7, the fact that rankings are averaged over $\mathcal{C}_0$ rather than $\mathcal{C}$ leads to different numbers of wins for the statistics.

While the data in Table 3.7 are indicative of the tests that are most powerful to test a specific null hypothesis, Table 3.8 provides information on the ability of the test statistics to pick up the characteristics of a given type of dependence structure.

One can see from Table 3.8 that all test statistics had approximately the same number of wins, except A_n and $T_n^{(K)}$. More specifically :

1. The best statistics were S_n, $S_n^{(C)}$ and $S_n^{(K)}$, with 10, 8.5 and 8 wins, respectively.
2. The statistics $S_n^{(B)}$ and T_n had an average performance, with 6 wins each.
3. The procedures based on A_n and $T_n^{(K)}$ were much less successful ; A_n won only 3.5 times, and $T_n^{(K)}$ was never preferred.

The overall ordering

$$T_n^{(K)} \quad \prec \quad A_n \quad \prec \quad T_n \quad \sim \quad S_n^{(B)} \quad \prec \quad S_n^{(K)} \quad \prec \quad S_n^{(C)} \quad \prec \quad S_n$$

is not as useful as that which is based on Table 3.7, unless one knows from experience what type of dependence structure to expect in the data at hand. Nevertheless, it is instructive to see that S_n is quite sensitive to the features of the Gumbel–Hougaard and Frank dependence structures, while $S_n^{(K)}$ picks up the characteristics of the Clayton model more than any other.

The main difference between statistics $S_n^{(C)}$ and $S_n^{(B)}$ is also apparent from Table 3.8 : while the successes of the former were mostly associated with the Plackett and the Student, the latter was just good overall. Finally, it is worth noting that T_n performed well for the normal and Pearson copulas.

3.7 Observations and recommendations

Based on the experience gained in carrying this comparative power study of the existing omnibus goodness-of-fit tests for copula models, the following general observations and specific recommendations can be made.

1. General observations :

   (a) In goodness-of-fit testing as in any other inferential context, the greater the sample size, the better. Large data sets not only help to distinguish between copula models but play a role in the reliability of the parametric bootstrap procedures used to approximate the statistics' null distribution.

   (b) In order for the double bootstrap to be efficient, the number m of repetitions must be substantially larger than the sample size n. In the present study, $m = 2500$ was found to be an acceptable minimum. While this is not a problem when using a test once, it quickly becomes computationally demanding in the context of a simulation study. In the present case, the recourse to a double bootstrap whenever C_θ or K_θ was not available in closed form made it totally impractical to run the experiment at a sample size of $n = 250$, for lack of sufficient computing resources.

   (c) In this regard, the tests based on A_n, $S_n^{(B)}$ and $S_n^{(C)}$ are at an advantage : because they rely on Rosenblatt's transform, a single bootstrap is enough to approximate their null distribution and extract P-values. However, these statistics are dependent on the order in which the variables are successively conditioned. While it is traditional to take $U_2|U_1$, $U_3|(U_1, U_2), \ldots$ as in (3.5), any other sequence could be used, leading in turn to different statistics.

   (d) When statistics based on Cramér–von Mises and Kolmogorov–Smirnov functionals of the same empirical process are compared, the former are almost invariably more powerful. The present simulations and those reported earlier by Genest et al. (2006) both point strongly in that direction.

2. Specific recommendations, based on the present state of knowledge :

   (a) Overall, statistics S_n and $S_n^{(B)}$ yield the best omnibus goodness-of-fit test procedures for copula models. While $S_n^{(B)}$ is slightly more consistent than S_n in its performance across models and can be implemented without ever calling upon a double bootstrap, it relies on a non-unique (and therefore somewhat arbitrary) Rosenblatt transform.

   (b) Statistics $S_n^{(C)}$ and $S_n^{(K)}$ are also recommendable, and the latter is especially convenient when the null hypothesis is Archimedean, since the Kendall distribution K is then available in closed form.

   (c) The jury is still out on the merits of the test based on A_n. Anderson–Darling type statistics have proved useful in many other contexts, particularly in circumstances where differences in the tail of a distribution were deemed to be important. While it seems plausible that the same would hold in a copula context, the simulation results are not convincing in this regard.

(d) There are no strong arguments in favor of using the tests based on T_n and $T_n^{(K)}$. As for the uncorrected version of the test proposed by Breymann et al. (2003), it should never be used.

In future work, it would be interesting to investigate the sensitivity of tests based on the Rosenblatt transform to the order in which conditioning is done. It would also be useful to expand the present study to include comparisons with general goodness-of-fit tests involving tuning parameters, as well as with procedures developed to test for specific dependence structures such as the Clayton or the normal copula.

Any extension promises to be a formidable task, however. Given the number of options, the need to rely on numerical algorithms and the sample sizes required to reach meaningful conclusions, the demand on computing resources will be enormous. In addition, the analysis and interpretation of the results may prove challenging.

On the theoretical front, several of the procedures that have been proposed recently for goodness-of-fit testing of copula models remain on shaky grounds. As illustrated by the appalling performance of the test proposed by Breymann et al. (2003), the dependence between pseudo-observations must imperatively be taken into account.

Non-trivial mathematics are required before one can conclude (or not) that the limiting distribution of a rank-based statistic is the same as in the classical multivariate context in which it was originally developed. Furthermore, conditions are required for the convergence of bootstrap algorithms, and failure to check them may lead to disaster. No sleight of hand will change that fact.

Given their relative simplicity and success, and notwithstanding their detractors (see notably Mikosch (2006) and the ensuing discussion), copulas are increasingly used, particularly in actuarial and financial practice. The question of “which copula model is the right one ?” is therefore a crucial one, and the temptation is great to improvise goodness-of-fit tests on the spur of the moment. While it may be relatively simple to come up with a procedure that seems sensible, end users should beware of non-validated tests that are described in a fling and for which bootstrapping is recommended without precautions and qualifications, as if it were a panacea.

Acknowledgments

Partial funding in support of this work was granted by the Natural Sciences and Engineering Research Council of Canada, by the Fonds québécois de la recherche sur la nature et les technologies, and by the Institut de finance mathématique de Montréal. The authors gratefully acknowledge the GERAD (Montréal) and the Salle des marchés at Université Laval for extensive use of their computing facilities.

Appendix A : A parametric bootstrap for S_n and T_n

The following procedure leads to an approximate P-value for the test based on S_n. The adaptations required for T_n or any other rank-based statistic are obvious.

1. Compute C_n as per formula (3.1).
2. Estimate θ with $\theta_n = \mathcal{T}_n(\mathbf{U}_1, \ldots, \mathbf{U}_n)$.
3. If there is an analytical expression for C_θ, compute the value of S_n, as defined in (3.2). Otherwise, proceed by Monte Carlo approximation. Specifically, choose $m \geq n$ and carry out the following extra steps :
   (a) Generate a random sample $\mathbf{U}_1^*, \ldots, \mathbf{U}_m^*$ from distribution C_{θ_n}.
   (b) Approximate C_{θ_n} by
   $$B_m^*(\mathbf{u}) = \frac{1}{m} \sum_{i=1}^{m} 1\left(\mathbf{U}_i^* \leq \mathbf{u}\right), \quad \mathbf{u} \in [0,1]^d.$$
   (c) Approximate S_n by
   $$S_n = n \int_{[0,1]^d} \{C_n(\mathbf{u}) - B_m^*(\mathbf{u})\}^2 \, dC_n(\mathbf{u}) = \sum_{i=1}^{n} \left\{C_n\left(\mathbf{U}_i\right) - B_m^*\left(\mathbf{U}_i\right)\right\}^2.$$
4. For some large integer N, repeat the following steps for every $k \in \{1, \ldots, N\}$:
   (a) Generate a random sample $\mathbf{Y}_{1,k}^*, \ldots, \mathbf{Y}_{n,k}^*$ from distribution C_{θ_n} and compute their associated rank vectors $\mathbf{R}_{1,k}^*, \ldots, \mathbf{R}_{n,k}^*$.
   (b) Compute $\mathbf{U}_{i,k}^* = \mathbf{R}_{i,k}^*/(n+1)$ for $i \in \{1, \ldots, n\}$ and let
   $$C_{n,k}^*(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^{n} 1\left(\mathbf{U}_{i,k}^* \leq \mathbf{u}\right), \quad \mathbf{u} \in [0,1]^d.$$
   (c) Estimate θ with $\theta_{n,k}^* = \mathcal{T}_n(\mathbf{U}_{1,k}^*, \ldots, \mathbf{U}_{n,k}^*)$.
   (d) If there is an analytical expression for C_θ, let
   $$S_{n,k}^* = \sum_{i=1}^{n} \{C_{n,k}^*\left(\mathbf{U}_{i,k}^*\right) - C_{\theta_{n,k}^*}\left(\mathbf{U}_{i,k}^*\right)\}^2.$$
   Otherwise, proceed as follows :
      i. Generate a random sample $\mathbf{Y}_{1,k}^{**}, \ldots, \mathbf{Y}_{m,k}^{**}$ from distribution $C_{\theta_{n,k}^*}$.
      ii. Approximate $C_{\theta_{n,k}^*}$ by
      $$B_{m,k}^{**}(\mathbf{u}) = \frac{1}{m} \sum_{i=1}^{m} 1\left(\mathbf{Y}_{i,k}^{**} \leq \mathbf{u}\right), \quad \mathbf{u} \in [0,1]^d$$
      and let
      $$S_{n,k}^* = \sum_{i=1}^{n} \left\{C_{n,k}^*\left(\mathbf{U}_{i,k}^*\right) - B_{m,k}^{**}\left(\mathbf{U}_{i,k}^*\right)\right\}^2.$$

An approximate P-value for the test is then given by

$$\frac{1}{N} \sum_{k=1}^{N} 1(S_{n,k}^* > S_n).$$

Appendix B : A parametric bootstrap for $S_n^{(K)}$ and $T_n^{(K)}$

For the sake of simplicity, the following algorithm is described in terms of statistic $S_n^{(K)}$. However, it is also valid *mutatis mutandis* for $T_n^{(K)}$ or any other rank-based statistic.

1. Compute K_n as per formula (3.3).
2. Estimate θ with $\theta_n = \mathcal{T}_n(\mathbf{U}_1, \ldots, \mathbf{U}_n)$.
3. If there is an analytical expression for K_θ, compute the value of $S_n^{(K)}$, as defined in (3.4). Otherwise, proceed by Monte Carlo approximation. Specifically, choose $m \geq n$ and carry out the following extra steps :
   (a) Generate a random sample $\mathbf{U}_1^*, \ldots, \mathbf{U}_m^*$ from distribution C_{θ_n}.
   (b) Approximate K_{θ_n} by
   $$B_m^*(t) = \frac{1}{m}\sum_{i=1}^m 1\left(V_i^* \leq t\right), \quad t \in [0,1],$$
   where
   $$V_i^* = \frac{1}{m}\sum_{j=1}^m 1\left(\mathbf{U}_j^* \leq \mathbf{U}_i^*\right), \quad i \in \{1, \ldots, m\}.$$
   (c) Approximate $S_n^{(K)}$ by
   $$S_n^{(K)} = n\int_{[0,1]} \{K_n(t) - B_m^*(t)\}^2\, dB_m^*(t) = \frac{n}{m}\sum_{i=1}^m \{K_n(V_i^*) - B_m^*(V_i^*)\}^2.$$
   Note in passing that $m \times B_m^*(V_i^*)$ is in fact nothing but the rank of V_i^* among $V_1^*, \ldots, V_m^*$.
4. For some large integer N, repeat the following steps for every $k \in \{1, \ldots, N\}$:
   (a) Generate a random sample $\mathbf{Y}_{1,k}^*, \ldots, \mathbf{Y}_{n,k}^*$ from distribution C_{θ_n} and compute their associated rank vectors $\mathbf{R}_{1,k}^*, \ldots, \mathbf{R}_{n,k}^*$.
   (b) Compute
   $$V_{i,k}^* = \frac{1}{n}\sum_{j=1}^n 1\left(\mathbf{Y}_{j,k}^* \leq \mathbf{Y}_{i,k}^*\right), \quad i \in \{1, \ldots, n\}$$
   and let
   $$K_{n,k}^*(t) = \frac{1}{n}\sum_{i=1}^n 1\left(V_{i,k}^* \leq t\right), \quad t \in [0,1].$$
   (c) Estimate θ with $\theta_{n,k}^* = \mathcal{T}_n\{\mathbf{R}_{1,k}^*/(n+1), \ldots, \mathbf{R}_{n,k}^*/(n+1)\}$.
   (d) If there is an analytical expression for K_θ, let
   $$S_{n,k}^{(K)*} = \int_0^1 \{C_{n,k}^*(t) - K_{\theta_{n,k}^*}(t)\}^2 dK_{\theta_{n,k}^*}(t),$$
   for which an explicit expression can easily be deduced from (3.4).
   Otherwise, proceed as follows :

i. Generate a random sample $\mathbf{Y}^{**}_{1,k}, \ldots, \mathbf{Y}^{**}_{m,k}$ from distribution $C_{\theta^*_{n,k}}$.

ii. Approximate $K^*_{\theta_{n,k}}$ by

$$B^{**}_{m,k}(t) = \frac{1}{m}\sum_{i=1}^{m} 1\left(V^{**}_{i,k} \le t\right), \quad t \in [0,1]$$

where

$$V^{**}_{i,k} = \frac{1}{m}\sum_{j=1}^{m} 1\left(\mathbf{Y}^{*}_{j,k} \le \mathbf{Y}^{*}_{i,k}\right), \quad i \in \{1, \ldots, m\}.$$

Then set

$$S^{(K)*}_{n,k} = \frac{n}{m}\sum_{i=1}^{m} \left\{K^*_{n,k}\left(V^*_{i,k}\right) - B^{**}_{m,k}\left(V^*_{i,k}\right)\right\}^2,$$

where $m \times B^{**}_{m,k}(V^*_{i,k})$ is in fact nothing but the rank of $V^*_{i,k}$ among $V^*_{1,k}$, …, $V^*_{m,k}$.

An approximate P-value for the test is then given by

$$\frac{1}{N}\sum_{k=1}^{N} 1(S^{(K)*}_{n,k} > S^{(K)}_{n}).$$

Appendix C : A parametric bootstrap for A_n

Although the following algorithm is described in terms of statistic A_n, it is also valid *mutatis mutandis* for any other rank-based statistic based on $\chi_1, \ldots, \chi_n$.

1. Compute
$$G_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1\left(\chi_i \leq t\right), \quad t \geq 0.$$
2. Estimate θ with $\theta_n = \mathcal{T}_n\left(\mathbf{U}_1, \ldots, \mathbf{U}_n\right)$.
3. Compute the value of A_n as per formula (3.7).
4. For some large integer N, repeat the following steps for every $k \in \{1, \ldots, N\}$:
   (a) Generate a random sample $\mathbf{Y}^*_{1,k}, \ldots, \mathbf{Y}^*_{n,k}$ from distribution C_{θ_n} and compute their associated rank vectors $\mathbf{R}^*_{1,k}, \ldots, \mathbf{R}^*_{n,k}$.
   (b) Compute $\mathbf{U}^*_{i,k} = \mathbf{R}^*_{i,k}/(n+1)$ for $i \in \{1, \ldots, n\}$,
   (c) Estimate θ with $\theta^*_{n,k} = \mathcal{T}_n(\mathbf{U}^*_{1,k}, \ldots, \mathbf{U}^*_{n,k})$, and compute $\chi^*_{1,k}, \ldots, \chi^*_{n,k}$, where
$$\chi^*_{i,k} = \sum_{j=1}^{d}\left\{\Phi^{-1}(E^*_{ij,k})\right\}^2 \quad \text{and} \quad \mathbf{E}^*_{i,k} = \mathcal{R}_{\theta^*_{n,k}}\left(\mathbf{U}^*_{i,k}\right), \quad i \in \{i, \ldots, n\}.$$
   (d) Let
$$G^*_{n,k}(t) = \frac{1}{n}\sum_{i=1}^{n} 1\left(\chi^*_{i,k} \leq t\right), \quad t \geq 0$$
   and define
$$A^*_{n,k} = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\cdot[\log\{G(\chi^*_{(i),k})\} + \log\{1 - G(\chi^*_{(n+1-i),k}\ )\}].$$

An approximate P-value for the test is then given by
$$\frac{1}{N}\sum_{k=1}^{N} 1(A^*_{n,k} > A_n).$$

Appendix D : A parametric bootstrap for $S_n^{(C)}$ and $S_n^{(B)}$

The following algorithm is described in terms of statistic $S_n^{(C)}$. However, it is also valid *mutatis mutandis* for $S_n^{(B)}$ or any other rank-based statistic.

1. Compute
$$D_n(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} 1\left(\mathbf{E}_i \le \mathbf{u}\right), \quad \mathbf{u} \in [0,1]^d.$$
2. Estimate θ with $\theta_n = \mathcal{T}_n\left(\mathbf{U}_1, \ldots, \mathbf{U}_n\right)$.
3. Compute the value of $S_n^{(C)}$, as defined in (3.8).

d) For some large integer N, repeat the following steps for every $k \in \{1, \ldots, N\}$:
   (a) Generate a random sample $\mathbf{Y}^*_{1,k}, \ldots, \mathbf{Y}^*_{n,k}$ from distribution C_{θ_n} and compute their associated rank vectors $\mathbf{R}^*_{1,k}, \ldots, \mathbf{R}^*_{n,k}$.
   (b) Compute $\mathbf{U}^*_{i,k} = \mathbf{R}^*_{i,k}/(n+1)$ for $i \in \{1, \ldots, n\}$.
   (c) Estimate θ with $\theta^*_{n,k} = \mathcal{T}_n(\mathbf{U}^*_{1,k}, \ldots, \mathbf{U}^*_{n,k})$ and compute $\mathbf{E}^*_{1,k}, \ldots, \mathbf{E}^*_{n,k}$, where
$$\mathbf{E}^*_{i,k} = \mathcal{R}_{\theta^*_{n,k}}\left(\mathbf{U}^*_{i,k}\right), \quad i \in \{i, \ldots, n\}.$$
   (d) Let
$$D^*_{n,k}(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} 1\left(\mathbf{E}^*_{i,k} \le \mathbf{u}\right), \quad \mathbf{u} \in [0,1]^d$$
   and set
$$S^{(C)*}_{n,k} = \sum_{i=1}^{n}\left\{D^*_{n,k}\left(\mathbf{E}^*_{i,k}\right) - C_\perp\left(\mathbf{E}^*_{i,k}\right)\right\}^2.$$

An approximate P-value for the test based on $S_n^{(C)}$ is then given by
$$\frac{1}{N}\sum_{k=1}^{N} 1(S^{(C)*}_{n,k} > S_n^{(C)}).$$

Transition

Le chapitre 3 a présenté des tests d'adéquation de modèles de copules en présence de données complètes. Seuls les tests dits *omnibus* ont été retenus dans le cadre d'une vaste étude de Monte-Carlo. Sept des huit tests considérés semblent respecter leur seuil nominal. Quant à l'étude de puissance, elle tend à montrer que les tests basés sur une statistique du type Cramér–von Mises sont plus puissants que ceux qui s'appuient sur une statistique de Kolmogorov–Smirnov ou de Anderson–Darling.

La présence de censure ou de troncation dans les données complique grandement les analyses statistiques. Récemment, plusieurs auteurs se sont intéressés au problème d'adéquation de modèles de copules dans le contexte de données censurées. Dans ces travaux, l'estimation du paramètre du modèle de copule passe souvent par celle du tau de Kendall. Le problème se pose donc de savoir comment estimer le tau de Kendall en présence de censure dans les données. Quelques travaux ont d'ailleurs déjà été réalisés récemment à ce sujet, notamment par Wang & Wells (2000a), Chen & Bandeen-Roche (2005) ainsi que Andersen et al. (2005).

L'article qui suit traite de ce problème particulier en se restreignant au cas où une seule des deux variables à l'étude est sujette au phénomène de censure. Soit Y la variable éventuellement censurée, X la covariable et $\hat{S}_{Y|X}$ un estimateur de la survie conditionnelle de Y sachant $X = x$. L'apport majeur du chapitre 4 réside dans le fait que nous proposons des estimateurs du tau de Kendall qui tirent profit de l'information fournie par $\hat{S}_{Y|X}$. Ceux-ci s'avèrent plus précis que les estimateurs se limitant à l'information contenue dans les lois marginales de X et de Y.

Chapitre 4

Improving the estimation of Kendall's tau when censoring affects only one of the variables

Résumé

Cet article considère l'estimation du tau de Kendall en présence de données bivariées (X, Y) où seule la variable Y est sujette au phénomène de censure à droite. Malgré le fait que τ soit estimable sous certaines conditions de régularité, les estimateurs proposés par Brown et al. (1974), Weier & Basu (1980) et Oakes (1982) ne sont pas convergents lorsque $\tau \neq 0$ parce qu'ils ne font usage que de l'information contenue dans les lois marginales. L'estimateur renormalisé de Oakes (2006) fait exception à cette règle. En effet, cette procédure mène à un estimateur convergent pour toute valeur de τ, mais seulement dans le contexte du modèle de susceptibilité gamma. Wang & Wells (2000a) ont été les premiers auteurs à développer un estimateur qui tient compte de l'information jointe. Quatre nouvelles méthodes sont détaillées ici : les trois premières sont en fait des extensions des procédures de Brown et al. (1974), Weier & Basu (1980) et Oakes (1982) qui prennent en compte l'information fournie par X, alors que le quatrième estimateur inverse une estimation de $\Pr(Y_i \leq y | X_i = x_i, Y_i > c_i)$ en vue d'obtenir une imputation de la vraie valeur de Y_i censurée en $C_i = c_i$. L'estimateur non paramétrique de Lim (2006) est également considéré. Ce dernier repose sur le calcul de la moyenne des valeurs $\hat{\tau}_i$ obtenues via un grand nombre de configurations possibles du jeu de données observé $(X_1, Z_1), \ldots, (X_n, Z_n)$, où $Z_i = \min(Y_i, C_i)$. Ces divers estimateurs du tau de Kendall sont comparés par l'entremise d'une vaste étude de simulation. Les méthodes sont aussi illustrées à l'aide de données de transplantation cardiaque de Stanford.

Abstract

This paper considers the estimation of Kendall's tau for bivariate data (X, Y) when only Y is subject to right-censoring. Although τ is estimable under weak regularity conditions, the estimators proposed by Brown et al. (1974), Weier & Basu (1980) and Oakes (1982), which are standard in this context, fail to be consistent when $\tau \neq 0$ because they only use information from the marginal distributions. An exception is the renormalized estimator of Oakes (2006), whose consistency has been established for all possible values of τ, but only in the context of the gamma frailty model. Wang & Wells (2000a) were the first to propose an estimator which accounts for joint information. Four more are developed here : the first three extend the methods of Brown et al. (1974), Weier & Basu (1980) and Oakes (1982) to account for information provided by X, while the fourth estimator inverts an estimation of $\Pr(Y_i \leq y | X_i = x_i, Y_i > c_i)$ to get an imputation of the value of Y_i censored at $C_i = c_i$. Following Lim (2006), a nonparametric estimator is also considered which averages the $\hat{\tau}_i$ obtained from a large number of possible configurations of the observed data $(X_1, Z_1), \ldots, (X_n, Z_n)$, where $Z_i = \min(Y_i, C_i)$. Simulations are presented which compare these various estimators of Kendall's tau. An illustration involving the well-known Stanford heart transplant data is also presented.

4.1 Introduction

Consider a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from a continuous random pair (X, Y) with joint cumulative distribution function $H(x, y)$ and margins $F(x)$ and $G(y)$. When the data arise from clinical studies, it may sometimes happen that Y is right-censored, while X is observed and treated as a covariate. For example, Y might represent time to relapse following a surgery, and X might be the age, weight or blood pressure of the patient. In such a context, the usual objective would be to construct a regression model for the distribution of Y given $X = x$.

Before attempting to model the relationship between two variables X and Y, it may be helpful to assess the level of association between them using model-free tools. Kendall's tau is useful for this purpose, because neither its definition nor its estimation require specific knowledge of the parametric form of the marginal distributions for X and Y.

Different estimators of Kendall's tau are available when neither variable is censored (Kendall, 1970) or when both are subject to right-censoring ; see Brown et al. (1974), Weier & Basu (1980), Oakes (1982), Wang & Wells (2000a), Oakes (2006), and Lim (2006). However, to the best of the authors' knowledge, no method has been especially designed to make inferences about Kendall's tau when only one of the variables is subject to censoring. As will be seen in the simulation study of Section 4.5, methods that take advantage of the fact that one of the variables is not censored perform better than estimators not specifically designed to handle this case.

In the absence of censoring, the sample value of Kendall's tau is given by

$$\hat{\tau}_n = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} w_{ij}, \tag{4.1}$$

where

$$w_{ij} = \begin{cases} -1 & \text{if } (X_i - X_j)(Y_i - Y_j) < 0, \\ 1 & \text{if } (X_i - X_j)(Y_i - Y_j) > 0. \end{cases}$$

The pairs (X_i, Y_i), (X_j, Y_j) are said to be discordant whenever $w_{ij} = -1$ and concordant otherwise. When X and Y are continuous, as assumed herein, the event $(X_i - X_j)(Y_i - Y_j) = 0$ occurs with probability zero.

From standard theory (Hoeffding, 1947), the statistic $\hat{\tau}_n$ is an unbiased, asymptotically normal estimator of the population parameter

$$\tau = \Pr\{(X - X')(Y - Y') > 0\} - \Pr\{(X - X')(Y - Y') < 0\},$$

where (X', Y') is an independent copy of (X, Y).

As is well known, $\tau \in [-1, 1]$ and vanishes in case of independence between X and Y. Furthermore, $\tau = \pm 1$ if and only if Y is a monotone function of X, i.e., $Y = G^{-1}\{F(X)\}$ or $Y = G^{-1}\{1 - F(X)\}$ almost surely. This is as opposed to Pearson's correlation coefficient, which only assumes its extreme values in the presence of *linear* dependence between the variables. For additional discussion concerning the limitations of standard correlation and alternate, margin-free measures of dependence, see, e.g., Embrechts et al. (2002) and Genest & Verret (2005).

Now suppose that because of censoring in Y, the observed data are of the form (X_i, Z_i, δ_i), where for $i \in \{1, \ldots, n\}$, $Z_i = \min(Y_i, C_i)$ and $\delta_i = 1(Y_i \leq C_i)$. Here, C is a censoring variable whose value is assumed to be independent of Y given X. Formula (4.1) is no longer applicable, because the concordance or discordance status of certain pairs (X_i, Y_i), (X_j, Y_j) is undetermined. This happens when $\delta_k = 0$ for

$$k = i \times 1(Z_i \leq Z_j) + j \times 1(Z_i > Z_j). \tag{4.2}$$

In other words, one cannot tell for sure whether the two pairs are concordant or discordant whenever $\min(Z_i, Z_j)$ is censored.

Can something still be done ? Yes ! Indeed, under mild regularity conditions, τ remains estimable from censored data. The appropriate conditions are recalled in Section 4.2, along with six existing estimators of τ in the presence of censoring. New estimating strategies are then proposed in Section 4.3. All of them involve the conditional distribution $H(y|x) = \Pr(Y \leq y|\ X = x)$. Kernel-based techniques for estimating the latter are considered in Section 4.4. In total, various combinations of estimates for τ and $H(y|x)$ lead to nine new estimators for τ under censoring.

Section 4.5 presents the results of a Monte Carlo study comparing the performance of the new estimators to the current ones. The comparisons are made under different choices of sample sizes and censoring fractions, as well as for various degrees and structures of dependence. The robustness of the conclusions to choices of kernel and bandwidth is then analyzed in Section 4.6, along with the impact of variations in the marginal distributions of the observations. The Stanford heart transplant data are then used in Section 4.7 to illustrate the different approaches to estimating τ. Concluding comments may be found in Section 4.8.

4.2 Existing estimators and conditions for their consistency

Six estimators of Kendall's tau under censoring are currently available in the literature. They are described below, along with conditions under which this parameter can be estimated consistently.

4.2.1 The estimators of Oakes (1982, 2006)

When censoring makes it impossible to determine the concordance or discordance status of some pairs $(X_i, Y_i), (X_j, Y_j)$, the simplest solution probably consists in setting $w_{ij} = 0$ in formula (4.1) whenever in doubt. This proposal was made by Oakes (1982), who shows that the resulting estimate is asymptotically normal and unbiased, under the null hypothesis of independence between X and Y. However, Wang & Wells (2000a) point out that this estimator is inconsistent when $\tau \neq 0$.

To improve the performance of his estimator when $\tau \neq 0$, Oakes (2006) suggests a renormalization. Specifically, the denominator in (4.1) should be replaced by the number of pairs whose concordance-discordance status is certain. The consistency of this renormalized estimator is established by Oakes (2006) in the context of the gamma frailty model, which has become very popular in survival analysis since it was introduced by Clayton (1978). Both versions of Oakes' estimator will be considered in the simulations.

4.2.2 The estimator of Brown et al. (1974)

An alternative approach has been suggested by Brown et al. (1974). Their estimator is defined by

$$\tilde{\tau}_n = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} b_{ij}}{\sqrt{\left(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2\right)\left(\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}^2\right)}}, \tag{4.3}$$

where $a_{ij} = 1(X_i > X_j) - 1(X_i < X_j)$ and

$$b_{ij} = \Pr\left(Y_i > Y_j | Z_i, Z_j, \delta_i, \delta_j; \bar{G}_n\right) - \Pr\left(Y_i < Y_j | Z_i, Z_j, \delta_i, \delta_j; \bar{G}_n\right),$$

with $\bar{G}_n(y)$ representing a Kaplan–Meier type estimate of $\bar{G}(y) = \Pr(Y > y)$.

In equation (4.3), b_{ij} reduces to $1(Z_i > Z_j) - 1(Z_i < Z_j)$ whenever $\min(Z_i, Z_j)$ is observed and, in particular, when $\delta_i = \delta_j = 1$. Note, however, that the information contained in X_i and X_j is totally disregarded when calculating b_{ij}.

4.2.3 The estimator of Weier and Basu (1980)

Another estimator was proposed by Weier & Basu (1980). In their approach,

$$\hat{Y}_i = Z_i + 1(\delta_i = 0) \times \mathrm{MRL}(Z_i), \tag{4.4}$$

where $\mathrm{MRL}(Z_i)$ stands for the mean residual life function evaluated at Z_i, viz.

$$\mathrm{MRL}(Z_i) = \mathrm{E}_{\bar{G}_n}(Y - Z_i | Y > Z_i) = \int_{Z_i}^{\infty} \bar{G}_n(t)dt \Big/ \bar{G}_n(Z_i).$$

Here again, $\bar{G}_n$ is usually taken to be a Kaplan–Meier estimator. Once the imputation step has been completed, Weier & Basu (1980) recommend that τ be estimated through formula (4.1), applied to the pairs $(X_i, \hat{Y}_i)$. Note, however, that this procedure also ignores information on X, making it inefficient when $\tau \neq 0$.

4.2.4 The estimator of Wang and Wells (2000)

An estimator of τ that accounts for information on both X and Y was first developed in the literature by Wang & Wells (2000a). Considering that the population value of Kendall's tau can be computed by

$$\tau = 4 \int_0^{\infty} \int_0^{\infty} \bar{H}(x,y) d\bar{H}(x,y) - 1,$$

where $\bar{H}(x,y) = \Pr(X > x, Y > y)$, these authors suggest to use

$$\bar{\tau}_n = 4 \sum_{i=1}^{n} \sum_{j=1}^{n} \bar{H}_n(X_{(i)}, Z_{(j)}) \bar{H}_n(\Delta X_{(i)}, \Delta Z_{(j)}) - 1.$$

Here, $\bar{H}_n$ is any estimator of the bivariate survival function $\bar{H}$, while $X_{(1)} < \cdots < X_{(n)}$ and $Z_{(1)} < \cdots < Z_{(n)}$ are the ordered statistics of the X and Z samples.

As for $\bar{H}_n(\Delta X_{(i)}, \Delta Z_{(j)})$, it stands for

$$\bar{H}_n(X_{(i)}, Z_{(j)}) - \bar{H}_n(X_{(i-1)}, Z_{(j)}) - \bar{H}_n(X_{(i)}, Z_{(j-1)}) + \bar{H}_n(X_{(i-1)}, Z_{(j-1)}),$$

with $X_{(0)} = Z_{(0)} = 0$ by convention. In their paper, Wang & Wells (2000a) used the estimator of Dabrowska (1988) for $\bar{H}_n$, so that in their case, $\bar{H}_n(0, Z_{(j)})$ was the Kaplan–Meier estimator on Z at $Z_{(j)}$, and $\bar{H}_n(X_{(i)}, 0)$ was the empirical survival function of X evaluated at $X_{(i)}$.

4.2.5 The estimator of Lim (2006)

Following Diaconis et al. (2001), Lim (2006) introduces a permutation procedure for the estimation of τ from doubly censored data. To see how it can be adapted to the present context, it is first necessary to define what is meant by a "configuration" of the data. Here, an example will come in handy. Suppose that $n = 5$ pairs of points have been observed, as per Table 4.1. These data are also illustrated in Fig. 4.1, in which an upward-pointing arrow symbolizes a censored point.

TAB. 4.1 – A fictitious data set.

i	1	2	3	4	5
X_i	7.9	38.7	23.3	15.0	31.7
Z_i	3.1	2.0	1.1	4.5	2.5
δ_i	0	1	1	1	0

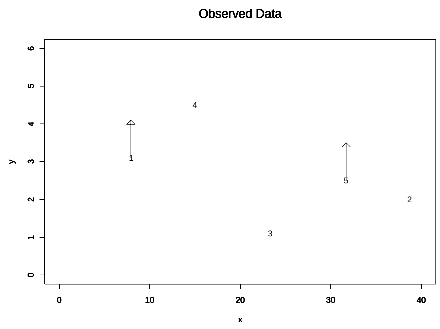


FIG. 4.1 – A Cartesian representation of the fictitious data set from Table 4.1.

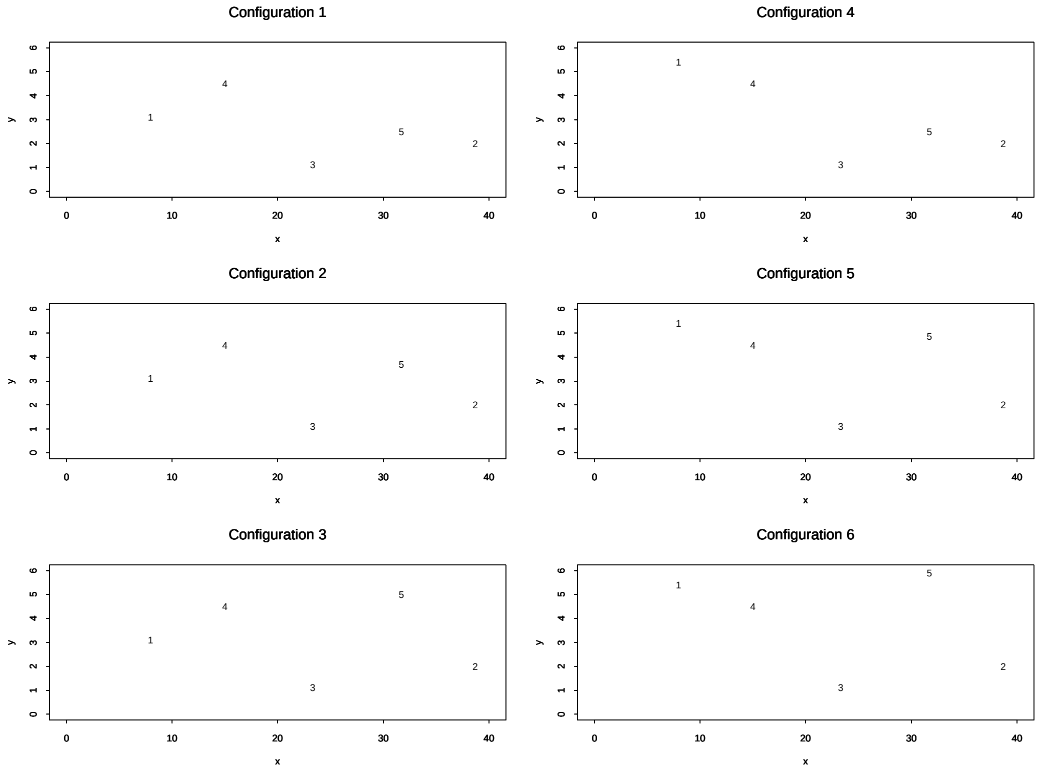


FIG. 4.2 – Possible configurations of the fictitious data set from Table 4.1.

If the two censored values of Y were revealed, six possible configurations of the data could ensue. These are depicted in Fig. 4.2. It is easy to see that, in general, the total number of possible configurations is given by

$$M_n = \prod_{i=1}^{n} \left\{ \delta_{(i)} + (1 - \delta_{(i)})(n - i + 1) \right\},$$

where $\delta_{(i)}$ is the indicator variable of the ordered statistic $Z_{(i)}$ for $i \in \{1, \ldots, n\}$. In other words, if there are c censored values of Y, then $M_n = G_1 \times \cdots \times G_c$, where G_i is the number of values of Z that are greater or equal to the ith censored Y.

A natural estimator of τ is then given by

$$\hat{\tau}_{\text{all}} = \frac{1}{M_n} \sum_{i=1}^{M_n} \hat{\tau}_i,$$

where $\hat{\tau}_i$ is the empirical version of Kendall's tau computed from (4.1), based on the ith configuration of the data. However, M_n can get extremely large, thus making it impossible computationally to get $\hat{\tau}_{\text{all}}$. As a compromise, one may use

$$\hat{\tau}_{\text{permut}} = \frac{1}{B} \sum_{i=1}^{B} \hat{\tau}_i,$$

where each summand is obtained from a randomly chosen configuration, and B is suitably large. In the simulations reported below, $B = 2{,}500$ was used.

4.2.6 Conditions for consistency

It was mentioned in Section 4.2.1 that the estimator of Oakes (1982) is inconsistent when $\tau \neq 0$. The same holds true for the estimators of Brown et al. (1974) and Weier & Basu (1980). In fact, Wang & Wells (2000a) show that their bias increases with the value of $\tau > 0$.

In contrast, the estimators of Wang & Wells (2000a) and Oakes (2006) can be consistent even when $\tau \neq 0$. While this was established under a specific model in the case of Oakes' renormalized estimator (ORE), Wang & Wells (2000a) invoke conditions on the censoring and sampling scheme to guarantee that their method yields a consistent solution. When only Y is subject to censoring, these conditions can be expressed as follows.

Lemma (Wang and Wells). *Let G and V be the cumulative distribution functions of Y and $Z = \min(Y, C)$, respectively. If $\{y : 1 - G(y) > 0\} = \{z : 1 - V(z) > 0\}$, then τ is estimable without bias using the method of Wang & Wells (2000a).*

In the simulation study described in Section 4.5, both survival and censoring times follow continuous distributions whose support is $(0, \infty)$. In these circumstances, the above lemma thus ensures that τ is estimable without bias.

4.3 New estimators

Four new estimators of τ are proposed below. All of them exploit information present in the conditional distribution of Y given X when the former is censored.

4.3.1 WePa : An extension of the estimator due to Oakes (1982)

Given that in the absence of censoring, one has $w_{ij} = 2\times 1\{(X_i - X_j)(Y_i - Y_j) > 0\} - 1$, the approach of Oakes (1982) amounts to estimating

$$\Pr\{(X_i - X_j)(Y_i - Y_j) > 0\}$$

by 1/2 whenever the concordance or discordance of the pairs (X_i, Y_i), (X_j, Y_j) cannot be determined. In practice, however, it is sometimes possible to get a better estimate of this probability, using information on the joint distribution of X and Y.

A natural extension of Oakes' method would thus consist of imputing a value of w_{ij} in (4.1) using an estimate of the probability of concordance of any such pair, conditional on

$$\mathcal{I}_n(i,j) = \{X_i, X_j, Z_i, Z_j, \delta_i, \delta_j; H_n(\cdot|X_i), H_n(\cdot|X_j)\},$$

where $H_n(y|x)$ is an estimate of $\Pr(Y \leq y | X = x)$. In other words, one would set $w_{ij} = 2P_n(i,j) - 1$, where

$$P_n(i,j) = \widehat{\Pr}\{(X_i - X_j)(Y_i - Y_j) > 0 | \mathcal{I}_n(i,j)\}.$$

The resulting estimator for τ is called the *Weighted Pair Estimator* (WePa) in the sequel.

In practice, the computation of $P_n(i,j)$ poses no difficulty when $\delta_k = 1$, with k defined as in (4.2). For, one then has $w_{ij} = \pm 1$. When $\delta_k = 0$, and assuming $Z_i = z_i < Z_j = z_j$ without loss of generality, there are four cases to consider :

a) If $\delta_j = 1$ and $X_i < X_j$, then

$$\begin{aligned} P_n(i,j) &= \widehat{\Pr}(Y_i \leq Y_j | Y_i > z_i, Y_j = z_j, X_i = x_i, X_j = x_j) \\ &= \widehat{\Pr}(Y_i \leq z_j | Y_i > z_i, X_i = x_i, X_j = x_j) \\ &= \frac{\widehat{\Pr}(z_i < Y_i \leq z_j | X_i = x_i)}{1 - \widehat{\Pr}(Y_i \leq z_i | X_i = x_i)} = \frac{H_n(z_j|x_i) - H_n(z_i|x_i)}{1 - H_n(z_i|x_i)}. \end{aligned}$$

b) If $\delta_j = 1$ and $X_i > X_j$, then

$$\begin{aligned} P_n(i,j) &= \widehat{\Pr}(Y_i > Y_j | Y_i > z_i, Y_j = z_j, X_i = x_i, X_j = x_j) \\ &= \frac{\widehat{\Pr}(Y_i > z_i \vee z_j | X_i = x_i)}{\widehat{\Pr}(Y_i > z_i | X_i = x_i)} = \frac{1 - H_n(z_j|x_i)}{1 - H_n(z_i|x_i)}, \end{aligned}$$

where in general $s \vee t = \max(s,t)$.

c) If $\delta_j = 0$ and $X_i < X_j$, then upon conditioning on the (unobserved) value of Y_i, one gets

$$\begin{aligned} P_n(i,j) &= \int_{z_i}^{\infty} h_n(t|Y_i > z_i, X_i = x_i) \widehat{\Pr}(Y_j > t | Y_j > z_j, X_j = x_j) dt \\ &= \frac{\int_{z_i}^{\infty} h_n(t|X_i = x_i) \widehat{\Pr}(Y_j > t \vee z_j | X_j = x_j) dt}{\widehat{\Pr}(Y_i > z_i | X_i = x_i) \widehat{\Pr}(Y_j > z_j | X_j = x_j)} \\ &= \frac{\int_{z_i}^{\infty} h_n(t|x_i)\{1 - H_n(t \vee z_j | x_j)\} dt}{\{1 - H_n(z_i|x_i)\}\{1 - H_n(z_j|x_j)\}}, \end{aligned}$$

where h_n is an estimate of the density associated with H_n.

d) If $\delta_j = 0$ and $X_i > X_j$, the same conditioning yields

$$\begin{aligned} P_n(i,j) &= \int_{z_j}^{\infty} h_n(t|Y_i > z_i, X_i = x_i)\ \widehat{\Pr}(Y_j \le t | Y_j > z_j, X_j = x_j) dt \\ &= \frac{\int_{z_j}^{\infty} h_n(t|X_i = x_i) \widehat{\Pr}(z_j < Y_j \le t | X_j = x_j) dt}{\widehat{\Pr}(Y_i > z_i | X_i = x_i) \widehat{\Pr}(Y_j > z_j | X_j = x_j)} \\ &= \frac{\int_{z_j}^{\infty} h_n(t|x_i)\{H_n(t|x_j) - H_n(z_j|x_j)\} dt}{\{1 - H_n(z_i|x_i)\}\{1 - H_n(z_j|x_j)\}}. \end{aligned}$$

4.3.2 CoBra : An extension of the estimator due to Brown et al. (1974)

A similar strategy leads to a natural extension of the estimator of Brown et al. (1974), to be called the *Conditional Brown et al. Estimator* (CoBra). To account for information given by the covariate X, one could replace b_{ij} in formula (4.3) by

$$\hat{b}_{ij} = \widehat{\Pr}\left\{Y_i > Y_j | \mathcal{I}_n(i,j)\right\} - \widehat{\Pr}\left\{Y_i < Y_j | \mathcal{I}_n(i,j)\right\},$$

so that $a_{ij}\hat{b}_{ij} = P_n(i,j)$. Note that if $\hat{b}_{ij} = \pm 1$ for all $i \neq j$, then

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \hat{b}_{ij}^2 = n(n-1),$$

in which case the estimators (4.1) and (4.3) coincide.

4.3.3 CoWeBa : An extension of the estimator due to Weier and Basu (1980)

The estimator of Weier & Basu (1980) is based on formula (4.1), in which each censored value Y_i is imputed using (4.4). To incorporate conditional information from X into this procedure, a simple way to proceed is to compute

$$\begin{aligned} \hat{Y}_i &= Z_i + 1(\delta_i = 0) \times \mathrm{MRL}(Z_i|X_i) = Z_i + 1(\delta_i = 0) \times \mathrm{E}_{H_n}\left(Y - Z_i | Y > Z_i, X_i\right) \\ &= Z_i + \frac{1(\delta_i = 0)}{1 - H_n(Z_i|X_i)} \int_{Z_i}^{\infty} \{1 - H_n(t|X_i)\} dt. \end{aligned}$$

This estimator is referred to henceforth as the *Conditional Weier–Basu Estimator* (CoWeBa).

4.3.4 Icdf : A simulation-based estimator

Formula (4.4) and its conditional extension described above are not the only way in which values could be imputed for each censored Y_i. A natural alternative would be as follows :

a) For each $i \in \{1, \ldots, n\}$ such that $\delta_i = 0$, use the conditional distribution of Y given $X_i = x_i$ and $Y_i > c_i$ to generate a random observation $\tilde{Y}_i$ (which is then superior to Z_i).

b) Use formula (4.1) with the completed set $(X_1, \hat{Y}_1), \ldots, (X_n, \hat{Y}_n)$, where $\hat{Y}_i = Z_i \times \delta_i + \tilde{Y}_i(1 - \delta_i)$;

c) As the estimate τ_ℓ depends on the data generated, repeat steps a) and b) a total of NSIM times, and take the average of the resulting estimates. In the simulations reported below, NSIM = 3000 was sufficient to insure the numerical stability of the *Inverted CDF Estimator* (Icdf), up to three decimals.

4.4 Estimators of $H(y|x)$

In order to calculate the estimators described in Section 4.3, an estimate $H_n(y|x)$ of $H(y|x) = \Pr(Y \leq y|X = x)$ is required. Two of them are considered here, which were introduced respectively by Leconte et al. (2002) and by Van Keilegom & Akritas (1999). These estimators, which are briefly described below for completeness, are consistent under the sampling conditions of Section 4.2.6.

4.4.1 LPT : The estimator of Leconte et al. (2002)

This estimator is a smoothed version of the generalized Kaplan–Meier estimator of Beran (1981). It is continuous both in the response variable and in the covariate.

Given a kernel function K with suitable bandwidth w, Beran's estimator is defined by

$$\bar{H}_{GKM}(y|x) = \begin{cases} \displaystyle\prod_{i=1}^{n} \left\{ 1 - \frac{B_i(x)}{\displaystyle\sum_{r=1}^{n} 1(Z_r \geq Z_i) B_r(x)} \right\}^{1(Z_i \leq y, \delta_i = 1)} & \text{if } y < Z_{(n)}; \\ 0 & \text{otherwise;} \end{cases}$$

where

$$B_i(x) = \frac{K\left(\dfrac{x - X_i}{w}\right)}{\displaystyle\sum_{j=1}^{n} K\left(\dfrac{x - X_j}{w}\right)}$$

is a Nadaraya–Watson type weight function, as in Dabrowska (1989).

While this estimator is continuous in the covariate X, it has jumps at each uncensored value of Z. In order to make it smooth in that variable also, Leconte et al. (2002) propose to write

$$H_{LPT}(y|x) = \int_0^\infty L\left(\frac{y - t}{\omega}\right) d\bar{H}_{GKM}(t|x), \tag{4.5}$$

where ω is another bandwidth and

$$L(t) = \int_{-\infty}^{t} K(u) du, \quad t \in \mathbb{R}.$$

Formula (4.5) may be written more explicitly as

$$H_{LPT}(y|x) = \sum_{i=1}^{I+1} \left\{ \bar{H}_{GKM}\left(Y^+_{(i-1)}|x\right) - \bar{H}_{GKM}\left(Y^+_{(i)}|x\right) \right\} L\left(\frac{y - Y^+_{(i)}}{\omega}\right),$$

where $I = \delta_1 + \cdots + \delta_n$ is the number of uncensored observations, $Y^+_{(0)} = 0$, $Y^+_{(i)}$ is the ith smallest uncensored value of Z, $i \in \{1, \ldots, I\}$, and $Y^+_{(I+1)}$ is the largest value of Z, whether censored or not.

4.4.2 VKA : The estimator of Van Keilegom and Akritas (1999)

This estimator is of the form

$$H_{VKA}(y|x) = 1 - \bar{H}_e \left\{ \frac{y - \hat{m}(x)}{\hat{\sigma}(x)} \right\},$$

where $\bar{H}_e(t)$ denotes the Kaplan–Meier estimator based on standardized versions of the Z values. The latter are given by

$$E_i = \frac{Z_i - \hat{m}(X_i)}{\hat{\sigma}(X_i)},$$

where $\hat{m}(x)$ and $\hat{\sigma}^2(x)$ are suitable estimators of $\mathrm{E}(Y|X = x)$ and $\mathrm{var}(Y|X = x)$, respectively.

Given a normalized score function J on $[0, 1]$, Van Keilegom & Akritas (1999) suggest the use of

$$\hat{m}(x) = \int_0^1 H_{GKM}^{-1}(s|x)J(s)ds$$

and

$$\hat{\sigma}^2(x) = \int_0^1 H_{GKM}^{-1}(s|x)^2 J(s)ds - \hat{m}^2(x),$$

with $H_{GKM}(t) = 1 - \bar{H}_{GKM}(t)$. In the sequel, J is taken to be

$$J(s) = \frac{1}{b-a} 1(a \leq s \leq b),$$

where $a = 0$ and

$$b = \min_{1 \leq i \leq n} H_{GKM}(+\infty|X_i) = \min_{1 \leq i \leq n} H_{GKM}(y^+_{(I)}|X_i).$$

As mentioned by Van Keilegom et al. (2001), the resulting local estimates of $m(x)$ and $\sigma^2(x)$ can then be expected to perform well, even under heavy censoring. While their claim was made under the assumption that X and Y are related via a nonparametric regression model

$$Y = m(X) + \sigma(X)\,\epsilon, \tag{4.6}$$

a similar empirical observation will be made in the simulation study that follows, despite the fact that this particular structure does not hold for the dependence models considered.

4.5 Simulations

This section reports the results of a simulation study carried out to investigate the relative performance of the various estimators of tau described in Sections 4.2 and 4.3. Details related to the implementation of the various estimators are given in Section 4.5.1, and the simulation design is described in Section 4.5.2. The simulation results per se are discussed in Section 4.5.3.

4.5.1 Implementation

The version of Leconte et al.'s estimator considered in the simulations is based on the tri-weight kernel function

$$K(x) = \frac{35}{32}(1-x^2)^3, \quad -1 \leq x \leq 1. \tag{4.7}$$

Following Leconte et al. (2002), the bivariate scale rule was used in selecting bandwidths

$$(w, \omega) = \left(2.978\hat{\sigma}_x n^{-1/6}, 2.978\hat{\sigma}_y n_+^{-1/6}\right).$$

Here, n_+ is the number of uncensored values of Z, $\hat{\sigma}_x^2$ is the sample variance computed with the values of X, and

$$\hat{\sigma}_y^2 = \int_0^\infty \left(t - \bar{Y}^+\right) d\bar{H}_n(t) = \sum_{i=1}^{I} s_{(i)} \left(Y_{(i)}^+ - \bar{Y}^+\right)^2,$$

where $s_{(i)}$ is the size of the jump at $Y_{(i)}^+$ of the standard Kaplan–Meier estimator $\bar{H}_n$ based on the values of Z, and

$$\bar{Y}^+ = \int_0^\infty t d\bar{H}_n(t) = \sum_{i=1}^{I} s_{(i)} Y_{(i)}^+.$$

For the estimator of Van Keilegom & Akritas (1999), the function J was taken to be uniform on the interval $(0, 1)$. This choice does not cause any difficulty, so long as the largest value of Z is uncensored, because one then has $H_{GKM}(y_{(I)}^+|x_i) = 1$ for every $i \in \{1, \ldots, n\}$, thus making

$$b = \min_{1\leq i\leq n} H_{GKM}\left(Y_{(I)}^+|X_i\right) = 1.$$

However, if the largest value of Z is censored, then $H_{GKM}(Y^+_{(I)}|x_i) < 1$ for some i. In this case,

$$\int_0^1 H^{-1}_{GKM}\left(s|X_i\right) J(s) ds$$

is undefined on $(b, 1)$. When this happened, it was arbitrarily decided to set

$$H_{GKM}(Z_{(n)}|X_i) = 1,$$

no matter whether $Z_{(n)}$ was censored or not. In this fashion, one gets $b = 1$ in this case also. This choice is of critical importance because when τ is large (say 0.8), one might get

$$b = \min_{1 \le i \le n} H_{GKM}\left(Y^+_{(I)}|X_i\right) \approx 0$$

for the largest X_i, so that $\hat{m}(x)$ and $\hat{\sigma}^2(x)$ would then be undefined.

4.5.2 Simulation plan

Simulations were carried out for two sample sizes ($n = 100, 200$) and two censoring proportions (20% and 40%). The variables X and Y were assumed to have the same log-normal distribution with mean 30 and variance 50.

To introduce dependence between X and Y, bivariate frailty models were considered. These models, originally introduced by Oakes (1989), are a natural bivariate extension of Cox's proportional hazards model. For recent applications, see, e.g., Braekers & Veraverbeke (2005), Choi & Matthews (2005), Oakes (2005) or Vandenhende & Lambert (2005) in the special issue of *The Canadian Journal of Statistics* devoted to the DeMoSTAFI meeting.

As shown by Oakes (1989), the joint survival function of a pair (X, Y) whose association is induced by a frailty γ can be expressed in the form $\bar{H}(x, y) = C\{\bar{F}(x), \bar{G}(y)\}$, where C is an Archimedean copula. In other words, C is a distribution function with uniform margins on the interval $(0, 1)$ that may be expressed in the form $C(u, v) = \psi\{\psi^{-1}(u) + \psi^{-1}(v)\}$ in terms of the Laplace transform ψ of the underlying frailty γ and its inverse ψ^{-1}. See, e.g., Genest & MacKay (1986) or (Nelsen, 1999, Chapter 4) for descriptions of this broad class of dependence models.

Three specific frailty models were considered in the simulation. They correspond to the following choices of Archimedean copula C or distribution D for γ :

a) the Clayton family (in which D is a gamma distribution)

$$C_\alpha(u,v) = \left(u^{-\alpha} + v^{-\alpha} - 1\right)^{-1/\alpha}, \quad \alpha > 0;$$

b) the Frank family (in which D is a logarithmic series distribution on the integers)

$$F_\alpha(u,v) = -\frac{1}{\alpha} \ln \left\{ 1 + \frac{\left(e^{-\alpha u} - 1\right)\left(e^{-\alpha v} - 1\right)}{\left(e^{-\alpha} - 1\right)} \right\}, \quad \alpha > 0;$$

c) the Gumbel–Hougaard family (in which D is a positive stable distribution)

$$G_\alpha(u,v) = \exp\left\{ - \left(|\ln(u)|^\alpha + |\ln(v)|^\alpha\right)^{1/\alpha} \right\}, \quad \alpha > 1.$$

In all cases, the independence copula $\Pi(u,v) = uv$ corresponds to the limit as α approaches its lower bound. By varying the value of α in the specified range, all possible degrees of association are covered between $\tau = 0$ and $\tau = 1$. As shown by Genest & MacKay (1986),

$$\tau\left(C_\alpha\right) = \frac{\alpha}{\alpha+2}, \quad \tau\left(G_\alpha\right) = 1 - \frac{1}{\alpha}$$

with a more complicated expression involving the Debye function for $\tau(F_\alpha)$.

For each of the above models, three different degrees of dependence were considered, namely $\tau = 0.2, 0.5, 0.8$. Coupled with the two previously mentioned choices of sample size and censoring proportion, this led to a full factorial design with $36 = 3 \times 3 \times 2 \times 2$ combinations, each of which was repeated 1000 times. This allowed for an estimation of both the bias and mean-squared error of the following 21 estimators of τ :

1 : the estimator of Oakes (1982) ;
2 : the ORE estimator from Oakes (2006) ;
3 : the estimator of Brown et al. (1974) with a Kaplan–Meier estimate for $G(y)$;
4 : the estimator of Brown et al. (1974) with the true marginal distribution G ;
5 : the estimator of Weier & Basu (1980) with a Kaplan–Meier estimate for G ;
6 : the estimator of Weier & Basu (1980) with the true G ;
7 : the estimator of Wang & Wells (2000a) with Dabrowska's estimator ;
8 : the WePa estimator with $H_{LPT}(y|x)$;
9 : the WePa estimator with $H_{VKA}(y|x)$;
10 : the WePa estimator with the true conditional distribution $H(y|x)$;
11 : the CoWeBa estimator with $H_{LPT}(y|x)$;
12 : the CoWeBa estimator with $H_{VKA}(y|x)$;
13 : the CoWeBa estimator with the true $H(y|x)$;
14 : the Icdf estimator with $H_{LPT}(y|x)$;
15 : the Icdf estimator with $H_{VKA}(y|x)$;
16 : the Icdf estimator with the true $H(y|x)$;
17 : the Icdf estimator with the survival estimator of Dabrowska (1988) ;
18 : the estimator of Lim (2006) ;
19 : the CoBra estimator with $H_{LPT}(y|x)$;
20 : the CoBra estimator with $H_{VKA}(y|x)$;
21 : the CoBra estimator with the true $H(y|x)$.

Estimators 4, 6, 10, 13, 16 and 21 require a full knowledge either of the marginal distribution $G(y)$ or of the conditional distribution $H(y|x)$ for every observed value $X = x$. Although this is not realistic, these estimators were included in order to assess the impact of estimating $H(y|x)$ by $H_{LPT}(y|x)$ or $H_{VKA}(y|x)$, especially for large values of τ.

This assessment is made in the following section, which also contains comparisons between those of the above estimators that could be used in practice, i.e., without any prior knowledge of either $G(y)$ or $H(y|x)$. Note, however, that in view of their poor performance, the CoBra estimators (19–21) are omitted in the sequel.

4.5.3 Results

Figs. 4.3–4.4 feature boxplots showing the dispersion of $R = 1000$ estimates of τ derived from samples of size $n = 100$ using estimators 1 to 18 in the above list. The figures correspond to 20% and 40% censoring, for which the percentage of indeterminate pairs averaged 8% and 32%, respectively. In each figure, the top and bottom panels correspond to values of $\tau = 0.5, 0.8$, respectively ; because of negligible differences between the estimators when $\tau = 0.2$, those plots are omitted. Results for the Clayton, Frank and Gumbel–Hougaard copulas are presented, left, center and right.

An overall look at the 12 graphs leads to the following general observations :

a) for a fixed value of τ and a given level of censoring, the general behavior of each of the 18 estimators is approximately the same from copula to copula ;

b) given the copula model and the level of censoring, the differences among the estimators become more accentuated as τ increases ;

c) for a fixed value of τ and a given copula model, the bias and dispersion of the 18 estimators become greater in absolute value as the proportion of censored observations goes from 20% to 40% ;

d) the performance of estimators WePa, CoWeBa and Icdf is not significantly improved when the estimator of $H(y|x)$ is replaced with the true distribution. This observation is also valid, mutatis mutandis, for the estimators of Brown et al. (1974) and Weier and Basu (1980).

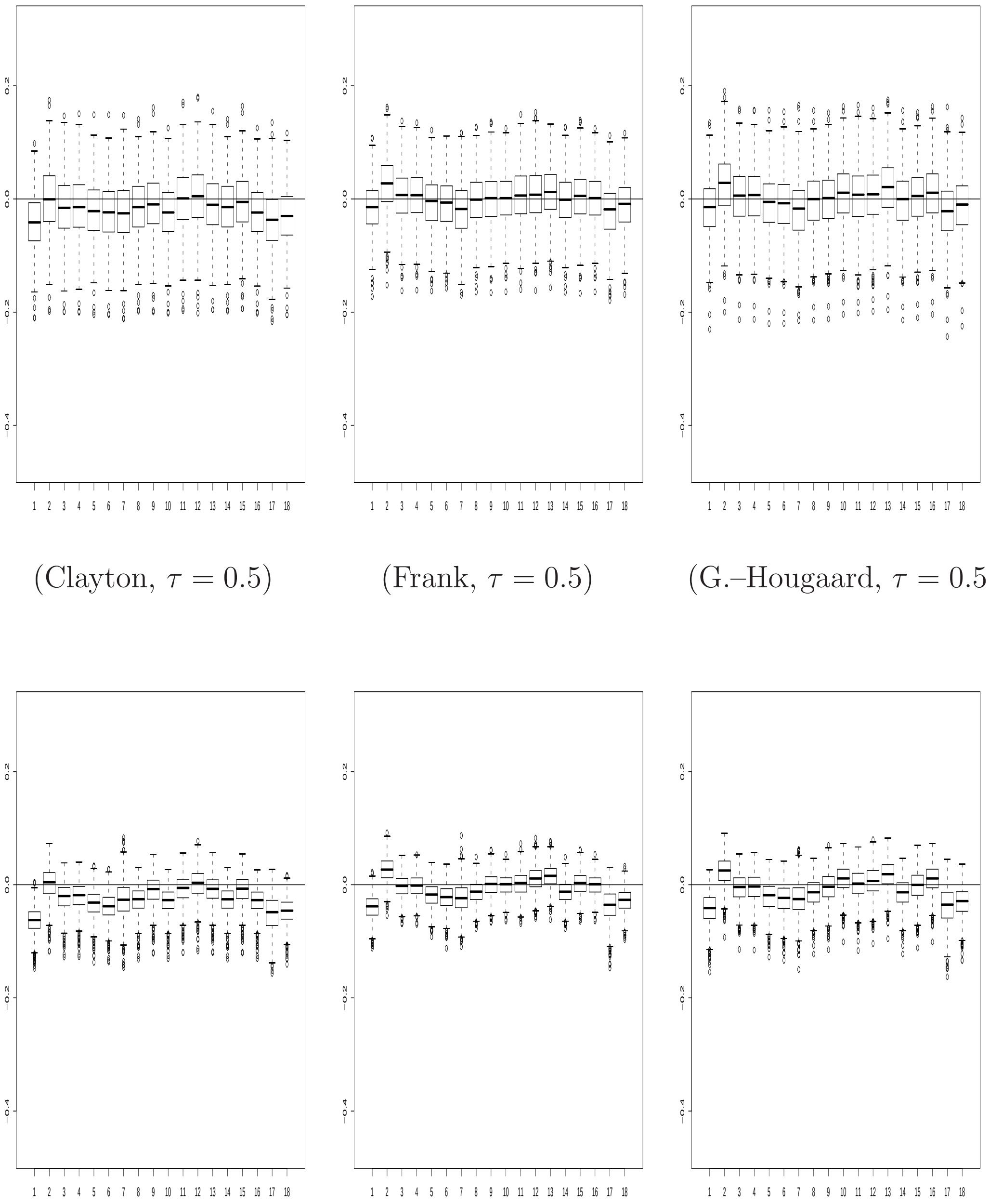
To refine these impressions, an analysis of variance was carried out on square-root absolute bias (SRAB) and the logarithm of the mean-squared error (LMSE), viz.

$$\text{SRAB} = \left| \frac{1}{R} \sum_{i=1}^{R} (\hat{\tau}_i - \tau) \right|^{1/2} \quad \text{and} \quad \text{LMSE} = \log \left\{ \frac{1}{R} \sum_{i=1}^{R} (\hat{\tau}_i - \tau)^2 \right\}.$$

Both response variables turned out to be approximately homoscedastic and normally distributed. They were analyzed separately for each copula model, as a function of sample size n, level of censoring C, level of dependence τ, and estimation method. *Oakes' renormalized estimator* (ORE) with $n = 100$, $C = 20\%$ and $\tau = 0.2$ was used as a reference point in all comparisons.

Because the purpose of the analysis was to determine which estimators perform best over all conditions, interactions involving estimation methods were removed at the outset from the models for SRAB and LMSE. Methods that required knowledge of distributions were also excluded to facilitate comparisons between estimators of practical use. A backward selection procedure was then called upon to remove main effects and interactions that were non-significant at the 1% level.

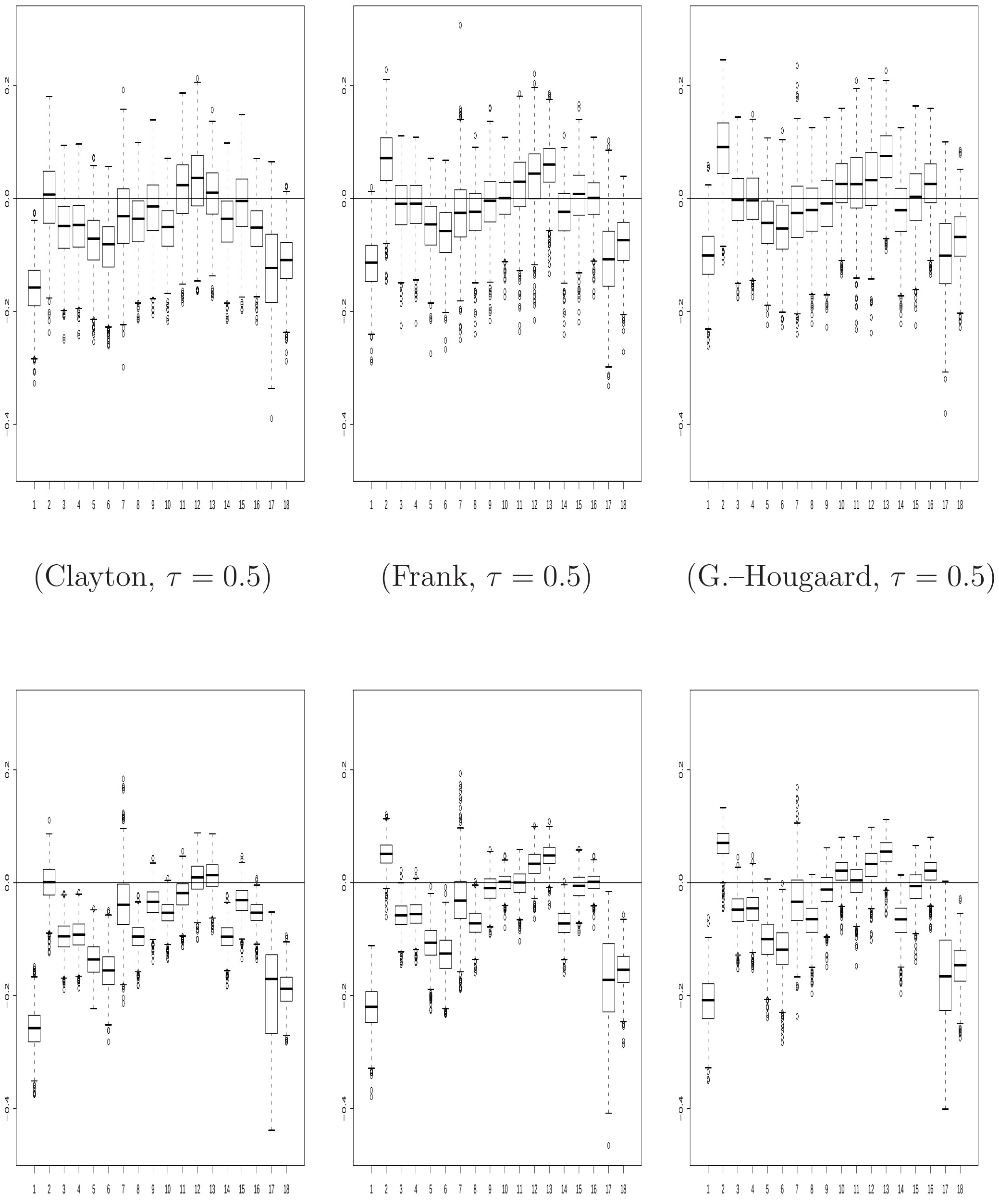
FIG. 4.3 – Comparison of 18 estimators of τ subject to $\mathbf{C} = \mathbf{20}\%$ censoring in 1000 samples of size $n = 100$ from distributions with different dependence structures and degrees of dependence. Both X and Y are assumed to be log-normal with mean 30 and variance 50.



(Clayton, $\tau = 0.5$) (Frank, $\tau = 0.5$) (G.–Hougaard, $\tau = 0.5$)

(Clayton, $\tau = 0.8$) (Frank, $\tau = 0.8$) (G.–Hougaard, $\tau = 0.8$)

Fig. 4.4 – Comparison of 18 estimators of τ subject to $\mathbf{C} = \mathbf{40}\%$ censoring in 1000 samples of size $n = 100$ from distributions with different dependence structures and degrees of dependence. Both X and Y are assumed to be log-normal with mean 30 and variance 50.



(Clayton, $\tau = 0.5$) (Frank, $\tau = 0.5$) (G.–Hougaard, $\tau = 0.5$)

(Clayton, $\tau = 0.8$) (Frank, $\tau = 0.8$) (G.–Hougaard, $\tau = 0.8$)

In both models, the 1% significance criterion suggested that factors C and τ should be retained, along with the interaction associated with $(C = 40\%) \times (\tau = 0.8)$. In the case of LMSE, the effect of n is also significant. The final models provide a comparison, either in terms of SRAB or LMSE, between the ORE (2) and the various other estimation methods, after adjustment for the different design factors.

Variable	Clayton	Frank	Gumbel–Hougaard
Intercept	0.001	0.080	0.133
Standard error (intercept)	*0.010*	*0.010*	*0.010*
Method 1	0.204	0.076	
Method 3	0.084	**−0.069**	**−0.122**
Method 5	0.117		−0.058
Method 7	0.053		−0.081
Method 8	0.068	−0.046	−0.103
Method 9		**−0.108**	**−0.145**
Method 11		**−0.068**	**−0.128**
Method 12			−0.091
Method 14	0.068	−0.046	−0.103
Method 15		**−0.092**	**−0.153**
Method 17	0.180	0.079	
Method 18	0.159		
Standard error (method)	*0.016*	*0.017*	*0.016*
$C = 40\%$	0.077	0.078	0.073
Standard error (C)	*0.010*	*0.010*	*0.009*
$\tau = 0.5$	0.050	0.041	0.041
$\tau = 0.8$	0.068	0.055	0.062
Average standard error (τ)	*0.012*	*0.012*	*0.011*
$(C = 40\%) \times (\tau = 0.8)$	0.045	0.050	0.049
Standard error (interaction)	*0.017*	*0.018*	*0.016*

TAB. 4.2 – Parameter estimates for the regression of SRAB on censorship, level of dependence and estimation methods not involving knowledge of the conditional distribution of Y given X. For each copula model, the four methods which provide the most significant improvement with respect to Oakes' renormalized estimator (ORE) are highlighted in boldface.

Table 4.2 reports parameter estimates and associated standard errors for the anova model on SRAB. Methods whose coefficient is negative (resp. positive) have lower (resp. higher) SRAB than ORE. Empty cells indicate that in terms of bias, the corresponding methods are not significantly different from ORE at the 1% level. For each scenario, the four best alternatives to ORE are highlighted in boldface.

The figures in Table 4.2 show clearly that the SRAB tends to increase with the level of dependence or censoring in the data. They also indicate that the SRAB associated with the ORE is smallest under Clayton's model. This is not surprising, given that this method is consistent in that case (Oakes, 2006). Nevertheless, the SRAB associated with estimators 9, 11, 12 and 15 is not significantly different from that of the ORE under this model.

For Frank or Gumbel–Hougaard dependence structures, Oakes (2006) shows that the ORE is inconsistent. This is in line with the simulation results, which suggest that most methods deliver better SRAB than the ORE. In particular, estimator 3 of Brown et al. (1974) and new methods 9, 11 and 15 give the best results.

More generally, Table 4.2 makes it possible to measure the impact of various estimators or levels of dependence and censorship. For example, method 9 performed best under Frank's model. When $C = 20\%$ and $\tau = 0.2$, its absolute bias is smaller than that of ORE by a factor of $(0.080)^2/(0.080 - 0.108)^2 \approx 8.2$ on average. Similarly, method 15 is best for the Gumbel–Hougaard copula and yields an average improvement of 44.2

			τ		
Model	C	Method	0.2	0.5	0.8
Clayton	20%	9, 15	0.05	3.25	3.72
	40%	9, 15	4.90	14.40	33.49
Frank	20%	9	0.78	0.17	0.73
		15	0.14	0.84	1.85
	40%	9	2.50	8.28	24.03
		15	4.36	11.45	29.24
Gumbel–Hougaard	20%	9	0.14	0.84	2.50
		15	0.40	0.44	1.76
	40%	9	3.72	10.40	29.58
		15	2.81	8.84	26.90

TAB. 4.3 – 1000 × |bias| for estimation methods 9 (WePa-VKA) and 15 (Icdf-VKA) under different simulation conditions based on the parameter estimates of the regression.

under the same conditions.

Table 4.3 provides an additional comparison between methods 9 and 15 (WePa-VKA and Icdf-VKA), which exhibit the lowest SRAB under the three models considered. Reported there is 1000 × |bias| for both methods, as estimated by the anova model under the different levels of censoring and dependence studied. The results show that |bias| tends to increase with the amount of dependence or censorship. They also confirm that method 9 is generally superior to method 15 for Frank's model, and vice versa for the Gumbel–Hougaard.

Variable	Clayton	Frank	Gumbel–Hougaard
Intercept	−5.930	−5.841	−5.776
Standard error (intercept)	*0.107*	*0.115*	*0.100*
Method 1	1.336	0.855	0.809
Method 3			
Method 5	0.595		
Methods 7–8			
Method 9		**−0.554**	
Methods 11, 12, 14			
Method 15		**−0.563**	
Method 17	1.235	0.938	0.913
Method 18	0.935		
Standard error (method)	*0.193*	*0.208*	*0.185*
$C = 40\%$	0.594	0.603	0.519
Standard error (C)	*0.125*	*0.134*	*0.120*
$n = 200$	−0.536	−0.599	−0.552
Standard error (n)	*0.102*	*0.109*	*0.098*
$\tau = 0.8$	−0.918	−1.343	−1.182
Standard error (τ)	*0.153*	*0.164*	*0.147*
$(C = 40\%) \times (\tau = 0.8)$	0.984	1.287	1.160
Standard error (interaction)	*0.216*	*0.232*	*0.208*

TAB. 4.4 – Parameter estimates for the regression of LMSE on sample size, censorship, level of dependence and estimation methods not involving knowledge of the conditional distribution of Y given X. The only two significant improvements with respect to Oakes' renormalized estimator (ORE) are highlighted in boldface.

Given in Table 4.4 are the parameter estimates and the associated standard errors for the model pertaining to LMSE. For the Clayton and Gumbel–Hougaard models, no estimator provides a significant improvement over the ORE. For Frank's dependence structure, however, methods 9 and 15 do perform better. The MSE is then reduced by an approximate factor of $e^{-5.841}/e^{-5.841-0.56} \approx 1.75$ in both cases.

One can also see from Table 4.4 that average savings of $1 - 1/1.709 = 41.5\%$ on the MSE obtain when the sample size goes from $n = 100$ to $n = 200$ under Clayton's model. The improvement is slightly better in the two other cases. This is in contrast with the absence of effect of n in the case of SRAB.

				τ		
Model	n	C	Method	0.2	0.5	0.8
Clayton	100	20%	9, 15	2.66	2.66	1.06
		40%	9, 15	4.82	4.82	5.14
	200	20%	9, 15	1.56	1.56	0.62
		40%	9, 15	2.82	2.82	3.01
Frank	100	20%	9	1.67	1.67	0.44
			15	1.65	1.65	0.43
		40%	9	3.05	3.05	2.89
			15	3.02	3.02	2.86
	200	20%	9	0.92	0.92	0.24
			15	0.91	0.91	0.24
		40%	9	1.68	1.68	1.59
			15	1.66	1.66	1.57
Gumbel–Hougaard	100	20%	9, 15	3.10	3.10	0.95
		40%	9, 15	5.21	5.21	5.10
	200	20%	9, 15	1.79	1.79	0.55
		40%	9, 15	3.00	3.00	2.94

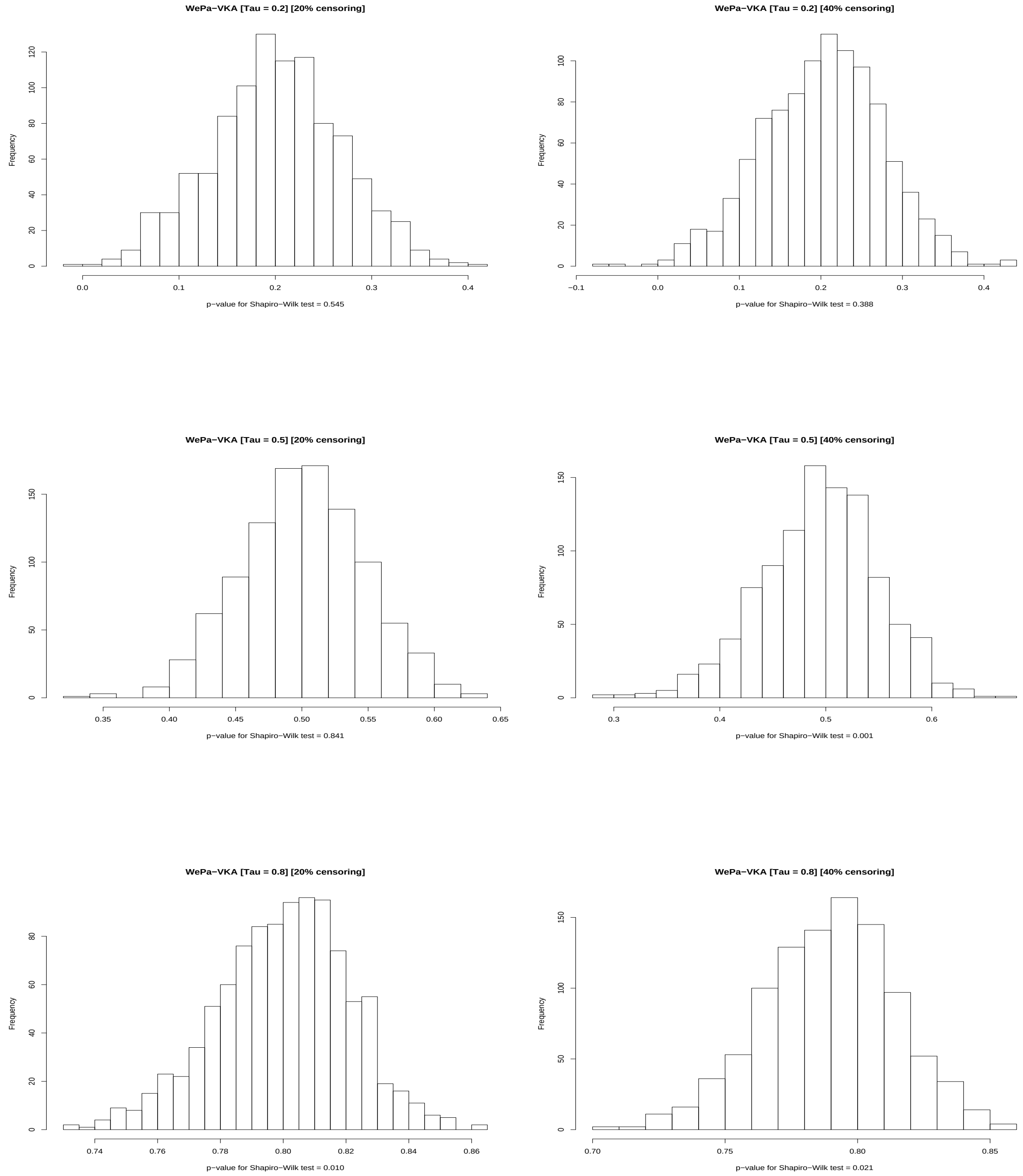
Tab. 4.5 – 1000 × MSE for estimation methods 9 (WePa-VKA) and 15 (Icdf-VKA) under different simulation conditions based on the parameter estimates of the regression
.

As for the combined effect of C and τ, it is more delicate to interpret than in the case of SRAB. When $n = 100$ and $C = 20\%$, the MSE tends to get smaller as dependence increases, but the effect is only significant (by a factor of 2.50) when $\tau = 0.8$. Because the bias is known to swell in this case, this phenomenon results from a sharp reduction in the variance which is likely due to the fact that τ is bounded. When censoring increases from 20% to 40%, the MSE increases by 77% on average when $\tau = 0.2$. This effect is approximately the same across copulas.

As a complement to Table 4.3, the values of 1000 × MSE are reported in Table 4.5 for methods 9 and 15 (WePa-VKA and Icdf-VKA), based on the final model for LMSE. As can be seen, the estimates are exactly the same for the two methods, both under the Clayton and the Gumbel–Hougaard model. The difference between the two is also negligible for Frank's dependence structure. This is in line with the results already reported in Table 4.4, where the main effects associated with these two estimators are practically equal for Frank's model, whereas they are null in the two other cases.

For inferential purposes, it is also of interest to examine the distribution across various scenarios, e.g., of the WePa–VKA estimator (9), which turned out to be one of the most successful among the new estimators. Displayed in Fig. 4.5 are histograms depicting the variation in the estimation of $\tau = 0.2, 0.5, 0.8$ for this method. They are based on 1000 random samples of size $n = 100$ from the Frank copula, subject either to 20% or to 40% censoring. As can be seen from this figure, the estimator is approximately normally distributed in all cases, except when $\tau = 0.8$ or $C = 40\%$ and $\tau = 0.5$. This is confirmed by the P-values of the Shapiro–Wilk test, using 5% as a cutoff point. Similar conclusions apply for other copulas and the two other main contenders, i.e., the CoWeBa–LPT (11) and the Icdf–VKA (15) estimators.

Fig. 4.5 – Histograms showing the dispersion of the WePa–VKA estimator (9) of tau, based on 1000 samples of size $n = 100$ from a Frank copula. The left and right columns correspond to $C = 20\%$ and $C = 40\%$, respectively. The theoretical value of τ is 0.2, 0.5 and 0.8 in the top, middle and bottom row, respectively.



4.6 Robustness issues

Because the proposed WePa, CoWeBa and Icdf estimators rely on information about the conditional distribution of Y given X, one may wonder to what extent their success can be affected by changes in the marginal distributions of the variables. One may also wish to check whether the specific choices of kernel and bandwidth used in the simulation could be responsible for the domination of estimators 9, 11 and 15, and for the relative poorer performance of variants 8, 12 and 14. Both of these robustness issues are considered below.

4.6.1 Robustness with respect to margins

Although Kendall's tau and existing estimators thereof are invariant to increasing transformations of X and Y, the newly proposed WePa $(8, 9)$, CoWeBa $(11, 12)$ and Icdf estimators $(14, 15)$ exploit information about the conditional distribution of Y given X. As the latter does not involve only the copula but also the marginal distributions of the variables, the simulation study described in Section 4.5 was reproduced for different choices of F and G to check whether the latter affect the performance of the new estimators.

Specifically, the Monte Carlo study of Section 4.5 was repeated $8 = 4 \times 2$ times, once for each possible combinations of $F \in \{\mathcal{L}_1, \mathcal{L}_2\}$ and $G \in \{\mathcal{L}_1, \mathcal{L}_2, \mathcal{G}_1, \mathcal{G}_2\}$, where for $i \in \{1, 2\}$, $\mathcal{L}_i$ is log-normal and $\mathcal{G}_i$ is Gamma with mean μ_i and variance σ_i^2. The values $\mu_1 = 30$, $\mu_2 = 20$, $\sigma_1^2 = 50$, $\sigma_2^2 = 10$ were used.

Following the same protocol as in Section 4.5, analysis of variance techniques were used to study the absolute bias and LMSE of estimators $8, 9, 11, 12, 14$ and 15 as a function of the sample size, the level of dependence and censoring, as well as the combination of marginal distributions. These response variables were chosen in order to meet as closely as possible the classical assumptions of homoscedasticity and normality for the error term. In both cases, the results indicated that neither the absolute bias nor the LMSE of the six estimators varied significantly as a function of the choice of margins. Neither this factor nor any of the interactions involving it turned out to be significant at the 1% level.

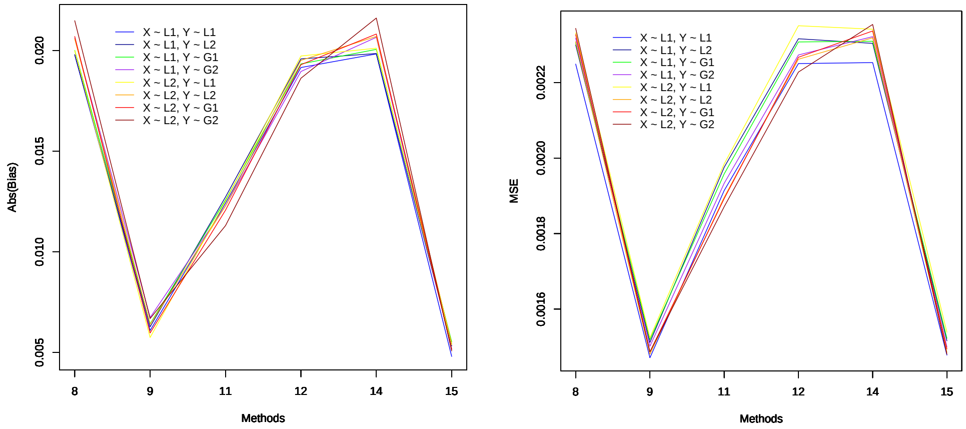


FIG. 4.6 – |Bias| and MSE of six estimators as a function of the choice of marginal distributions for X and Y.

Displayed in the left and right panels of Figure 4.6 are the absolute bias and MSE of the six estimators, averaged over all experimental conditions. It is clear from these graphs that the results reported in Section 4.5 are practically invariant to the selection of margins for X and Y in the given sets, except possibly for estimators 12 and 14. However, the earlier conclusions concerning estimators 9, 11 and 15 are thus vindicated.

4.6.2 Robustness with respect to choices of kernel and bandwidth

In order to assess the impact of the choice of kernel and bandwidth on the performance of estimators WePa, CoWeBa and Icdf, the simulation study of Section 4.5 was again repeated for $6 = 3 \times 2$ possible choices of three kernels and two bandwidths. The kernels considered were as follows :

a) the tri-weight kernel defined in (4.7) ;

b) an Epanechnikov kernel $K(x) = 3(1 - x^2)/4$ with $x \in (-1, 1)$;

c) a uniform kernel with $K(x) = 1/2$ with $x \in (-1, 1)$.

Two choices of bandwidth were also used, namely

a) a bandwidth ignoring a possible dependence between X and Y :

$$w_1 = c\hat{\sigma}_x n^{-1/6}, \quad \omega_1 = c\hat{\sigma}_y n_+^{-1/6}$$

b) a bandwidth taking into account a possible dependence between X and Y, as measured by the correlation coefficient ρ :

$$\begin{aligned} w_2 &= c\hat{\sigma}_x n^{-1/6}(1-\hat{\rho}^2)^{5/12}(1+\hat{\rho}^2/2)^{-1/6}, \\ \omega_2 &= c\hat{\sigma}_y n_+^{-1/6}(1-\hat{\rho}^2)^{5/12}(1+\hat{\rho}^2/2)^{-1/6}. \end{aligned}$$

Here, $\hat{\sigma}_x$, $\hat{\sigma}_y$ and n_+ are defined as in Section 4.5.1, and the constant c was taken equal to 2.978, 2.214 or 1.740, according as the kernel was tri-weight, Epanechnikov or uniform. According to Scott (1992), these choices of bandwidths (and associated constants) are close to optimal for joint density estimation when the pair (X, Y) follows a bivariate normal distribution with correlation 0 or ρ. In the simulation study, this correlation was estimated by plugging in the ORE estimate for τ in the relation $\rho = \sin(\pi\tau/2)$ which links Pearson's correlation and Kendall's tau in the bivariate normal model ; see, e.g., Kruskal (1958).

Proceeding once again by analysis of variance, it was seen that neither the absolute bias nor the LMSE of the six estimators was significantly affected by the choice of kernel. Neither this factor nor any of the interactions involving it turned out to be significant at the 1% level.

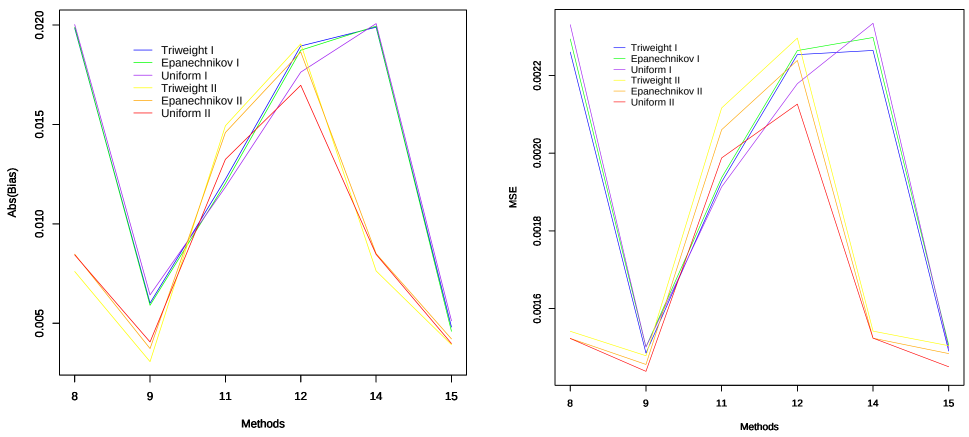


FIG. 4.7 – |Bias| and MSE of six estimators as a function of the choice of kernel and bandwidth.

In contrast, the final analysis of variance models revealed that the absolute bias and LMSE of the six estimators could be improved by an appropriate choice of bandwidth. These effects are depicted in the left and right panels of Figure 4.7, where the reddish and blueish curves correspond to a choice of dependent or independent bandwidth, respectively. Although their scales are different, the two graphs are strikingly similar in form. This suggests that the improvement in LMSE is in fact driven by a reduction of the absolute bias in all instances.

Keeping in mind that all existing estimators are (trivially) unaffected by changes of kernel and bandwidth, one can derive the following conclusions from these graphs :

a) the results reported in Section 4.5, which correspond to the dark blue curve, are close to being least favorable for estimators 8, 9, 12, 14 and 15, both in terms of absolute bias and LMSE ;

b) in particular, estimators 9 and 15 would have continued to provide the best performance under any combination of kernel and bandwidth considered ;

c) except for estimator 11, significant improvements in performance accrue from the use of the dependent-bandwidth rule, especially for estimators 8 and 14.

That proper bandwidth selection is much more crucial than the choice of kernel is a well known phenomenon in the nonparametric smoothing literature. The present study provides yet another illustration. As for point c), it makes perfect sense in the current context, where dependence between the variables is of the essence.

4.7 Illustrative example

As an application of the methods presented here, consider the classical Stanford heart transplant data. Patients who had received such a transplantation between October 1967 and February 1980 were considered in this study. Their age at transplantation was measured as a covariate, whereas their survival time (in days) represented the dependent variable. The latter was subject to right-censoring.

Following Van Keilegom et al. (2001), individuals with incomplete tissue typing (i.e., having a mismatch score $= 9999$) were dropped. There were 157 observations left, 55 of which were censored. The same kernel function and bandwidths were used as in Section 4.5.1. As the largest observation (3695 days) was censored, values of $a = 0$ and $b = 1$ were picked by forcing $H_{GKM}(3695|x_i) = 1$ for $i = 1, \ldots, 157$.

Tab. 4.6 – Various estimates of Kendall's τ and corresponding bootstrap 95% confidence intervals for the Stanford heart transplant data.

Method	Name	$\hat{\tau}$	Bootstrap 95% confidence interval
1	Oakes (1982)	-0.154	$(-0.264, -0.042)$
2	ORE (Oakes 2006)	-0.197	$(-0.341, -0.052)$
3	Brown et al. (1974)	-0.178	$(-0.306, -0.053)$
5	Weier and Basu (1980)	-0.163	$(-0.278, -0.050)$
7	Wang and Wells (2000)	-0.235	$(-0.370, -0.118)$
8	WePa with $H_{LPT}(x\|y)$	-0.168	$(-0.292, -0.049)$
9	WePa with $H_{VKA}(x\|y)$	-0.174	$(-0.288, -0.053)$
11	CoWeBa with $H_{LPT}(x\|y)$	-0.172	$(-0.304, -0.049)$
12	CoWeBa with $H_{VKA}(x\|y)$	-0.179	$(-0.293, -0.055)$
14	Icdf with $H_{LPT}(x\|y)$	-0.168	$(-0.281, -0.055)$
15	Icdf with $H_{VKA}(x\|y)$	-0.170	$(-0.287, -0.061)$
17	Icdf with Dabrowska (1988)	-0.195	$(-0.308, -0.071)$
18	Lim (2006)	-0.161	$(-0.273, -0.049)$

In this context, negative values are expected for Kendall's tau, because age at transplantation should be inversely related to remaining lifetime. The conditions of the Lemma from Section 4.2.6 are not unreasonable in this case. The estimates are reported in Table 4.6 for the 13 methods not involving any knowledge of a true, underlying distribution.

Also included in Table 4.6 are 95% confidence intervals obtained via a bootstrapping algorithm. Specifically, the following steps were repeated 1000 times :

a) draw at random (with replacement) 157 observations from the original data set ;
b) estimate Kendall's tau using each of the 13 methods that do not require any knowledge of the true underlying distribution.

For each method considered, the limits of the bootstrap interval are then given by the 25th and 975th ordered estimates of tau.

One can see from Table 4.6 that the estimates and the bootstrap confidence intervals returned by each method are comparable for these data. Exceptions are Oakes' estimator and its renormalized version, as well as the two techniques 7 and 17 that call on Dabrowska's estimator of $\bar{H}(x, y)$. It is hard to speculate on the exact cause of these discrepancies.

4.8 Conclusion

This paper has proposed several new estimators for Kendall's tau between two survival times X and Y, when only the latter component is subject to right-censoring. Through simulations, the new methods have been compared to the existing ones, both in terms of bias and mean-squared error. These comparisons were made under a wide variety of conditions, whether in terms of sample size, censoring fraction, structure and strength of dependence between the variables, nature of their marginal distributions, as well as choice of kernel and bandwidth involved in the construction of the new estimators.

When the degree of dependence or censoring in the data is small, the difference between the various estimation methods was negligible in the simulations. This point is illustrated by the Stanford heart transplant data set considered in Section 4.7. As the dependence and the censoring fraction increase, however, most of the existing estimators have a strong tendency to underestimate the true magnitude of τ.

The WePa–VKA (9) and Icdf–VKA (15) estimators, introduced here for the first time, performed significantly better than the others under conditions of strong dependence and heavy censoring. It seems, therefore, that conditioning on the fully observed covariate truly allows for a more efficient use of the information contained in the indeterminate pairs. As these pairs constitute a substantial portion of the sample under heavy censoring, the gain in precision provided by the new methods is then non-negligible.

Estimators WePa–VKA (9) and Icdf–VKA (15) can both be recommended with confidence under all scenarios considered herein, particularly in view of their relative robustness to the choice of kernel and bandwidth. Estimators WePa–LPT (8), CoWeBa–LPT (11), CoWeBa–VKA (12) and Icdf–LPT (14) are also promising, especially for highly dependent variables, but their performance is somewhat sensitive to the choice of bandwidth, although not on the choice of kernel. Additional tuning will be required before they can be recommended without qualification.

In future work, it would also be of interest to examine the distributional properties of the new estimators of τ. An even more ambitious project would be to try and extend the present estimation strategies to situations involving double censoring. While this would pose no great challenge from a theoretical point of view, the computational investigation associated with it would be prohibitive.

Acknowledgements

The authors are grateful to Lajmi Lakhal-Chaieb for useful methodological suggestions and to six anonymous referees for constructive criticisms on earlier drafts of this work. Thanks are also due to Dr Weijing Wang for providing a program that estimates Kendall's tau under the method described in Wang & Wells (2000a). Partial funding in support of this work was provided by the Natural Sciences and Engineering Research Council of Canada, the Fonds québécois de la recherche sur la nature et les technologies, and the Institut de finance mathématique de Montréal.

Transition

L'article précédent s'intéressait à la problématique d'estimation du tau de Kendall lorsque l'une des deux variables d'intérêt est sujette au phénomène de censure. De nouveaux estimateurs y sont proposés et une vaste étude de simulation tend à démontrer que la précision des estimations de tau par ces méthodes innovatrices est largement meilleure que celle obtenue via les méthodes déjà existantes. Cette amélioration s'explique par l'emploi de l'information jointe entre les deux variables, par opposition au simple recours aux distributions marginales privilégié par divers auteurs dans le passé.

Le chapitre suivant porte également sur l'estimation du tau de Kendall, mais cette fois en émettant l'hypothèse que les deux variables étudiées sont sujettes à la censure. L'angle sous lequel le problème est attaqué s'avère cependant fort différent. Au lieu de faire appel à l'information jointe des variables, l'estimation proposée s'appuie exclusivement sur les paires d'observations dont le statut de concordance ou discordance peut être établi de façon certaine (on parle alors de paires *ordorables*). En vue d'éliminer le biais, ces paires d'observations sont pondérées par l'inverse de la probabilité d'être ordorables. Cet article compare cette nouvelle méthode à celles proposées dans la littérature, en plus de démontrer la convergence et la normalité asymptotique du nouvel estimateur.

Chapitre 5

IPCW estimator for Kendall’s tau under bivariate censoring

Résumé

Nous investiguons l’estimation non paramétrique du coefficient de concordance de Kendall, τ, qui mesure le niveau d’association entre deux variables sujettes à la censure bivariée. L’estimateur proposé constitue en fait une modification de celui introduit par Oakes (1982) en utilisant une correction du type Horvitz-Thompson pour les paires non ordorables. En présence de données censurées, une paire est dite ordorable si l’on peut établir avec certitude la nature concordante ou discordante de la paire non censurée qui est associée à cette paire, et ce, en se basant sur l’information disponible. Notre estimateur s’avère convergent et asymptotiquement normalement distribué. Une étude de simulation démontre la bonne performance de l’estimateur proposé dans cet article vis-à-vis les estimateurs déjà existants. Ces diverses méthodes d’estimation sont par la suite illustrées à l’aide de deux jeux de données réels.

Abstract

We investigate the nonparametric estimation of Kendall's coefficient of concordance, τ, for measuring the association between two variables under bivariate censoring. The proposed estimator is a modification of the estimator introduced by Oakes (1982), using a Horvitz-Thompson-type correction for the pairs that are not orderable. With censored data, a pair is orderable if one can establish whether the uncensored pair is discordant or concordant using the data available for that pair. Our estimator is shown to be consistent and asymptotically normally distributed. A simulation study shows that the proposed estimator performs well when compared with competing alternatives. The various methods are illustrated with two real data sets.

5.1 Introduction

In many biomedical experiments, the prime interest of the study is to investigate the relationship between two random variables. Kendall's coefficient of concordance τ, is a simple measure of association between a pair of lifetime random variables (X, Y). This measure is independent of the marginal distributions for X and Y. Its rank invariance property makes it particularly suitable, especially when only the association between the random variables is of interest. Semi-parametric models for bivariate data, see for instance Clayton (1978), Oakes (1986), and Genest (1987), use a copula to model the dependency between the variables. Kendall's τ is a function of the copula parameters ; thus estimators of Kendall's τ naturally yield estimators of the copula parameter in semi-parametric models (Genest & Rivest, 1993; Wang & Wells, 2000b). Nonparametric estimation of τ from n identical independent replications $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ of (X, Y) has been extensively studied, see for instance Kendall & Gibbons (1990) and Gibbons (1971).

In recent years, a substantial research effort has been devoted to the estimation of the dependence between lifetime random variables from incomplete data, see for instance Lin et al. (1999), Wang & Wells (1997), Wang & Wells (1998) and Fine et al. (2001). This paper focuses on the nonparametric estimation of τ under bivariate censoring, that is, when only $(\tilde{X}, \tilde{Y}, \delta_X, \delta_Y)$ are observable, where $\tilde{X} = \min(X, C_X)$, $\delta_X = I_{\{X<C_X\}}$ is a censoring indicator, C_X is a censoring random variable independent of X, and $\tilde{Y}$, C_Y, and δ_Y are defined in a similar way for Y. Brown et al. (1974), Weier & Basu (1980) and Oakes (1982) modified the estimation of τ to account for censoring in both coordinates. It turns out that none of these estimators is consistent when $\tau \neq 0$. Alternatively, Wang & Wells (2000a) derived an estimator for τ expressed as an integral of an estimate of the bivariate survival function. The performance of this von Mises-type estimator depends heavily on the selection of the survival function estimator.

In this paper, we propose a new nonparametric estimator for τ under bivariate censoring based on a modification of Oakes (1982). The contribution of each orderable pair to the coefficient is weighted by the inverse probability that the pair is orderable. A pair is orderable if the concordance-discordance status of the pair can be established using the information available in the censored sample. Our estimator takes different expressions depending on the censoring scheme. Four situations are considered : independent censoring variables (C_X and C_Y independent), censoring on X only ($C_Y = \infty$), univariate censoring ($C_X = C_Y = C$) and the general case (C_X and C_Y dependent). Wang & Wells (1997) investigated the estimation of the bivariate survival function under these simplified censoring schemes. The proposed Horvitz-Thompson-type estimator is an extension of the Inverse Probabilily Censoring Weighted (IPCW) estimators class of Robins & Rotnitzky (1992) to multivariate selection probabilities. This class of estimators allows the estimation of the joint survival function for successive events (Lin et al., 1999), the estimation of the mean quality adjusted lifetime with censored data (Zhao & Tsiatis, 1997, 2000) and the estimation of regression parameters in a multiplicative intensity model (van der Laan & Robins, 2003).

This new estimator for τ is shown to be consistent under suitable regularity conditions. It is also asymptotically normally distributed for the first three cases. Simulations comparing the proposal with existing estimators show its good performances. In Section 5.2, we present our estimator and investigate its asymptotic behavior in Section 5.3. Simulation results are provided in Section 5.4 and we conclude with a discussion in Section 5.5.

5.2 Estimation of τ

Let (X_1, Y_1) and (X_2, Y_2) be two independent replications of (X, Y), a bivariate lifetime random variable with continuous marginals $S_X(x) = P(X > x)$ and $S_Y(y) = P(Y > y)$. This pair is said to be concordant if $(X_1 - X_2)(Y_1 - Y_2) > 0$ and discordant if $(X_1 - X_2)(Y_1 - Y_2) < 0$. Kendall's tau (Kendall & Gibbons, 1990) is defined by

$$\begin{aligned}\tau &= P\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - P\{(X_1 - X_2)(Y_1 - Y_2) < 0\} \\ &= 2P\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - 1 \\ &= E(a_{12}b_{12}) \\ &= 4\int_0^\infty \int_0^\infty \pi(x,y)\frac{\partial^2 \pi(x,y)}{\partial x \partial y} dxdy - 1 \qquad (5.1)\end{aligned}$$

where $a_{ij} = 2 \times I_{\{X_i - X_j > 0\}} - 1$, $b_{ij} = 2 \times I_{\{Y_i - Y_j > 0\}} - 1$ and $\pi(.,.)$ is the bivariate survival function of (X, Y) defined by $\pi(x, y) = P(X > x, Y > y)$.

In the absence of censoring, one can estimate τ by its sample version

$$\hat{\tau}_K = \binom{n}{2}^{-1} \sum_{i<j} a_{ij} b_{ij}. \qquad (5.2)$$

5.2.1 Previous estimators under bivariate censoring

With censored values, the concordance/discordance status can be established only for orderable pairs. Oakes (1982) showed that a pair is orderable if $\{\tilde{X}_{ij} < \tilde{C}_X^{ij}, \tilde{Y}_{ij} < \tilde{C}_Y^{ij}\}$ where $\tilde{X}_{ij} = \min\,(X_i, X_j)$, $\tilde{Y}_{ij} = \min\,(Y_i, Y_j)$, $\tilde{C}_X^{ij} = \min\,(C_X^i, C_X^j)$ and $\tilde{C}_Y^{ij} = \min\,(C_Y^i, C_Y^j)$. Let $L_{ij} = 1$ if the pairs i and j are orderable, and $L_{ij} = 0$ otherwise. Several alternatives to (5.2) have been proposed to estimate τ under bivariate censoring.

Brown et al. (1974) modified the definitions of a_{ij} and b_{ij} for non orderable pairs :

$$\begin{aligned} a'_{ij} &= 2\hat{P}(X_i - X_j > 0|\tilde{X}_i, \tilde{X}_j, \delta_X^i, \delta_X^j) - 1 \\ b'_{ij} &= 2\hat{P}(Y_i - Y_j > 0|\tilde{Y}_i, \tilde{Y}_j, \delta_Y^i, \delta_Y^j) - 1. \end{aligned}$$

A pair of points is not orderable as soon as one of the events $(\tilde{X}_{ij} > \tilde{C}_X^{ij})$ or $(\tilde{Y}_{ij} > \tilde{C}_Y^{ij})$ is true. The condition on the X's holds if either one of the following mutual exclusive events is true : $(\tilde{X}_i > \tilde{X}_j; \delta_X^i = 1; \delta_X^j = 0), (\tilde{X}_i > \tilde{X}_j; \delta_X^i = 0; \delta_X^j = 0), (\tilde{X}_i < \tilde{X}_j; \delta_X^i = 0; \delta_X^j = 1)$ or $(\tilde{X}_i < \tilde{X}_j; \delta_X^i = 0; \delta_X^j = 0)$. It is easy to see that

$$P(X_i - X_j > 0|\tilde{X}_i > \tilde{X}_j; \delta_X^i = 1; \delta_X^j = 0) = 1 - \frac{S_X(\tilde{x}_i)}{S_X(\tilde{x}_j)} \tag{5.3}$$

$$P(X_i - X_j > 0|\tilde{X}_i > \tilde{X}_j; \delta_X^i = 0; \delta_X^j = 0) = 1 - \frac{S_X(\tilde{x}_i)}{2\ S_X(\tilde{x}_j)} \tag{5.4}$$

$$P(X_i - X_j > 0|\tilde{X}_i < \tilde{X}_j; \delta_X^i = 0; \delta_X^j = 1) = \frac{S_X(\tilde{x}_j)}{S_X(\tilde{x}_i)} \tag{5.5}$$

$$P(X_i - X_j > 0|\tilde{X}_i < \tilde{X}_j; \delta_X^i = 0; \delta_X^j = 0) = \frac{S_X(\tilde{x}_j)}{2\ S_X(\tilde{x}_i)}, \tag{5.6}$$

where $\tilde{x}_i$ and $\tilde{x}_j$ are the observed values of $\tilde{X}_i$ and $\tilde{X}_j$.

The coefficients a'_{ij} for non orderable pairs are obtained by substituting the probabilities in the righthand side of (5.3), (5.4), (5.5) and (5.6) by their Kaplan-Meier estimates based on $\{(\tilde{X}_k, \delta_X^k),\ k = 1, \cdots, n\}$. For orderable pairs, set $a'_{ij} = a_{ij}$. The b'_{ij} are defined analogously.

This yields the estimator

$$\hat{\tau}_B = \frac{\sum_{i,j} a'_{ij} b'_{ij}}{(\sum_{i,j} a'^2_{ij} \sum_{i,j} b'^2_{ij})^{1/2}}. \tag{5.7}$$

Oakes (1982) modified (5.2) by summing over orderable pairs only :

$$\hat{\tau}_O = \binom{n}{2}^{-1} \sum_{i<j} L_{ij} a_{ij} b_{ij}. \tag{5.8}$$

It turns out that none of these estimators is consistent when $\tau \neq 0$, since they ignore the dependence between X and Y when estimating the concordance-discordance status of a non orderable pair.

Alternatively, Wang & Wells (2000a) used (5.1) to estimate τ by

$$\begin{aligned}\hat{\tau}_W &= 4\int_0^\infty\int_0^\infty \hat{\pi}(x,y)\frac{\partial^2\hat{\pi}(x,y)}{\partial x\partial y} - 1\\ &= 4\sum_{i=1}^n\sum_{j=1}^n \hat{\pi}(x_{(i)},y_{(j)})\hat{\pi}(\Delta x_{(i)},\Delta y_{(j)}) - 1, \qquad (5.9)\end{aligned}$$

where $\{x_{(0)} = 0 < x_{(1)} < x_{(2)} < \cdots < x_{(n)}\}$ and $\{y_{(0)} = 0 < y_{(1)} < y_{(2)} < \cdots < y_{(n)}\}$ are the ordered samples of $\{\tilde{X}_k,\ k = 1,\cdots,n\}$ and $\{\tilde{Y}_k,\ k = 1,\cdots,n\}$, $\hat{\pi}(.,.)$ is a nonparametric estimator for $\pi(.,.)$ and $\hat{\pi}(\Delta x_{(i)},\Delta y_{(j)}) = \hat{\pi}(x_{(i)},y_{(j)}) - \hat{\pi}(x_{(i-1)},y_{(j)}) - \hat{\pi}(x_{(i)},y_{(j-1)}) + \hat{\pi}(x_{(i-1)},y_{(j-1)})$ is the estimated probability mass of the rectangle $[x_{(i-1)},x_{(i)}] * [y_{(j-1)},y_{(j)}]$. Several nonparametric estimators for the bivariate survival function can be plugged in (5.9), for instance those of Campbell (1981), Dabrowska (1988), and Prentice & Cai (1992) ; the resulting $\hat{\tau}_W$ critically depends on this selection.

5.2.2 New Estimators

One can consider the orderable pairs as a sample selected from the population of $\binom{n}{2}$ possible pairs. A common technique in survey sampling to correct the bias consists of weighting the pairs in (5.8) by the inverse estimated probabilities $\hat{p}_{ij}$ of being selected (Horvitz & Thompson, 1952). This yields

$$\hat{\tau}_{mo} = \binom{n}{2}^{-1}\sum_{i<j}\frac{L_{ij}a_{ij}b_{ij}}{\hat{p}_{ij}} \qquad (5.10)$$

When $|\tau|$ is close to 1, $\hat{\tau}_{mo}$ might lay outside $[-1,1]$. An alternative to address this issue is to consider $\hat{\tau}_{mo2}$ defined by

$$\hat{\tau}_{mo2} = \{\sum_{i<j}\frac{L_{ij}a_{ij}b_{ij}}{\hat{p}_{ij}}\}/\{\sum_{i<j}\frac{L_{ij}}{\hat{p}_{ij}}\} \qquad (5.11)$$

It is easy to see that $\hat{\tau}_{mo2}$ always lays inside $[-1,1]$.

Note that $\tilde{X}_{i,j}$ and $\tilde{Y}_{i,j}$ are observed for all orderable pairs so the individual selection probabilities are expressed as

$$\begin{aligned} p_{ij} &= P\{L_{ij} = 1|\tilde{X}_{ij}, \tilde{Y}_{ij}\} \\ &= P\{\tilde{C}_X^{ij} > \tilde{X}_{ij}, \tilde{C}_Y^{ij} > \tilde{Y}_{ij}|\tilde{X}_{ij}, \tilde{Y}_{ij}\} \\ &= P\{C_X^i > \tilde{X}_{ij}, C_X^j > \tilde{X}_{ij}, C_Y^i > \tilde{Y}_{ij}, C_Y^j > \tilde{Y}_{ij}|\tilde{X}_{ij}, \tilde{Y}_{ij}\} \\ &= P\{C_X > \tilde{X}_{ij}, C_Y > \tilde{Y}_{ij}|\tilde{X}_{ij}, \tilde{Y}_{ij}\}^2 \end{aligned} \tag{5.12}$$

The estimation of p_{ij} for orderable pairs depends on the censoring scheme.

Independent Censoring Variables

In this case, (C_X, C_Y) are independent so that the selection probability p_{ij} is written as

$$p_{ij} = \{S_X^c(\tilde{X}_{ij})\ S_Y^c(\tilde{Y}_{ij})\}^2, \tag{5.13}$$

where S_X^c and S_Y^c are the survival functions of respectively C_X and C_Y. These can be estimated by the Kaplan-Meier estimators based respectively on $\{(\tilde{X}_k, 1-\delta_X^k), k = 1, \ldots, n\}$ and $\{(\tilde{Y}_k, 1-\delta_Y^k), k = 1, \ldots, n\}$. This yields

$$\hat{p}_{ij} = \{\hat{S}_X^c(\tilde{X}_{ij})\ \hat{S}_Y^c(\tilde{Y}_{ij})\}^2. \tag{5.14}$$

Censoring on X only

When only one coordinate, say X, is censored, $C_Y = \infty$. Therefore, this scenario can be viewed as a particular case of the previous section, where $S_Y^c(\tilde{Y}_{ij}) = 1$. Clearly, the individual selection probability can thus be written as

$$p_{ij} = \{S_X^c(\tilde{X}_{ij})\}^2, \tag{5.15}$$

which can be estimated by the Kaplan-Meier estimator based on $\{(\tilde{X}_k, 1-\delta_X^k), k = 1, \ldots, n\}$.

Univariate Censoring

In this case, $C_X = C_Y = C$ so that the selection probability p_{ij} is expressed as

$$\begin{aligned} p_{ij} &= P\{\tilde{X}_{ij} < C, \tilde{Y}_{ij} < C\}^2 \\ &= [S^c\{\max\ (\tilde{X}_{ij}, \tilde{Y}_{ij})\}]^2, \end{aligned}$$

where S^c is the survival function of the common censoring variable C. It can be estimated by the Kaplan-Meier estimator based on $[\{\max\ (\tilde{X}_k, \tilde{Y}_k), 1-\delta_X^k\delta_Y^k\}, k = 1, \ldots, n]$.

General Case

In the general case, the relationship between C_X and C_Y is left completely unspecified. The selection probability is expressed in terms of the bivariate survival function $\pi^c(.,.)$ of the censoring variables :

$$p_{ij} = \{\pi^c(\tilde{X}_{ij}, \tilde{Y}_{ij})\}^2. \tag{5.16}$$

As for $\hat{\tau}_W$, several nonparametric estimators are available for the bivariate survival function. In our procedure, we shall use the Dabrowska estimator. It is clear that plugging a different estimator for the survival function leads to a different estimator of Kendall's tau.

5.3 Asymptotic Behavior

In this section, we investigate the asymptotic properties of our estimator. Define $\tilde{\tau}$ by

$$\tilde{\tau} = \binom{n}{2}^{-1} \sum_{i<j} \frac{L_{ij}}{p_{ij}} a_{ij} b_{ij} \tag{5.17}$$

and write $\sqrt{n}(\hat{\tau} - \tau)$ as $\sqrt{n}(\hat{\tau} - \tilde{\tau}) + \sqrt{n}(\tilde{\tau} - \tau)$. Assume for now that $p_{ij} > 0$ for all (i, j). This assumption holds if the support of $(\tilde{X}, \tilde{Y})$ is equal to that of (X, Y), that is if all the possible values of (X, Y) are observable. Under this assumption, we show in Appendix B that, under certain regularity conditions on $\hat{p}_{ij}$, $\sqrt{n}(\hat{\tau} - \tilde{\tau})$ is asymptotically equivalent to a zero mean U-statistics of order 3. In Appendix A, we show that $\sqrt{n}(\tilde{\tau} - \tau)$ is a zero mean U-statistics of order 2. And thus, $\sqrt{n}(\hat{\tau} - \tau)$ is asymptotically equivalent to a zero mean U-statistics of order 3. Consistency and asymptotic normality follow (van der Vaart, 1998). However, the computation of the variance involves complex formulas so we use the jackknife resampling procedure to estimate it. In the general case of Section 5.2.2, $\hat{p}_{ij}$ does not meet the regularity conditions required in Appendix A. Therefore, the proof of asymptotic normality does not hold for this case. Nevertheless, the consistency of the Dabrowska estimator insures that of $\hat{\tau}$.

The condition $p_{ij} > 0$ for $i, j = 1, \ldots, n$ fails to hold when the support of the censoring variables is shorter than that of the failure times of interest. In this case, all the observed points belong to $\mathcal{B}$, the support of the censoring variables ; $\mathcal{B}$ is the observable region. This may happen if the termination of the study causes data points to be censored. The concordance/discordance relationship outside $\mathcal{B}$ is missing. Wang & Wells

(2000a) discuss this issue and show that the parameter estimated by their estimator is

$$\tau^* = 4 \int_{\mathcal{B}} \pi(x,y) \frac{\partial^2 \pi(x,y)}{\partial x \partial y} dxdy - 1.$$

The new estimator is also vulnerable to this problem because the missing information is not recoverable.

Assume that X and Y have the same support, say $[0,1]$, and that the supports of C_X and C_Y are $[0,A]$ and $[0,B]$ respectively, where $0 < A, B < 1$. The square $[0,1]^2$ is then divided into four regions :

$$\begin{aligned} R_1 &= \{0 \le x \le A, 0 \le y \le B\} \\ R_2 &= \{0 \le x \le A, B \le y \le 1\} \\ R_3 &= \{A \le x \le 1, 0 \le y \le B\} \\ R_4 &= \{A \le x \le 1, B \le y \le 1\}. \end{aligned}$$

The observable region is R_1. Consider now a pair (i,j) of observations. The pair $\{(X_i,Y_i),(X_j,Y_j)\}$ has a positive probability of being orderable, e.g. $p_{ij} > 0$, if and only if at least one point falls in R_1, or one point belongs to R_2 and the other to R_3. Denote this event by O_{ij}. Kendall's tau is then expressed as :

$$\begin{aligned} \tau &= E(a_{12}b_{12}) \\ &= E(a_{12}b_{12}|O_{12})P(O_{12}) + E(a_{12}b_{12}|O_{12}^c)P(O_{12}^c) \end{aligned} \tag{5.18}$$

where O_{ij}^c is O_{ij}'s complement. Thus, we have no information about $E(a_{12}b_{12}|O_{12}^c)$. In this case, $\hat{\tau}_{mo}$ and $\hat{\tau}_{mo2}$ are estimating τ^c and τ_2^c respectively instead of estimating τ, where

$$\tau^c = E(a_{12}b_{12}|O_{12})P(O_{12}) \quad \text{and} \quad \tau_2^c = E(a_{12}b_{12}|O_{12}) \tag{5.19}$$

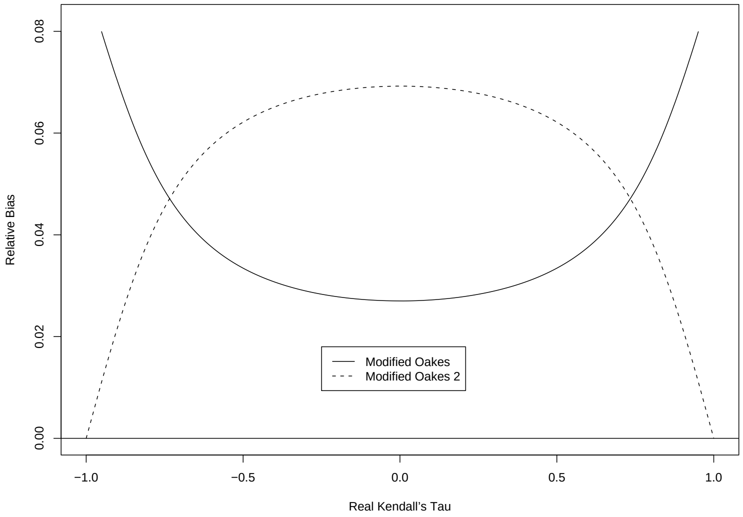
Elementary algebra yields

$$\begin{aligned} \tau^c - \tau &= -E(a_{12}b_{12}|O_{12}^c)P(O_{12}^c) \\ \tau_2^c - \tau &= P(O_{12}^c)\{E(a_{12}b_{12}|O_{12}) - E(a_{12}b_{12}|O_{12}^c)\}. \end{aligned}$$

Consider the case where (X,Y) follows a Frank copula (Genest, 1987). This Archimedean copula is indexed by a one-dimensioned parameter α measuring the dependence between X and Y. It is related to Kendall's tau by

$$\tau = 1 + \frac{4}{\alpha}\left\{\int_0^\alpha \frac{t}{e^t-1}dt - 1\right\}. \tag{5.20}$$

For simplicity, assume that only X is subject to censoring ($C_Y = \infty$). Under these conditions, $P(O_{12}^c) = (1-A)^2$, $P(O_{12}) = 1 - P(O_{12}^c)$ and $E(a_{12}b_{12}|O_{12}^c)$ corresponds



to *the truncated Kendall's tau* as defined and expressed by Manatunga & Oakes (1996) for an arbitrary Archimedean copula. For Frank's copula, we obtain

$$E(a_{12}b_{12}|O_{12}^c) = \frac{4}{t} + 1 + \text{dilog}(\exp(t))\frac{4}{t^2}.$$

where $t = \alpha(1 - A)$ and $\text{dilog}(v) = \int_1^v \log(s)/(1-s)ds$.

Expression for $E(a_{12}b_{12}|O_{12})$, and therefore for $(\tau^c - \tau)/\tau$ and $(\tau_2^c - \tau)/\tau$ are derived from (5.18) and (5.20). In Figure 1, we present these relative biases for $A = 0.7$. This figure suggests to use τ_{mo} for small values of $|\tau|$ and τ_{mo2} for large values of $|\tau|$.

5.4 Numerical examples

5.4.1 Simulations

A series of simulations were carried out to assess the finite-sample performance of the proposal and to compare it to existent estimators. The pairs (X, Y) were generated from a Clayton copula (Clayton, 1978) with unit exponential marginals. The censoring variables (C_X, C_Y) were generated with exponential marginals with a parameter controlling the censoring fraction. Samples were simulated with a size of 100, 200 and 400, a Kendall's tau of 0.2, 0.5 and 0.8 and a censoring fraction (% C) of 20% and 40%. Under each scheme, 1000 replications were performed. Results are reported in Tables 5.1-5.4. In each cell, we reported the bias $(*10^3)$ and the root mean squared error ($*10^3$) for 3 estimators (modified Oakes, Wang and Wells, and Brown et al). We considered four situations : independent censoring variables (Table 5.1), censoring on X only (Table 5.2), univariate censoring (Table 5.3) and dependent censoring variables (Table 5.4). For this last case, (C_X, C_Y) were simulated from a Clayton copula with $\tau = 0.5$.

The proposed estimator, called modified Oakes, is compared to that of Brown et al. (1974) and of Wang & Wells (2000a). The latter used the bivariate survival estimator proposed by Dabrowska (1988).

% C	n	τ	Modified Oakes	Brown et al	Wang & Wells
20	100	0.2	-0.18 (65.60)	9.77 (66.75)	-20.80 (69.30)
		0.5	0.62 (58.24)	23.50 (61.02)	-20.90 (62.52)
		0.8	-4.22 (38.25)	20.38 (32.92)	-27.91 (49.52)
	200	0.2	-1.71 (52.24)	8.18 (53.04)	-11.35 (53.74)
		0.5	0.39 (41.36)	22.67 (45.45)	-10.07 (42.43)
		0.8	-1.10 (27.64)	21.43 (28.34)	-11.74 (30.19)
	400	0.2	1.05 (34.36)	10.93 (36.17)	-4.03 (34.66)
		0.5	-0.21 (28.18)	22.53 (35.33)	-4.76 (29.02)
		0.8	-0.93 (20.86)	21.25 (24.89)	-4.35 (19.59)
40	100	0.2	2.94 (95.75)	29.52 (84.99)	-20.82 (99.43)
		0.5	-8.74 (91.79)	44.88 (76.58)	-29.76 (97.28)
		0.8	-14.63 (88.73)	-12.11 (35.66)	-24.54 (116.25)
	200	0.2	-2.37 (64.59)	26.70 (59.80)	-11.77 (65.42)
		0.5	0.79 (62.78)	48.80 (64.25)	-7.23 (66.76)
		0.8	-6.01 (60.04)	-11.16 (26.74)	-9.35 (76.22)
	400	0.2	-1.74 (44.11)	26.44 (46.45)	-5.82 (45.50)
		0.5	-1.62 (40.72)	46.79 (55.07)	-3.72 (41.15)
		0.8	-1.94 (42.83)	-10.32 (19.77)	-0.48 (46.68)

TAB. 5.1 – Independent censoring : Biases ($\times 10^3$) and root mean squared errors ($\times 10^3$, in parentheses) for three estimators of τ.

Simulations confirm the substantial bias reduction associated with our estimator, especially with independent censoring variables, censoring on X only and univariate censoring. As expected, the performance of our estimator improves when the sample size increases and Kendall's tau or the censoring fraction decrease. In the dependent censoring variables case, $\hat{\tau}_{mo}$ and $\hat{\tau}_W$ suffer from the high variability of $\hat{\pi}^c$ and $\hat{\pi}$ obtained from the method of Dabrowska (1988). This is the main drawback of these estimators. Using a better estimator for the bivariate survival function will substantially improve

% C	n	τ	Modified Oakes	Brown et al	Wang & Wells
20	100	0.2	-1.81 (67.65)	3.45 (67.95)	-17.52 (70.21)
		0.5	-2.09 (56.35)	9.03 (56.44)	-21.11 (60.80)
		0.8	-2.04 (30.05)	7.67 (27.60)	-23.78 (40.04)
	200	0.2	-1.31 (47.44)	3.87 (47.68)	-9.26 (48.45)
		0.5	0.24 (39.45)	11.30 (40.46)	-9.30 (40.86)
		0.8	-1.52 (21.88)	7.58 (20.41)	-11.87 (25.65)
	400	0.2	-1.00 (33.21)	4.01 (33.52)	-5.08 (33.62)
		0.5	-0.98 (27.14)	10.04 (28.87)	-5.52 (27.80)
		0.8	-0.15 (15.00)	8.25 (15.50)	-5.47 (16.42)
40	100	0.2	4.94 (74.01)	18.31 (72.86)	-15.44 (77.36)
		0.5	-3.50 (64.31)	15.64 (58.67)	-24.52 (71.90)
		0.8	-6.12 (38.38)	-29.70 (43.42)	-25.99 (58.81)
	200	0.2	-0.47 (54.71)	12.34 (51.83)	-10.48 (56.69)
		0.5	-0.02 (44.45)	18.87 (43.44)	-11.46 (47.50)
		0.8	-4.03 (27.58)	-29.42 (36.89)	-14.14 (38.67)
	400	0.2	1.42 (37.05)	14.45 (38.05)	-3.58 (37.88)
		0.5	-1.53 (31.02)	18.15 (32.92)	-6.69 (33.36)
		0.8	-0.80 (17.90)	-27.82 (31.66)	-5.23 (25.15)

TAB. 5.2 – Censoring on X only : Biases ($\times 10^3$) and root mean squared errors ($\times 10^3$, in parentheses) for three estimators of τ.

$\hat{\tau}_{mo}$. When τ is near 1 or -1, $\hat{\tau}_{mo}$ may lie outside $[-1, 1]$. In that case, one can estimate τ by $\hat{\tau}_{mo2}$ given by (5.11). Simulations, not reported here, show that $\hat{\tau}_{mo}$ performs, in general, better than $\hat{\tau}_{mo2}$.

5.4.2 Real Data

Two real data sets were used to illustrate the proposed estimator.

Heart Transplant Data

We consider first a data set issued from the Stanford heart transplantation program (Miller & Halpern, 1982). From October 1967 to February 1980, 157 patients received heart transplantation. The relationship between the survival time after transplantation

% C	n	τ	Modified Oakes	Brown et al	Wang & Wells
20	100	0.2	0.88 (71.61)	11.11 (72.64)	-20.23 (74.34)
		0.5	-1.08 (60.27)	22.76 (62.55)	-25.74 (66.40)
		0.8	-1.78 (31.93)	26.19 (37.31)	-28.38 (45.19)
	200	0.2	-0.80 (49.11)	8.71 (50.30)	-11.20 (50.90)
		0.5	1.34 (40.66)	24.08 (46.91)	-11.03 (42.13)
		0.8	-1.29 (22.23)	26.06 (32.52)	-15.32 (28.32)
	400	0.2	0.61 (35.05)	10.56 (37.18)	-4.44 (35.61)
		0.5	0.31 (28.99)	23.12 (36.82)	-5.61 (29.58)
		0.8	-1.01 (15.19)	25.97 (29.11)	-8.03 (18.09)
40	100	0.2	3.32 (83.67)	33.86 (86.89)	-28.66 (91.16)
		0.5	0.12 (73.55)	70.95 (95.74)	-36.99 (87.70)
		0.8	-7.99 (42.29)	61.30 (67.36)	-46.43 (70.60)
	200	0.2	-2.05 (61.74)	28.96 (64.22)	-18.36 (65.43)
		0.5	-0.55 (49.99)	69.30 (82.24)	-18.98 (55.75)
		0.2	-3.53 (28.94)	63.14 (66.05)	-24.11 (43.50)
	400	0.2	-0.57 (41.77)	30.53 (50.35)	-8.83 (43.01)
		0.5	-1.28 (34.94)	68.03 (74.45)	-9.60 (38.01)
		0.8	-2.08 (19.59)	62.23 (63.66)	-12.66 (26.57)

TAB. 5.3 – Univariate censoring : Biases ($\times 10^3$) and root mean squared errors ($\times 10^3$, in parentheses) for three estimators of τ.

and the patient's age at the time of transplantation is of interest. Of the 157 patients, 55 were still alive at the end of the study and 102 were deceased. Since age is available for all subjects, only the survival time is censored. Using the proposed method, we obtain $\hat{\tau}_{mo} = -0.185$ *(s.e. 0.061)*. The estimates by Wang & Wells (2000a) and Brown et al. (1974) are respectively $\hat{\tau}_W = -0.235$ *(s.e 0.062)* and $\hat{\tau}_B = -0.178$ *(s.e 0.062)*. Without knowledge of the true τ, it is very difficult to compare these estimates. We simulated 1000 samples under similar conditions to the data set ($n = 157$, $\tau = -0.2$, a single censored coordinate with a censoring fraction of 0.35). The biases and root mean squared errors of the three estimators are $(.001, .056)$, $(.008, .057)$, $(.010, .057)$, for $\hat{\tau}_{mo}$ $\hat{\tau}_W$, and $\hat{\tau}_B$ respectively. Thus the proposed estimator is marginally better than its competitors in this case.

% C	n	τ	Modified Oakes	Brown et al	Wang & Wells
20	100	0.2	-1.34 (72.01)	9.57 (71.25)	-20.95 (73.62)
		0.5	-0.91 (67.12)	22.46 (60.46)	-24.15 (62.99)
		0.8	17.64 (66.01)	25.76 (36.58)	-27.42 (45.18)
	200	0.2	0.20 (49.43)	10.39 (49.28)	-9.88 (49.95)
		0.5	2.64 (43.70)	23.45 (45.52)	-10.94 (41.63)
		0.8	21.69 (46.18)	25.36 (31.57)	-12.79 (28.01)
	400	0.2	-0.73 (35.16)	8.71 (35.48)	-6.23 (34.80)
		0.5	0.45 (32.62)	21.79 (35.78)	-7.02 (29.98)
		0.8	22.14 (37.58)	25.15 (28.22)	-6.33 (17.86)
40	100	0.2	-0.23 (128.97)	27.04 (83.96)	-37.58 (95.09)
		0.5	46.47 (157.49)	60.71 (85.70)	-38.83 (88.71)
		0.8	18.58 (117.75)	36.87 (47.51)	-40.00 (82.94)
	200	0.2	11.83 (92.34)	32.59 (62.81)	-14.13 (60.48)
		0.5	67.66 (122.95)	65.27 (79.01)	-15.14 (57.55)
		0.8	4.72 (113.80)	39.29 (44.32)	-15.33 (52.98)
	400	0.2	12.97 (91.70)	30.61 (49.34)	-7.13 (42.28)
		0.5	66.42 (99.15)	63.77 (70.91)	-6.64 (37.85)
		0.8	-27.81 (110.27)	39.11 (41.66)	-6.90 (32.75)

TAB. 5.4 – Dependent censoring (Clayton, $\tau(C_X, C_Y) = 0.5$) : Biases ($\times 10^3$) and root mean squared errors ($\times 10^3$, in parentheses) for three estimators of τ.

Kidney Data

The second data set concerns a study of the recurrence time of infection in kidney patients using a portable dialysis machine (McGilchrist & Aisbett, 1991). Once an infection occurs, the catheter is removed and is not reinserted until the infection is cleared up. Let X and Y be respectively the times to the first 2 infections. Censoring can occur due to removal for other reasons or the end of the study. Of the 38 observations, 6 were censored in X, 12 in Y and 3 in both coordinates. We tested the independence assumption between the censoring variables using a permutation test with $B = 10000$ iterations. We got $p - value = 0.79$ providing no evidence against independence. Our estimate of Kendall's tau between X and Y is $\hat{\tau}_{mo} = 0.245$ *(s.e. 0.119)*, while we have $\hat{\tau}_B = 0.243$ *(s.e 0.120)* and $\hat{\tau}_W = 0.269$ *(s.e 0.184)*. All methods give similar results. We simulated 1000 samples under similar conditions ($n = 38$, $\tau = 0.25$, independent censoring variables with censoring fractions of 0.28 and 0.18). The biases and root mean squared errors of the three estimators are $(.002, .118)$, $(-.060, .135)$, $(.015, .114)$, for $\hat{\tau}_{mo}$ $\hat{\tau}_W$, and $\hat{\tau}_B$ respectively.

Thus, in these two examples, $\hat{\tau}_{mo}$ has the smallest bias and a small root mean squared error. Note also that the jackknife method provides reasonable variance estimates, close to the root mean squared error estimated in the simulations.

5.5 Discussion

In this paper, we derived an Horvitz-Thompson-type estimator for Kendall's tau under bivariate censoring. Under suitable regularity conditions, this estimator is consistent and asymptotically normal. This is confirmed by the simulation results in Tables 5.1, 5.2 and 5.3. Simulations conducted to compare the proposed estimator with alternatives show clearly that there is no uniformly best estimator for all situations. Still, except in the simulations of Table 5.4 the proposed modified Oakes estimator has the smallest bias and a relatively small root mean squared error. It appears to be the best estimator available if the censoring is either independent, on X only, or univariate. Applying IPCW procedures to derive estimating equations in semi-parametric models is a promising research direction.

Acknowledgements

Partial funding for this work was provided by the Natural Sciences and Engineering Research Council of Canada and the Fonds québécois de la recherche sur la nature et les technologies.

Appendix A : A U-statistics expression for $\sqrt{n}(\tilde{\tau} - \tau)$

It is clear that $\tilde{\tau}$ is a U-statistics of order 2 whose expectation is equal to

$$\begin{aligned} E(\tilde{\tau}) &= \binom{n}{2}^{-1} \sum_{i<j} E\left\{ \frac{L_{ij}}{p_{ij}} a_{ij} b_{ij} \right\} \\ &= \binom{n}{2}^{-1} \sum_{i<j} E\left\{ E[\frac{L_{ij}}{p_{ij}} a_{ij} b_{ij} | \tilde{X}_{ij}, \tilde{Y}_{ij}] \right\} \\ &= \binom{n}{2}^{-1} \sum_{i<j} E\left\{ \frac{1}{p_{ij}} E[L_{ij} a_{ij} b_{ij} | \tilde{X}_{ij}, \tilde{Y}_{ij}] \right\}. \end{aligned}$$

By Oakes (1986), the concordance/discordance status and the orderability event are unconditionally independent. Once $\tilde{X}_{ij}$ and $\tilde{Y}_{ij}$ are fixed, by (5.12) the orderability event depends only on the censoring variables C_X and C_Y, while the concordance/discordance status depends only on the original pairs and hence these events are conditionally independent. So we have

$$E\{L_{ij}a_{ij}b_{ij}|\tilde{X}_{ij},\tilde{Y}_{ij}\} = E\{L_{ij}|\tilde{X}_{ij},\tilde{Y}_{ij}\}E\{a_{ij}b_{ij}|\tilde{X}_{ij},\tilde{Y}_{ij}\}$$

and, if $p_{ij} > 0$ for all pairs,

$$\begin{aligned} E(\tilde{\tau}) &= \binom{n}{2}^{-1}\sum_{i<j} E\{E[a_{ij}b_{ij}|\tilde{X}_{ij},\tilde{Y}_{ij}]\} \\ &= \binom{n}{2}^{-1}\sum_{i<j} E\{a_{ij}b_{ij}\} = \tau \end{aligned}$$

And thus $\sqrt{n}(\tilde{\tau}-\tau)$ is a zero mean U-statistics of order 2.

Appendix B : A U-statistics expression for $\sqrt{n}(\hat{\tau}-\tilde{\tau})$

For $k = 1,\cdots,n$, let $P_k = (\tilde{X}_k, \tilde{Y}_k, \delta_X^k, \delta_Y^k)$. For the special cases presented in Section 5.2.2, the asymptotic expansion

$$\sqrt{n}\left(\frac{1}{\hat{p}_{ij}} - \frac{1}{p_{ij}}\right) = \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\Phi_{ij}(P_k) + 0_p\left(\frac{1}{\sqrt{n}}\right)$$

holds for some function Φ_{ij} such that $E\{\Phi_{ij}(P_k)\} = 0$. This yields

$$\begin{aligned} \sqrt{n}(\hat{\tau}-\tilde{\tau}) &= \frac{1}{\binom{n}{2}}\sum_{i<j} L_{ij}a_{ij}b_{ij}\sqrt{n}\left(\frac{1}{\hat{p}_{ij}} - \frac{1}{p_{ij}}\right) \\ &\simeq \frac{1}{\binom{n}{2}}\sum_{i<j} L_{ij}a_{ij}b_{ij}\left\{\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\Phi_{ij}(P_k)\right\} \\ &= \frac{1}{\sqrt{n}\binom{n}{2}}\sum_{i<j}\sum_{k=1}^{n}\xi(P_i,P_j,P_k) \\ &= \frac{1}{\sqrt{n}\binom{n}{2}}\left\{\sum_{i<j}\xi_2(P_i,P_j) + \sum_{i<j<k}\xi_3(P_i,P_j,P_k)\right\}, \end{aligned} \tag{5.21}$$

where $\xi_2(P_i,P_j) = \xi(P_i,P_j,P_i)+\xi(P_i,P_j,P_j)$ and $\xi_3(P_i,P_j,P_k) = \xi(P_i,P_j,P_k)+\xi(P_i,P_k,P_j)+\xi(P_i,P_j,P_k)$.

Note that $\xi(\cdot,\cdot,\cdot)$ is symmetric with respect to its first two arguments. From (5.21), $\sqrt{n}(\hat{\tau}-\tilde{\tau})$ is asymptotically equivalent to a zero mean U-statistics of order 3 since $\xi_3(\cdot,\cdot,\cdot)$ is symmetric with respect to its arguments and since the term in $\xi_2(\cdot,\cdot)$ is negligible.

Transition

Le chapitre 5 décrit une nouvelle méthode d'estimation du tau de Kendall, τ, dans le contexte de censure bivariée. Si l'on choisit de modéliser la relation entre les deux variables à l'étude par une famille de copules, on peut alors estimer facilement le paramètre d'association α de cette famille à l'aide de la relation $\hat{\alpha} = g^{-1}(\hat{\tau})$, où

$$\tau = g(\alpha) = -1 + 4 \int_0^1 \int_0^1 C_\alpha(u, v) dC_\alpha(u, v).$$

On peut aussi estimer les fonctions de survie marginales S_X et S_Y en ayant recours à la méthode de Kaplan & Meier (1958).

La tâche d'estimer α, S_X et S_Y se complique énormément si seule la variable Y est sujette à la censure et si de plus, on ne peut observer que les paires respectant la condition $Y > X$. Si X et Y sont indépendantes, on peut estimer S_X et S_Y via les méthodes décrites dans Lynden-Bell (1971) et Lagakos et al. (1988). À l'opposé, si l'hypothèse d'indépendance entre X et Y est violée, alors la seule option possible pour l'estimation de α, S_X et S_Y est celle proposée par Lakhal-Chaieb et al. (2006). Les travaux du chapitre suivant tirent profit des estimations livrées par la procédure de Lakhal-Chaieb et al. (2006) afin de définir un tout nouveau critère de sélection de modèle en présence de troncation dépendante.

Chapitre 6

Archimedean copula model selection under dependent truncation

Résumé

Des données tronquées peuvent résulter de la présence d'une paire de durées de vie (X, Y) observée seulement si $X < Y$. Les méthodes d'analyse existantes sont fondées sur l'hypothèse de quasi-indépendance entre X et Y. Récemment, Lakhal-Chaieb et al. (2006) ont modélisé la dépendance potentielle entre ces variables aléatoires via une copule archimédienne de semi-survie. Dans cet article, nous présentons une procédure de sélection de modèle qui classe un ensemble de familles de copules archimédiennes de semi-survie selon leur niveau d'adéquation en regard d'un jeu de données particulier soumis à la troncation dépendante. La procédure proposée est basée sur la version tronquée du tau de Kendall (Manatunga & Oakes, 1996). La performance de notre méthode est illustrée à l'aide de simulations et de trois jeux de données réels.

Abstract

One-sided truncated survival data arise when a pair of time-to-event variables (X, Y) is observed only when $X < Y$. Existing methods of analysis rely on the assumption of quasi-independence between X and Y. Recently, Lakhal-Chaieb et al. (2006) modeled potential dependency between these random variables via a semi-survival Archimedean copula. In this paper, we present a model selection procedure to rank a set of semi-survival Archimedean copula families according to their ability to fit a given data set subject to dependent truncation. The proposed procedure is based on a truncated version of Kendall's tau (Manatunga & Oakes, 1996). The performance of the proposal is illustrated through simulations and three real data sets.

6.1 Introduction

In many applications of survival analysis, patients are subject to multiple correlated events. The dependence structure between the times of occurrence of these events is of prime interest for practitioners. This paper focuses on the selection of an appropriate dependence model for such data in the presence of truncation.

Truncated data arise when an individual is observed only if the failure time falls within a subject-specific truncation set. Left and right truncation times usually denote the lower and upper limits of the truncation set, respectively. When both limits are finite, the data are said to be doubly-truncated. Truncation is said to be *one-sided* when the data are left or right truncated depending on the variable of interest.

In the Canadian prevalence study of dementia, for example, only individuals who were diagnosed with the disease and who are still alive at recruitment time are included in the sample. The time from onset of the disease until death is thus left-truncated by the time from onset until the recruitment date. Similarly, in the famous Channing data set originally considered by Hyde (1977), men who did not live long enough to enter the retirement house in Palo Alto, CA, are excluded from the study. In this case, age at death is left-truncated by the age at entry in the retirement center. The patients are also subject to censoring if they were lost to follow-up or survived past the end of the study.

In this paper, we focus on one-sided cases. We consider pairs (X, Y) of time-to-event random variables which can be observed only if $X < Y$. These variables are said to be left-truncated if Y is the variable of interest and right-truncated otherwise. Many authors derived statistical techniques to handle such data. Procedures developed by these authors include survival estimation (Turnbull, 1976; Efron & Petrosian, 1999), linear regression (Bhattacharya et al., 1983; Tsui et al., 1988) and hazard rate modeling (Alioum & Commenges, 1996). These various methods rely on the quasi-independence assumption, i.e., the independence between X and Y in the observable region $X < Y$.

In many situations, however, the quasi-independence assumption is questionable. For example, people who were diagnosed in recent years with dementia may tend to live longer, due to improvements in the medical practice. Dependence could also be induced if people who enter a retirement home at a younger age receive better medical attention, which may induce increased longevity. Tsai (1990) developed a test of the quasi-independence assumption based on a conditional version of Kendall's tau ; see Martin & Betensky (2005) for an extension.

Truncation is a common phenomenon in bivariate survival analysis. In the last decade, copulas have become a popular tool in modeling the dependence between a pair of time-to-event variables ; recent contributions include Shih & Louis (1995), Zheng & Klein (1995), Rivest & Wells (2001), and Fine et al. (2001). Recently, Lakhal-Chaieb et al. (2006) modeled potential dependency between X and Y via a semi-survival Archimedean copula C. Under their model, the joint bivariate probability $\pi(x, y) = \Pr(X \leq x, Y > y | Y > X)$ is written as

$$\pi(x, y) = \frac{C_\alpha\{F_X(x), S_Y(y)\}}{c}, \quad y \geq x \tag{6.1}$$

where c is a normalizing constant, $1 - F_X$ and S_Y are survival functions related to the marginal behavior of X and Y respectively, and C_α is an Archimedean copula defined by

$$C_\alpha(u, v) = \phi_\alpha^{-1}\{\phi_\alpha(u) + \phi_\alpha(v)\}, \quad 0 \leq u, v \leq 1.$$

In order for the latter to be a distribution function with uniform margins, the copula generator ϕ_α must be a convex decreasing function such that $\phi_\alpha(1) = 0$; see, e.g., Genest & MacKay (1986). The resulting copula model involves a parameter α that measures the dependence between X and Y in the observable region $X < Y$.

It is desirable to interpret $1 - F_X$ and S_Y as the survival functions of X and Y, respectively. Unfortunately, this is not true in general. Indeed, the marginal functions of X and Y depend on both observable and unobservable regions. Hence, additional assumptions regarding the unobservable region are required to have the desired interpretations for $1 - F_X$ and S_Y. A simple case where these conditions are fulfilled is when (6.1) holds in the whole positive quadrant, in which case we also have $c = \Pr(Y > X)$. Even when these conditions fail, Lakhal-Chaieb et al. (2006) show that model (6.1) is still useful because it allows for the computation of conditional probabilities such as $\Pr(Y > y | X = x)$. Also, comparisons between groups based on F_X or S_Y can be made under this model.

Lakhal-Chaieb et al. (2006) studied inference procedures for α, c, S_Y and F_X under the above setting. For convenience, they restricted their analysis to the case where C_α belongs to Frank's family of copulas (Genest, 1987). Other Archimedean copulas could be used, however, and a natural question is : Which Archimedean copula family is the most suitable to a given data set ?

The main goal of this paper is to address this question and to present a model selection procedure to classify a set of copula families according to their ability to fit a given data set subject to dependent truncation. Several Archimedean copula selection and goodness-of-fit procedures already exist for complete data ; see Genest et al.

(2007) for a review. A few procedures also exist for censored observations, most notably Wang & Wells (2000b), Chen & Bandeen-Roche (2005), and Andersen et al. (2005). Unfortunately, none of these methods can be applied directly to data subject to dependent truncation, where observations fall in the upper wedge. The need to derive specific procedures thus exists. In this paper, we derive such a procedure, based on the truncated Kendall's tau due to Manatunga & Oakes (1996).

Model selection procedures should be distinguished from goodness-of-fit tests. A model selection procedure consists on ranking models according to their statistic values, whereas a goodness-of-fit test consists of determining whether a specific model fits the data in the light of a P-value. Such a P-value can be difficult to obtain, especially when the sampling scheme is complex. In particular, Genest et al. (2006) show that the P-value derived by Wang & Wells (2000b) for bivariate censored data is not valid.

In many situations, data are also subject to independent censoring in addition to dependent truncation. Lakhal-Chaieb et al. (2006) extended their model and inference procedures to this case but noted that this extension does not work well numerically. The secondary goal of this paper is to address this issue. Indeed, simulations show the good performance of the inference procedures in presence of censoring.

The rest of this paper is organized as follows. In Section 6.2, we present the model and estimation procedures of Lakhal-Chaieb et al. (2006). In Section 6.3, we derive the model selection procedure and in Section 6.4, we present numerical investigations. Concluding comments can be found in Section 6.5.

6.2 Data, model and statistical inference

Consider the general case where observations are subject to independent censoring in addition to dependent truncation. Under this setting, the data set consists of n independent replications of the observable variables (X, Z, δ), where X is the left-truncation time, $Z = \min(Y, C)$, Y is the failure time, C is the censoring variable and δ is the censoring indicator defined by $\delta = 1(Y < C)$.

Note that a standard setting for such data does not exist. Indeed, Martin & Betensky (2005) assume that the failure and censoring times are conditionally independent given the truncation time and the observable region, while Tsai (1990) assumes that the failure and censoring times are conditionally independent given only the truncation time. Both papers also assume that truncation precedes censoring with probability one and hence that the events $X < Y$ and $X < Z$ are identical. In this paper, we relax this assumption and simply assume that C and (X, Y) are independent given the observable region, denoted by $C \perp (X, Y)|(X < Z)$.

Under this setting and in the presence of censoring, model (6.1) becomes

$$\begin{aligned}\pi(x, y) &= \Pr(X \leq x, Z > y | Z > X) \\ &= \frac{S_C(y)\phi_\alpha^{-1}[\phi_\alpha\{F_X(x)\} + \phi_\alpha\{S_Y(y)\}]}{c}, \qquad (6.2)\end{aligned}$$

where $S_C(t) = \Pr(C > t|Z > X)$ is the conditional survival function of the censoring variable in the observable region.

Lakhal-Chaieb et al. (2006) derive estimation equations for α and c, namely

$$\underline{\mathrm{H}}(\alpha, c) = \begin{pmatrix} H_1(\alpha, c) \\ H_2(\alpha, c) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \qquad (6.3)$$

Denote by $(\hat{\alpha}, \hat{c})$ the solutions of these equations. Marginal distributions are then estimated by

$$\begin{aligned}\phi_{\hat{\alpha}}\{\hat{S}_Y(t, \hat{\alpha}, \hat{c})\} &= -\sum_{z_i \leq t, \delta_i = 1} \left[\phi_{\hat{\alpha}}\left\{\frac{\hat{c}\tilde{R}(z_i)}{n\hat{S}_C(z_i)}\right\} - \phi_{\hat{\alpha}}\left\{\frac{\hat{c}(\tilde{R}(z_i) - 1)}{n\hat{S}_C(z_i)}\right\}\right], \\ \phi_{\hat{\alpha}}\{\hat{F}_X(t, \hat{\alpha}, \hat{c})\} &= \sum_{x_i \geq t} \left[\phi_{\hat{\alpha}}\left\{\frac{\hat{c}\tilde{R}(x_i)}{n\hat{S}_C(x_i)}\right\} - \phi_{\hat{\alpha}}\left\{\frac{\hat{c}(\tilde{R}(x_i) - 1)}{n\hat{S}_C(x_i)}\right\}\right], \qquad (6.4)\end{aligned}$$

where $\tilde{R}(t)$ is the risk set at time t and $\hat{S}_C(.)$ is an estimate of $S_C(.)$.

They claim that their estimation procedure requires a large sample size ($n > 500$) to obtain reasonable variances for $\hat{\alpha}$ and $\hat{c}$, and therefore for $\hat{F}_X$ and $\hat{S}_Y$. As will be seen in Section 6.4.1, however, this assertion is not true. In fact, the method works very well even when the sample size gets as low as 100.

6.3 Model selection procedure

Let t be a fixed time. Define by $\tau^*(t)$ the truncated version of Kendall's tau as defined by Manatunga & Oakes (1996). In other words, the computation of $\tau^*(t)$ is restricted to individuals satisfying $A_t = \{X \leq t, Y > t\}$. The model selection procedure presented here consists on comparing a model-based and a nonparametric estimator of $\tau^*(t)$.

6.3.1 A model-based estimator

For all $t > 0$, the set A_t is included in the observable region $X < Z$. Thus a slight modification of equation (4) in Manatunga & Oakes (1996) yields

$$\tau^*(t, \alpha, c) = \tau^*(v_t, \alpha) = 1 + \frac{4}{v_t^2} \int_0^{v_t} \frac{\phi_\alpha(s) - \phi_\alpha(v_t)}{\phi'_\alpha(s)} \, ds, \tag{6.5}$$

where $v_t = c\pi(t,t)/S_C(t)$.

Let $(\hat{\alpha}, \hat{c})$ be the estimator of (α, c) given by Lakhal-Chaieb et al. (2006). Let also

$$\hat{\pi}(x, y) = \sum_{i=1}^{n} 1(X_i \le x, Z_i > y)/n,$$

and denote by $\hat{S}_C$ the Kaplan–Meier estimator with delayed entry for C based on $\{X_i, Z_i, 1 - \delta_i, i = 1, \ldots, n\}$. Finally, let $\hat{v}_t$ be the pseudo-observation defined by $\hat{v}_t = \hat{c}\hat{\pi}(t,t)/\hat{S}_C(t)$.

Substituting (α, v_t) in (6.5) by $(\hat{\alpha}, \hat{v}_t)$ yields a model-based estimator $\tau^*(t, \hat{\alpha}, \hat{c})$ for $\tau^*(t)$.

6.3.2 A nonparametric estimator

The nonparametric estimation of tau in the presence of censoring is a non-trivial problem ; see Beaudoin et al. (2007) for a review and new proposals. Indeed, in the presence of censoring, the concordance/discordance status can be established only for “orderable” pairs.

In the present context, an additional difficulty arises. Consider a point satisfying $X < Z \le t$ with $\delta = 0$. It is not clear whether this point belongs to A_t or not. Denote by $\tilde{A}_t$ the set of points satisfying $X \le t$ and $Z > t$. Note that $\tilde{A}_t \subseteq A_t$. Chen & Bandeen-Roche (2005) faced the same problem while devising a nonparametric estimator for Kendall’s tau restricted to individuals in the upper-left corner. They proposed to consider only observations for which the concordance/discordance status is fully determined.

In this paper, we adopt the same solution and estimate $\tau^*(t)$ nonparametrically by

$$\hat{\tau}^*(t) = \frac{C_t - D_t}{C_t + D_t}\,, \tag{6.6}$$

where C_t (respectively D_t) is the number of concordant (respectively discordant) "orderable" pairs among individuals satisfying $\tilde{A}_t$.

6.3.3 Model selection procedure

Several metrics can be used to measure the distance between the model-based and the nonparametric estimators of $\tau^*(t)$. A natural choice is the weighted L^2-norm distance

$$S_n(\hat{\alpha}, \hat{c}) = \int_H \hat{w}(t)\{\tau^*(t, \hat{\alpha}, \hat{c}) - \hat{\tau}^*(t)\}^2 dt, \tag{6.7}$$

where $H = \{t : v_t > 0\}$ and $\hat{w}$ is an estimator of a weight function w. Quantities inside the integral are bounded, which insures the integrability of the right-hand side of (6.7).

Given a data set, one can estimate α and c following Lakhal-Chaieb et al. (2006) and compute $S_n(\hat{\alpha}, \hat{c})$ according to (6.7) under several Archimedean copula families. These values can then be used to rank the copula families according to their ability to fit the data at hand.

6.4 Numerical investigations

6.4.1 Simulations

Estimation procedures

Simulations were carried out to evaluate the performance of the estimation procedure derived by Lakhal-Chaieb et al. (2006).

Triplets of observable variables (X, Z, δ) were simulated according to (6.2) with Frank's copula. The simulations were performed using two sample sizes ($n = 100, 250$), two values for c $(0.6, 0.8)$, two censoring fractions $(20\%, 40\%)$ and three values for the copula parameter ($\alpha = 2.372, 5.736, 14.139$) corresponding to an unconditional Kendall's tau of $0.25, 0.50, 0.75$. Marginals had exponential distributions with parameters controlling the values of c and the censoring fraction for a given copula parameter.

In Table 6.1, we report the mean and the mean squared error (MSE) for $\hat{c}$ and $\tau(\hat{\alpha})$

based on 400 replications of each scenario. These results show the good performance of the estimation method reported by Lakhal-Chaieb et al. (2006), even with moderate sample sizes ($n = 100, 250$). This contradicts the claim Lakhal-Chaieb et al. (2006) made about the performance of this method in the presence of censoring. This was due to programming errors on their part. Hence, the model selection procedure will use the estimators $\hat{\alpha}$ and $\hat{c}$ provided by this estimating procedure.

Model selection

The model selection procedure consists in estimating α and c by the inference method of Lakhal-Chaieb et al. (2006) and then computing the statistic $S_n(\hat{\alpha}, \hat{c})$ given by (6.7). In our simulation study, we set $w(t)$ to be proportional to the probability of the event A_t. It is estimated by $\hat{w}(t) = \hat{\pi}(t,t)/\hat{S}_C(t)$. We consider a set of five semi-survival Archimedean copulas numbered 1 to 5. They are presented in Table 6.2.

Simulations were performed using a single sample size ($n = 200$), a single value for the normalizing constant ($c = 0.8$), three values for τ $(0.25, 0.50, 0.75)$ and two values for the censoring fraction $(20\%, 40\%)$. For each combination of the conditions above, 1000 samples were simulated under each copula model. The statistic $S_n(\hat{\alpha}, \hat{c})$ was then computed for every sample under the various dependence structures. The copula families were then ranked according to the values of the statistic. Finally, a count was made of the number of times that each copula family was declared the best-fitting model. Results are reported in Table 6.3.

As expected, some copula models are easier to detect than others. For instance, the model selection procedure has a good ability to recognize copula families 1 and 2. The correct detection rate for models 3 and 4 lies between 60% and 70%, while dependence structures 4 and 5 are hard to distinguish. The performance of the model selection procedure improves with the dependency level under model 1. This is expected as it is known that this copula model can be confused with model 2 near independence. Surprisingly, the converse is true for copula families 2 and 3. For the remaining copula models, the dependence level does not seem to have an impact on the performance of the model selection procedure. Finally, the censoring fraction does not seem to affect the performance of the proposal, whatever the dependence structure.

Some care must be taken in interpreting these results, as they should not be compared to those obtained from goodness-of-fit tests. The performance of the proposal depends on the set of copula families. If the latter contains models which differ greatly from each other, the method can look performant. For instance, copula families 4 and 5 are very similar. Removing model 4 from the set would dramatically improve the ability of the model selection procedure to detect copula family 5.

Generally speaking, a look at the values of the statistic $S_n(\hat{\alpha}, \hat{c})$ can be informative. If they are very close for a given data set under two or more copula models, conclusions can be difficult to reach. Of course, the difference between the models may not be crucial. Indeed, several copula models having close values for $S_n(\hat{\alpha}, \hat{c})$ are very likely to yield similar estimators for S_Y and F_X and hence for probabilities written in terms of these functionals, such as $\Pr(Y > y|X = x)$.

6.4.2 Examples

The Channing house

We consider the Channing house retirement community data set from Hyde (1977). Truncation and failure times are respectively the retirement center entry age and the age at death. Censoring is due to withdrawal from the community or the end of the study. This data set consists of 97 males, of which 51 are censored. The nonparametric test of Tsai (1990) yields $Z_{\text{obs}} = 2.02$ (P-value $= 0.04$) providing evidence against quasi-independence. Two observations were removed because they died in the community before any other individual entered the risk set (Klein & Moeschberger (1997) discuss this issue in Chapter 4 of their book for the Lyndell-Bell estimator).

We estimated the parameters c and α using each of the five copula models presented in Table 6.2. The corresponding values of $S_n(\hat{\alpha}, \hat{c})$ are equal to 0.0639, 0.0651, 0.0959 and 0.1307 for models 1, 2, 3 and 4, respectively. Therefore, the model selection procedure suggests that copula family 1 (Frank's copula) is the best fitting-model for the data set at hand. The estimation procedure developed by Lakhal-Chaieb et al. (2006) did not converge under model 5. As a result, the statistic $S_n(\hat{\alpha}, \hat{c})$ could not be evaluated in this case. The development of a better estimation method would positively affect the performance of the model selection criterion presented in this paper.

Under Frank's model, we obtain $\tau(\hat{\alpha}) = 0.307$ and $\hat{c} = 0.515$. This suggests the presence of a moderate positive association between the age at entry and the age at death. A possible explanation could be that patients entering the retirement center at a relatively young age already have a deteriorated health, while people coming into the Channing house at an older age maintained a healthier lifestyle throughout the years.

The interpretation of S_Y as a proper survival function is somewhat questionable in this example. Indeed, S_Y relies on assumptions related to the association between X and Y in the unobserved region, which corresponds here to people entering the Channing house after their death. Instead, we focus our analysis on the observable region.

Let us consider the conditional survival function $S_Y(y|X = x)$, for $x \leq y$. This conditional probability can be estimated from (6.2). Easy algebra yields

$$S_Y(y|X = x) = \frac{S_C(y)}{c} \frac{\phi'_\alpha\{F_X(x)\}}{\phi'_\alpha\{\frac{c\pi(x,y)}{S_C(y)}\}}, \quad x \leq y. \tag{6.8}$$

An estimator for this conditional survival function is then obtained by plugging in estimators for unknown quantities in the relationship above. $S_Y(y|X = x)$ is then used to compute the residual lifetime $m(x) = \mathrm{E}(Y - x|X = x)$ once a person enters the Channing house.

In Figure 6.1, we present $\hat{m}$ with its associated confidence limits, along with the US population residual lifetime in 1971 obtained from the US Census Bureau web site (www.census.gov). These curves are fairly similar. Notice that the US population curve falls within the confidence bands on the whole domain (except perhaps at 65 years of age). This result is expected since, to our knowledge, there is no evidence that entering a retirement house has an impact on overall survival. Under independence, $m(x) = \mathrm{E}(Y) - x$ is a line with slope -1 (see Figure 6.1). Failure to account for dependence leads to severe underestimation of the residual lifetime.

Dementia

In 1991, Health Canada carried out a nationwide prevalence study on dementia as part of the *Canadian Study on Health and Aging* (www.csha.ca). This disease causes a progressive decline in cognitive functions such as memory, attention, language and problem solving. The study ended in 1996. A total of 10,263 Canadians aged 65 years or older were randomly selected across the country. Of these, 821 were diagnosed with dementia. The final data set is reduced to $n = 807$ observations because of typing errors in the original file.

Characterizing survival following the onset of this disease is becoming a very important issue, given that the number of elderly will only grow in the years to come due to the increase in life expectancy. This is a cross-sectional study because only people diagnosed with the disease at the beginning of the study, in 1991, were included in the

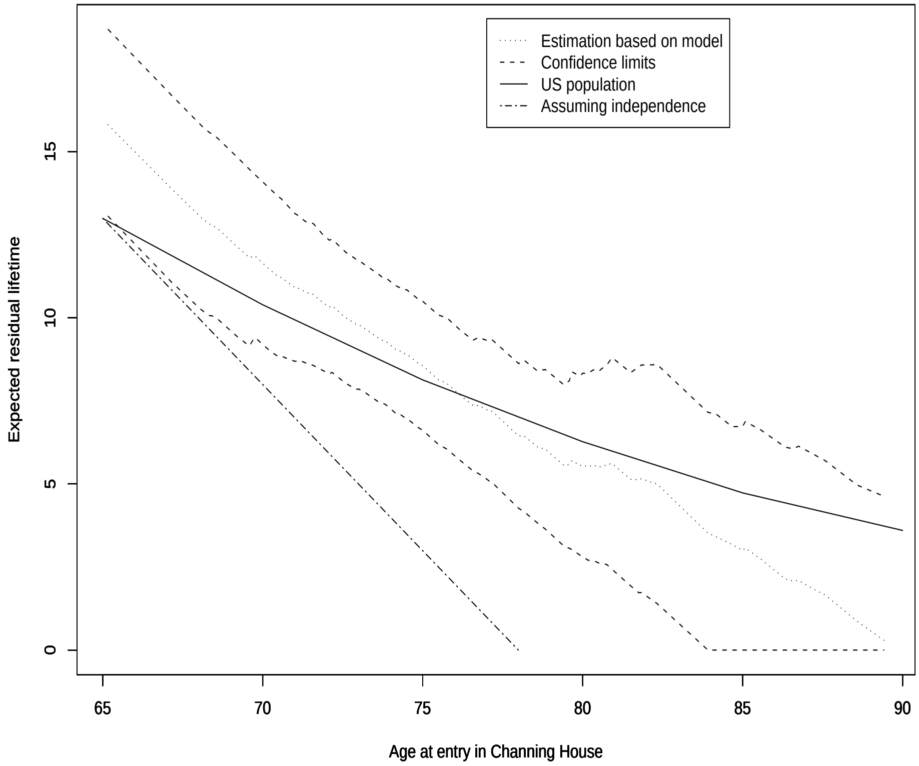


FIG. 6.1 – Residual lifetime using Frank's copula model for the Channing house data set. The US population curve and the remaining lifetime assuming independence between age at entry into the retirement center and age at death are also presented.

investigation. Times (in days) from onset of dementia to 1991 and death are truncation and failure times respectively. Subjects were censored if they were still alive at the end of the study or lost to follow-up. This affects 22% of the patients.

The nonparametric test of Tsai (1990) yields a P-value of 0.025, providing evidence against quasi-independence. Changes in medical practice over the years may explain this dependence. We fitted each of the copula models presented in Table 6.2 to the data at hand. We obtained $S_n(\hat{c}, \hat{\alpha})$ equal to 0.008886, 0.008925 and 0.009939 for copula families 2, 3 and 1, respectively. The algorithm proposed by Lakhal-Chaieb et al. (2006) did not converge under models 4 and 5. This suggests that the Clayton copula is the best-fitting model. The latter yields $\hat{c} = 0.140$ and $\tau(\hat{\alpha}) = 0.057$.

In this example, there is no reason for the dependence structures between failure and truncation times to differ before and after 1991. Thus, we can assume that (6.2) holds in both observable and unobservable regions and we thus interpret S_Y as the survival function of Y and c as $\Pr(X < Y)$. We obtained $\hat{c}$ below 0.3 under all fitted models. This suggests that the majority of observations were missed by the data collection process.

In order to show the impact of the copula model misspecification on survival estimation, we fitted the copula families giving the largest and smallest values for the statistic $S_n(\hat{\alpha}, \hat{c})$, namely the Clayton and Frank models respectively. Estimation of S_Y under both models is reported in Figure 6.2. The estimator under the Frank copula lays outside the confidence bands provided by the Clayton copula. Also, pointwise jackknife equality tests between the two curves at $Y = 1000$, 2000, 3000, 4000, 5000 and 6000 give $Z_{\text{obs}} = 4.92$, 5.25, 4.65, 3.63, 2.90 and 2.19 respectively. Hence, there is a significant difference between the two curves.

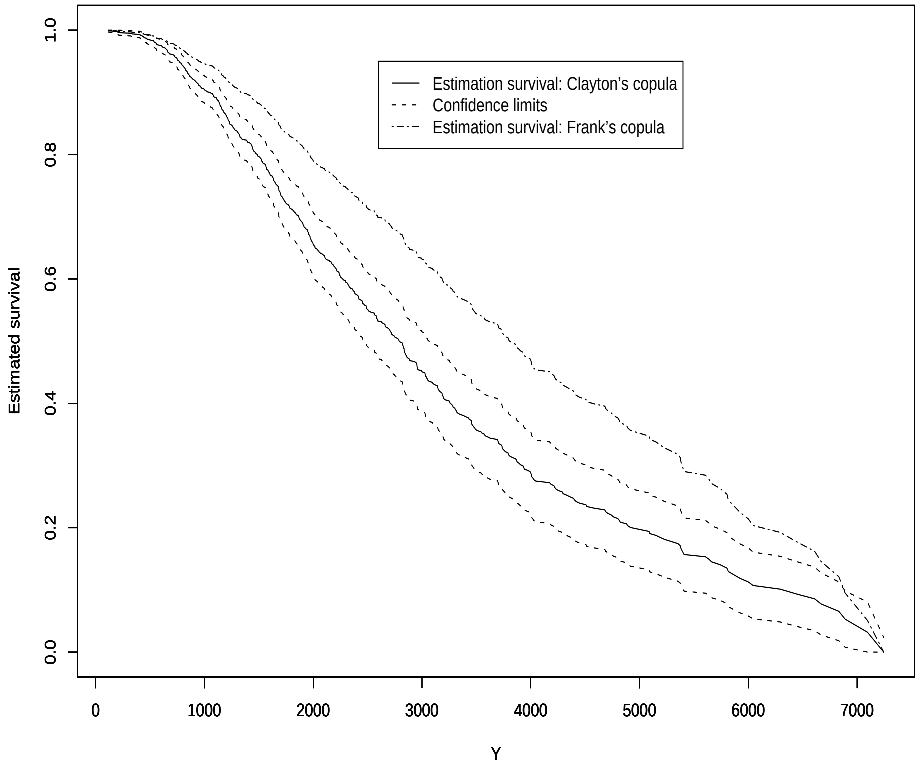


FIG. 6.2 – Estimated survival functions obtained from the Dementia data set using the Clayton and Frank models, along with Clayton's associated confidence limits.

Transfusion-related AIDS

Consider the transfusion-related AIDS data set provided by the Centers for Disease Control in Atlanta, GA. The study started on January 1978. Only individuals who developed the disease prior to July 1986 are ascertained for the study.

Denote by T and X the successive times in months from the start of the study until infection and from infection until the development of the disease respectively. Subjects are observed only if $T + X < 102$, e.g., if $X < Y$, where $Y = 102 - T$. Hence, the induction time X is right-truncated by Y. Tsai (1990) tested and rejected the quasi-independence assumption between X and Y.

We fitted each of the five copula models to this data set. The statistic values of $\hat{S}(\hat{c}, \hat{\alpha})$ are 0.004265, 0.004293 and 0.005874 for models 2, 3 and 1, respectively. The algorithm of Lakhal-Chaieb et al. (2006) did not converge for models 4 and 5. Model 2 (Clayton) seems to be the best-fitting model to this data set.

In order to show the impact of the copula model misspecification on conditional cumulative distribution function estimation, we fitted the copula families giving the largest and smallest values for the statistic $S_n(\hat{\alpha}, \hat{c})$, namely the Frank and Clayton models respectively. Estimation of $\Pr(X \leq x | X \leq x_0, Y = y_0)$ under both models, based only on the observable region, is reported in Figure 6.3 for $x_0 = 40$ and $y_0 = 80$. A pointwise jackknife equality test between the two curves at $x = 20$ yields $Z_{\text{obs}} = 2.16$. Hence, there is a significant difference between the two curves.

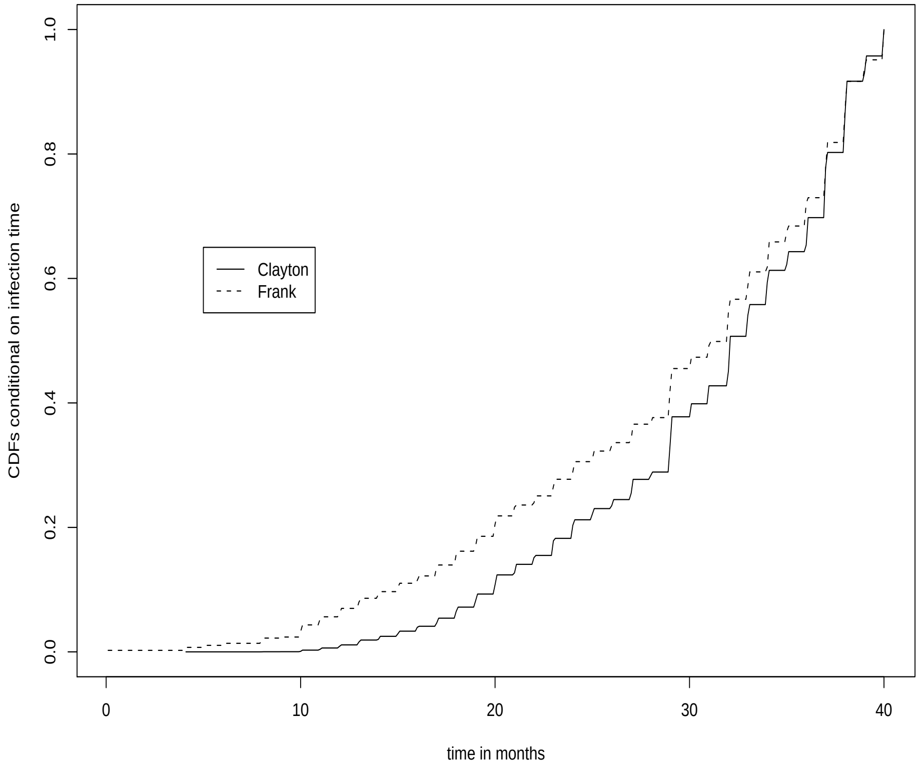


FIG. 6.3 – Estimation of $\Pr(X \leq x | X \leq 40, Y = 80)$ from the AIDS data set using the Clayton and Frank models.

6.5 Conclusion

In this paper, we present a simple model selection procedure which ranks a set of copula families according to their ability to fit a given data set. This method is evaluated through a simulation study and illustrated using three data sets.

The nonparametric estimation of the truncated Kendall's tau presented in Section 6.3.2 is somewhat ad hoc, and hence can be improved upon by adapting modern techniques of nonparametric estimation of Kendall's tau in presence of censoring ; see Beaudoin et al. (2007) and Lakhal-Chaieb et al. (2007b). It is not clear how this would improve the performance of the selection procedure.

The context of data subject to dependent truncation and independent censoring is quite complex because a significant piece of information is missing, whence the difficulty in providing a P-value with a formal goodness-of-fit test. Genest et al. (2006) pointed out that standard jackknife and bootstrap methods used by Wang & Wells (2000b) to estimate the variance of their statistics under the null hypothesis are not valid. An alternative parametric bootstrap developed by Genest et al. (2006) is available for the estimation of variances with complete observations ; its validity is established by Genest & Rémillard (2005). However, it is not clear how to adapt their method to estimate the variance of $S_n(\hat{\alpha}, \hat{c})$ under the null hypothesis. This is the object of current research.

Finally, the model selection procedure presented here can be adapted to other incomplete data schemes such as semi-competing risks (Fine et al., 2001; Lakhal-Chaieb et al., 2007a) and compared to other already existing procedures for such data.

Censoring fraction	c	τ	$\hat{c}$		$\tau(\hat{\alpha})$	
			$n = 100$	$n = 250$	$n = 100$	$n = 250$
20%	0.8	0.25	0.792	0.789	0.240	0.241
			(77.58)	(58.47)	(85.12)	(58.11)
		0.50	0.795	0.792	0.480	0.491
			(74.41)	(51.95)	(68.31)	(40.98)
		0.75	0.795	0.794	0.723	0.742
			(85.79)	(48.68)	(47.32)	(20.40)
	0.6	0.25	0.592	0.597	0.240	0.246
			(126.10)	(95.25)	(100.26)	(75.20)
		0.50	0.584	0.583	0.469	0.487
			(135.31)	(84.47)	(99.87)	(58.20)
		0.75	0.623	0.590	0.709	0.742
			(177.24)	(62.27)	(87.97)	(26.48)
40%	0.8	0.25	0.794	0.798	0.237	0.249
			(72.92)	(49.08)	(94.53)	(54.21)
		0.50	0.794	0.799	0.482	0.494
			(70.52)	(43.58)	(66.74)	(39.75)
		0.75	0.801	0.797	0.730	0.741
			(79.30)	(55.00)	(41.93)	(21.60)
	0.6	0.25	0.593	0.593	0.247	0.241
			(107.98)	(74.49)	(109.10)	(86.01)
		0.50	0.596	0.592	0.476	0.490
			(121.22)	(56.64)	(89.26)	(51.42)
		0.75	0.618	0.593	0.718	0.741
			(137.36)	(67.92)	(70.23)	(28.82)

Tab. 6.1 – Mean and 1000*square root(MSE) of $\hat{c}$ and $\tau(\hat{\alpha})$ by the estimation procedure of Lakhal-Chaieb et al. (2006).

No	$\phi_\alpha(t)$	Name
1	$-\log\{\frac{e^{-\alpha t}-1}{e^{-\alpha}-1}\}$	Frank
2	$\frac{t^{-\alpha}-1}{\alpha}$	Clayton
3	$(1-t^{1/\alpha})^{\alpha}$	
4	$-\log\{(1-\alpha)t+\alpha\}$	
5	$\frac{1-t}{1+(\alpha-1)t}$	

Tab. 6.2 – Five semi-survival Archimedean copula models considered in the model selection simulation study.

		Chosen copula									
		20% Censoring					40% Censoring				
τ	True copula	1	2	3	4	5	1	2	3	4	5
0.25	1	**839**	157	4	0	0	**805**	193	2	0	0
	2	0	**874**	126	0	0	5	**877**	113	5	0
	3	0	195	**773**	32	0	0	217	**684**	85	14
	4	0	9	256	**649**	86	0	57	371	**520**	52
	5	0	0	31	**563**	406	0	2	106	**638**	254
0.50	1	**945**	55	0	0	0	**953**	47	0	0	0
	2	0	**790**	209	1	0	0	**829**	171	0	0
	3	0	305	**654**	41	0	0	303	**651**	43	3
	4	0	0	137	**695**	168	0	6	287	**616**	91
	5	0	0	61	**620**	319	0	6	163	**655**	176
0.75	1	**998**	2	0	0	0	**996**	4	0	0	0
	2	0	**646**	348	6	0	0	**723**	275	2	0
	3	0	409	**495**	94	2	0	409	**489**	87	15
	4	0	13	166	**534**	287	0	37	266	**512**	185
	5	0	14	149	**503**	334	0	24	217	**532**	227

TAB. 6.3 – Selection frequency of the 5 copula models. The largest frequency is highlighted in boldface

Chapitre 7

Conclusion

Dans le premier article de la thèse, plusieurs tests d'adéquation de modèles de copules à deux dimensions ont été présentés. Ces procédures s'appliquent dans les situations où les données sont complètes. L'étude de Monte-Carlo que nous avons réalisée a notamment mis en valeur la bonne performance des tests fondés sur une statistique de type Cramér–von Mises. Par ailleurs, nos simulations ont également démontré l'importance d'identifier correctement la loi asymptotique d'une statistique donnée. À cet effet, les résultats désastreux du test décrit par Breymann et al. (2003) sont éloquents.

L'étude de tests d'adéquation de modèles de copules en présence de données censurées et/ou tronquées offre des perspectives intéressantes et naturelles de généralisation de ce thème de recherche. Le domaine est actuellement en pleine ébullition, comme en font foi les récents travaux de Chen & Bandeen-Roche (2005), ainsi que ceux de Andersen et al. (2005).

Pour sa part, le quatrième article de la thèse a innové dans le contexte de la troncation dépendante par l'introduction d'un critère de sélection de modèle archimédien. Ce dernier s'est d'ailleurs avéré très performant dans le cadre de notre étude de Monte-Carlo. Cette méthode permet de déterminer la copule archimédienne la mieux adaptée à un jeu de données quelconque. Contrairement aux tests de l'article 1, toutefois, la procédure ne garantit pas la qualité de l'adéquation. Afin d'y parvenir, il paraît naturel de procéder comme dans le premier article, c'est-à-dire d'implanter une procédure de bootstrap paramétrique en vue d'obtenir un seuil observé quantifiant plus formellement l'adéquation du modèle choisi. Malheureusement, nul n'est parvenu à ce jour à développer une procédure de bootstrap paramétrique en présence de censure et de troncation. Cette avenue mériterait d'être explorée dans des travaux futurs.

Les deuxième et troisième articles de la thèse ont décrit un ensemble de méthodes d'inférence sous divers schémas de censure. Dans les deux cas, une étude de simulation a permis de démontrer la supériorité de nos propositions par rapport aux estimateurs existants.

La performance des nouveaux estimateurs du tau de Kendall du deuxième article ne laisse planer aucun doute quant à leur valeur. Ceux-ci dépendent de façon critique de l'estimation de la loi conditionnelle $S_{Y|X}$ de Y sachant $X = x$. Aussi la précision de nos procédures d'estimation du tau de Kendall est-elle fortement liée au choix de $\hat{S}_{Y|X}$. Dans notre projet, nous avons opté pour les estimateurs $\hat{S}_{Y|X}$ de Leconte et al. (2002) et de Van Keilegom & Akritas (1999). Or, une étude de la performance de ces deux estimateurs nous a permis de découvrir qu'ils éprouvent beaucoup de difficulté à approcher la loi sous-jacente lorsque le niveau d'association entre les variables étudiées est fort. Ceci nous porte à croire que nos nouvelles méthodes d'estimation du tau de Kendall pourraient se révéler encore plus performantes si on parvenait à raffiner davantage l'estimation de $S_{Y|X}$.

Bibliographie

Abdous, B., Genest, C., & Rémillard, B. (2005). Dependence properties of meta-elliptical distributions. In *Statistical Modeling and Analysis for Complex Data Problems*, volume 1 of *GERAD 25th Anniv. Ser.*, pages 1–15. Springer, New York.

Alioum, A. & Commenges, D. (1996). A proportional hazards model for arbitrarily censored and truncated data. *Biometrics*, 52 :512–524.

Andersen, P. K., Ekstrøm, C. T., Klein, J. P., Shu, Y., & Zhang, M.-J. (2005). A class of goodness of fit tests for a copula based on bivariate right-censored data. *Biom. J.*, 47 :815–824.

Ané, T. & Kharoubi, C. (2003). Dependence structure and risk measure. *J. Business*, 76 :411–438.

Barbe, P., Genest, C., Ghoudi, K., & Rémillard, B. (1996). On Kendall's process. *J. Multivariate Anal.*, 58 :197–229.

Beaudoin, D., Duchesne, T., & Genest, C. (2007). Improving the estimation of Kendall's tau when censoring affects only one of the variables. *Comput. Statist. Data Anal.*, 50 :in press.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical report, University of California, Berkeley, CA.

Berg, D. & Bakken, H. (2005). A goodness-of-fit test for copulae based on the probability integral transform. Technical report SAMBA/41/05, Norsk Regnesentral, Oslo, Norway.

Bhattacharya, P. K., Chernoff, H., & Yang, S. S. (1983). Nonparametric estimation of the slope of a truncated regression. *Ann. Statist.*, 11(2) :505–514.

Braekers, R. & Veraverbeke, N. (2005). A copula-graphic estimator for the conditional survival function under dependent censoring. *Canad. J. Statist.*, 33 :429–447.

Breymann, W., Dias, A., & Embrechts, P. (2003). Dependence structures for multivariate high-frequency data in finance. *Quant. Finance*, 3 :1–14.

Brown, B., Hollander, M., & Korwar, R. M. (1974). Nonparametric tests of independence for censored data, with applications to heart transplant studies. *Reliability and Biometry*, pages 327–354.

Burzykowski, T., Molenberghs, G., & Buyse, M. (2004). The validation of surrogate end points by using data from randomized clinical trials : a case-study in advanced colorectal cancer. *J. Roy. Statist. Soc. Ser. A*, 167 :103–124.

Campbell, G. (1981). Nonparametric bivariate estimation with randomly censored data. *Biometrika*, 68(2) :417–422.

Chen, M.-C. & Bandeen-Roche, K. (2005). A diagnostic for association in bivariate survival models. *Lifetime Data Anal.*, 11(2) :245–264.

Chen, X., Fan, Y., & Tsyrennikov, V. (2006). Efficient estimation of semiparametric multivariate copula models. *J. Amer. Statist. Assoc.*, 101 :1228–1240.

Cherubini, U., Luciano, E., & Vecchiato, W. (2004). *Copula Methods in Finance.* Wiley, New York.

Choi, Y.-H. & Matthews, D. E. (2005). Accelerated life regression modelling of dependent bivariate time-to-event data. *Canad. J. Statist.*, 33 :449–464.

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

Cook, R. D. & Johnson, M. E. (1981). A family of distributions for modelling nonelliptically symmetric multivariate data. *J. Roy. Statist. Soc. Ser. B*, 43 :210–218.

Cui, S. & Sun, Y. (2004). Checking for the gamma frailty distribution under the marginal proportional hazards frailty model. *Statist. Sinica*, 14 :249–267.

Dabrowska, D. M. (1987). Nonparametric regression with censored survival time data. *Scand. J. Statist.*, 14 :181–197.

Dabrowska, D. M. (1988). Kaplan-Meier estimate on the plane. *Ann. Statist.*, 16 :1475–1489.

Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional Kaplan–Meier estimate. *Ann. Statist.*, 17 :1157–1167.

Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés : Un test non paramétrique d'indépendance. *Acad. Royal Bel., Bull. Class. Sci., 5^{e} série*, 65 :274–292.

Diaconis, P., Graham, R., & Holmes, S. P. (2001). Statistical problems involving permutations with restricted positions. In *State of the art in probability and statistics (Leiden, 1999)*, volume 36 of *IMS Lecture Notes Monograph Series*, pages 195–222. Institute of Mathematical Statistics, Beachwood, OH.

Dobrić, J. & Schmid, F. (2005). Testing goodness of fit for parametric families of copulas : Application to financial data. *Comm. Statist. Simulation Comput.*, 34 :1053–1068.

Dobrić, J. & Schmid, F. (2007). A goodness of fit test for copulas based on Rosenblatt's transformation. *Comput. Statist. Data Anal.*, 52 :in press.

Drouet-Mari, D. & Kotz, S. (2001). *Correlation and Dependence.* Imperial College Press, London.

Efron, B. & Petrosian, V. (1999). Nonparametric methods for doubly truncated data. *J. Amer. Statist. Assoc.*, 94(447) :824–834.

Embrechts, P., McNeil, A. J., & Straumann, D. (2002). Correlation and dependence in risk management : properties and pitfalls. In *Risk management : value at risk and beyond (Cambridge, 1998)*, pages 176–223. Cambridge Univ. Press, Cambridge.

Fang, H.-B., Fang, K.-T., & Kotz, S. (2002). The meta-elliptical distributions with given marginals. *J. Multivariate Anal.*, 82 :1–16.

Fang, H.-B., Fang, K.-T., & Kotz, S. (2005). Corrigendum to : "The meta-elliptical distributions with given marginals" [J. Multivariate Anal. 82 : 1–16 (2002)]. *J. Multivariate Anal.*, 94 :222–223.

Faraggi, D. & Korn, E. L. (1996). Competing risks with frailty models when treatment affects only one failure type. *Biometrika*, 83 :467–471.

Fermanian, J.-D. (2005). Goodness-of-fit tests for copulas. *J. Multivariate Anal.*, 95 :119–152.

Fermanian, J.-D., Radulović, D., & Wegkamp, M. J. (2004). Weak convergence of empirical copula processes. *Bernoulli*, 10 :847–860.

Fine, J. P., Jiang, H., & Chappell, R. (2001). On semi-competing risks data. *Biometrika*, 88(4) :907–919.

Frahm, G., Junker, M., & Szimayer, A. (2003). Elliptical copulas : applicability and limitations. *Statist. Probab. Lett.*, 63 :275–286.

Frank, M. J. (1979). On the simultaneous associativity of $F(x, y)$ and $x+y-F(x, y)$. *Aequationes Math.*, 19 :194–226.

Frees, E. W. & Valdez, E. A. (1998). Understanding relationships using copulas. *N. Am. Actuar. J.*, 2 :1–25.

Gänßler, P. & Stute, W. (1987). *Seminar on Empirical Processes*. Birkhäuser Verlag, Basel, Switzerland.

Genest, C. (1987). Frank's family of bivariate distributions. *Biometrika*, 74 :549–555.

Genest, C. & Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *J. Hydrol. Eng.*, 12 :in press.

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

Genest, C. & MacKay, R. J. (1986). Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données. *Canad. J. Statist.*, 14 :145–159.

Genest, C., Quessy, J.-F., & Rémillard, B. (2006). Goodness-of-fit procedures for copula models based on the probability integral transformation. *Scand. J. Statist.*, 33 :337–366.

Genest, C. & Rémillard, B. (2005). Validity of the parametric bootsrap for goodness-of-fit testing in semiparametric models. Technical Report G–2005–51, GERAD, Montréal, Canada.

Genest, C., Rémillard, B., & Beaudoin, D. (2007). Omnibus goodness-of-fit tests for copulas : A review and a power study. *Insurance : Mathematics & Economics*, 41 :in press.

Genest, C. & Rivest, L.-P. (1993). Statistical inference procedures for bivariate Archimedean copulas. *J. Amer. Statist. Assoc.*, 88 :1034–1043.

Genest, C. & Verret, F. (2005). Locally most powerful rank tests of independence for copula models. *J. Nonparametr. Stat.*, 17 :521–539.

Genest, C. & Werker, B. J. M. (2002). Conditions for the asymptotic semiparametric efficiency of an omnibus estimator of dependence parameters in copula models. In *Distributions with Given Marginals and Statistical Modelling*, pages 103–112. Kluwer, Dordrecht, The Netherlands.

Ghoudi, K. & Rémillard, B. (1998). Empirical processes based on pseudo-observations. In *Asymptotic Methods in Probability and Statistics (Ottawa, ON, 1997)*, pages 171–197. North-Holland, Amsterdam.

Ghoudi, K. & Rémillard, B. (2004). Empirical processes based on pseudo-observations. II. The multivariate case. In *Asymptotic Methods in Stochastics*, volume 44 of *Fields Inst. Commun.*, pages 381–406. Amer. Math. Soc., Providence, RI.

Gibbons, J. D. (1971). *Nonparametric statistical inference.* McGraw-Hill Book Co., New York.

Glidden, D. V. (1999). Checking the adequacy of the gamma frailty model for multivariate failure times. *Biometrika*, 86 :381–393.

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

Hoeffding, W. (1947). On the distribution of the rank correlation coefficient τ when the variates are not independent. *Biometrika*, 34 :183–196.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, 47 :663–685.

Hougaard, P., Harvald, B., & Holm, N. V. (1992). Measuring the similarities between the lifetimes of adult Danish twins born between 1881-1930. *J. Amer. Statist. Assoc.*, 87 :17–24.

Hyde, J. (1977). Testing survival under right censoring and left truncation. *Biometrika*, 64(2) :225–230.

Isobe, T., Feigelson, Eric, D., & Nelson, P. I. (1986). Statistical methods for astronomical data with upper limits : Ii. correlation and regression. *The American Astronomical Society*, 306 :490–507.

Joe, H. (1997). *Multivariate Models and Dependence Concepts.* Chapman & Hall, London.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *J. Multivariate Anal.*, 94 :401–419.

Jouini, M. N. & Clemen, R. T. (1996). Copula models for aggregating expert opinions. *Oper. Res.*, 44 :444–457.

Junker, M. & May, A. (2005). Measurement of aggregate risk with copulas. *Econom. J.*, 8 :428–454.

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, 53 :457–481.

Kendall, M. & Gibbons, J. D. (1990). *Rank correlation methods.* A Charles Griffin Title. Edward Arnold, London, fifth edition.

Kendall, M. G. (1970). *Rank Correlation Methods.* Charles Griffin, London.

Klaassen, C. A. J. & Wellner, J. A. (1997). Efficient estimation in the bivariate normal copula model : Normal margins are least favourable. *Bernoulli*, 3 :55–77.

Klein, J. P. & Moeschberger, M. L. (1997). *Survival Analysis : Techniques for Censored and Truncated Data.* Springer, New York.

Klugman, S. & Parsa, R. (1999). Fitting bivariate loss distributions with copulas. *Insurance Math. Econom.*, 24 :139–148.

Kruskal, W. H. (1958). Ordinal measures of association. *J. Amer. Statist. Assoc.*, 53 :814–861.

Lagakos, S. W., Barraj, L. M., & De Gruttola, V. (1988). Nonparametric analysis of truncated survival data, with applications to AIDS. *Biometrika*, 75(3) :515–523.

Lakhal-Chaieb, L., Rivest, L.-P., & Abdous, B. (2006). Estimating survival under a dependent truncation. *Biometrika*, 93 :655–669.

Lakhal-Chaieb, L., Rivest, L.-P., & Abdous, B. (2007a). Estimating association and survival in a semi-competing risks model. *Biometrics*, page in press.

Lakhal-Chaieb, L., Rivest, L.-P., & Beaudoin, D. (2007b). IPCW estimator for Kendall's tau under bivariate censoring. Technical report, Département de mathématiques et de statistique, Université Laval, Québec, Canada.

Leconte, E., Poiraud-Casanova, S., & Thomas-Agnan, C. (2002). Smooth conditional distribution function and quantiles under random censorship. *Lifetime Data Anal.*, 8 :229–246.

Lim, J. (2006). Permutation procedures with censored data. *Comput. Statist. Data Anal.*, 50 :332–345.

Lin, D. Y., Sun, W., & Ying, Z. (1999). Nonparametric estimation of the gap time distributions for serial events with censored data. *Biometrika*, 86(1) :59–70.

Lynden-Bell, D. (1971). A method of allowing for known observational selection in small samples applied to 3cr quasars. *Mon. Not. R. Astr. Soc.*, 155 :95–118.

Malevergne, Y. & Sornette, D. (2003). Testing the Gaussian copula hypothesis for financial assets dependences. *Quant. Finance*, 3 :231–250.

Manatunga, A. K. & Oakes, D. (1996). A measure of association for bivariate frailty distributions. *J. Multivariate Anal.*, 56 :60–74.

Mardia, K. V. (1970). *Families of Bivariate Distributions*. Hafner, Darien, CT.

Martin, E. C. & Betensky, R. A. (2005). Testing quasi-independence of failure and truncation times via conditional Kendall's tau. *J. Amer. Statist. Assoc.*, 100(470) :484–492.

McGilchrist, C. & Aisbett, C. (1991). Regression with frailty in survival analysis. *Biometrics*, 47 :461–466.

McNeil, A. J., Frey, R., & Embrechts, P. (2005). *Quantitative Risk Management*. Princeton University Press, Princeton, NJ.

Mikosch, T. (2006). Copulas : Tales and facts. *Extremes*, 9 :3–20.

Miller, R. & Halpern, J. (1982). Regression with censored data. *Biometrika*, 69(3) :521–531.

Nelsen, R. B. (1986). Properties of a one-parameter family of bivariate distributions with specified marginals. *Comm. Statist. A—Theory Methods*, 15 :3277–3285.

Nelsen, R. B. (1999). *An Introduction to Copulas*. Springer, New York.

Oakes, D. (1982). A concordance test for independence in the presence of censoring. *Biometrics*, 38 :451–455.

Oakes, D. (1986). Semiparametric inference in a model for association in bivariate survival data. *Biometrika*, 73(2) :353–361.

Oakes, D. (1989). Bivariate survival models induced by frailties. *J. Amer. Statist. Assoc.*, 84 :487–493.

Oakes, D. (2005). On the preservation of copula structure under truncation. *Canad. J. Statist.*, 33 :465–468.

Oakes, D. (2006). On consistency of Kendall's tau under censoring. Technical report, Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY.

Panchenko, V. (2005). Goodness-of-fit test for copulas. *Phys. A*, 355 :176–182.

Plackett, R. L. (1965). A class of bivariate distributions. *J. Amer. Statist. Assoc.*, 60 :516–522.

Prentice, R. L. & Cai, J. (1992). Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika*, 79(3) :495–512.

Rivest, L.-P. & Wells, M. T. (2001). A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. *J. Multivariate Anal.*, 79(1) :138–155.

Robins, J. & Rotnitzky, A. (1992). Recovery of information and adjustement for dependent censoring using surrogate markers. *AIDS Epidemiology - Methodological Issues*, pages 297–331.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Ann. Math. Statist.*, 23 :470–472.

Scaillet, O. (2007). Kernel based goodness-of-fit tests for copulas with fixed smoothing parameters. *J. Multivariate Anal.*, 98 :533–543.

Scott, D. W. (1992). *Multivariate density estimation.* Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. Wiley, New York.

Shih, J. H. (1998). A goodness-of-fit test for association in a bivariate survival model. *Biometrika*, 85 :189–200.

Shih, J. H. & Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 51 :1384–1399.

Shorack, G. R. & Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics.* Wiley, New York.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8 :229–231.

Tsai, W.-Y. (1990). Testing the assumption of independence of truncation time and failure time. *Biometrika*, 77(1) :169–177.

Tsui, K.-L., Jewell, N. P., & Wu, C.-F. J. (1988). A nonparametric approach to the truncated regression problem. *J. Amer. Statist. Assoc.*, 83(403) :785–792.

Tsukahara, H. (2005). Semiparametric estimation in copula models. *Canad. J. Statist.*, 33 :357–375.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser. B*, 38(3) :290–295.

van der Laan, M. J. & Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality.* Springer Series in Statistics. Springer-Verlag, New York.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Van Keilegom, I. & Akritas, M. G. (1999). Transfer of tail information in censored regression models. *Ann. Statist.*, 27 :1745–1784.

Van Keilegom, I., Akritas, M. G., & Veraverbeke, N. (2001). Estimation of the conditional distribution in regression with censored data : a comparative study. *Comput. Statist. Data Anal.*, 35 :487–500.

Vandenhende, F. & Lambert, P. (2005). Local dependence estimation using semiparametric archimedean copulas. *Canad. J. Statist.*, 33 :377–388.

Wang, W. & Wells, M. T. (1997). Nonparametric estimators of the bivariate survival function under simplified censoring conditions. *Biometrika*, 84(4) :863–880.

Wang, W. & Wells, M. T. (1998). Nonparametric estimation of successive duration times under dependent censoring. *Biometrika*, 85(3) :561–572.

Wang, W. & Wells, M. T. (2000a). Estimation of Kendall's tau under censoring. *Statist. Sinica*, 10 :1199–1215.

Wang, W. & Wells, M. T. (2000b). Model selection and semiparametric inference for bivariate failure-time data. *J. Amer. Statist. Assoc.*, 95 :62–72.

Weier, D. R. & Basu, A. P. (1980). An investigation of Kendall's τ modified for censored data with applications. *J. Statist. Plann. Inference*, 4 :381–390.

Zhao, H. & Tsiatis, A. A. (1997). A consistent estimator for the distribution of quality adjusted survival time. *Biometrika*, 84(2) :339–348.

Zhao, H. & Tsiatis, A. A. (2000). Estimating mean quality adjusted lifetime with censored data. *Sankhyā Ser. B*, 62(1) :175–188. Biostatistics.

Zheng, M. & Klein, J. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82 :127–138.

Annexe A

Caractéristiques des dix copules considérées dans cette thèse

Les dix familles de copules utilisées dans le cadre de cette thèse ont été énumérées à la section 2.1.4. Nous exposons ici cinq caractéristiques de chacune d'entre elles, à savoir :

- appartenance ou non à la classe archimédienne ;
- générateur de la copule (si archimédienne) ;
- tau de Kendall ;
- distribution de Kendall ;
- algorithme de génération de données.

Les tableaux ci-dessous rappellent les quatre premières caractéristiques de ces familles. Des algorithmes de simulation de données sont ensuite fournis pour chacune d'entre elles.

TAB. A.1 – Quelques caractéristiques des 10 copules utilisées dans cette thèse.

Copule	Archimédienne ?	Générateur de la copule
Clayton	Oui	$\phi_\alpha(t) = \frac{1}{\alpha}(t^{-\alpha} - 1)$
Frank	Oui	$\phi_\alpha(t) = -\ln\left(\frac{e^{-\alpha t}-1}{e^{-\alpha}-1}\right)$
G.-Hougaard	Oui	$\phi_\alpha(t) = \{-\ln(t)\}^\alpha$
Plackett	Non	-
Normale	Non	-
Student	Non	-
Pearson type II	Non	-
Copule 3	Oui	$(1 - t^{1/\alpha})^\alpha$
Copule 4	Oui	$-\ln\{(1-\alpha)t + \alpha\}$
Copule 5	Oui	$\frac{1-t}{1+(\alpha-1)t}$

TAB. A.2 – Autres caractéristiques des 10 copules utilisées dans cette thèse.

Copule	Tau de Kendall	Processus de Kendall, $K_\alpha(t)$
Clayton	$\frac{\alpha}{\alpha+2}$	$t + \frac{t}{\alpha}(1 - t^\alpha)$
Frank	$1 - \frac{4}{\alpha}\left(1 - \frac{1}{\alpha}\int_0^\alpha \frac{tdt}{e^t-1}\right)$	$t + \frac{1}{\alpha}(1 - e^{\alpha t})\ln\left(\frac{e^{-\alpha t}-1}{e^{-\alpha}-1}\right)$
G.-Hougaard	$\frac{\alpha-1}{\alpha}$	$t\left\{1 - \frac{\ln(t)}{\alpha}\right\}$
Plackett	$4\int_0^1\int_0^1 C_\alpha(u,v)\frac{\partial^2}{\partial u\partial v}C_\alpha(u,v)dudv - 1$	(*)
Normale	$\frac{2}{\pi}\text{arcsin}(\rho)$	(*)
Student	$\frac{2}{\pi}\text{arcsin}(\rho)$	(*)
Pearson type II	$\frac{2}{\pi}\text{arcsin}(\rho)$	(*)
Copule 3	$\frac{2\alpha-3}{2\alpha-1}$	$t^{(\alpha-1)/\alpha}$
Copule 4	$-\frac{2\alpha}{(1-\alpha)^2}(1 - \alpha + \alpha\ln\alpha)$	$t - \alpha^{-1}(\alpha t + 1 - \alpha)\ln(\alpha t + 1 - \alpha)$
Copule 5	$\frac{\alpha-4}{3\alpha}$	$t + \frac{(1-t)\{1+(\alpha-1)t\}}{\alpha}$

(*) : le processus de Kendall est trop complexe à évaluer. Pour y remédier, un échantillon bootstrap de paramètre α_n (obtenu à l'aide de l'inversion du tau de Kendall empirique) est généré et on calcule le processus de Kendall empirique sur cet échantillon.

1. Algorithme de simulation de la famille de Clayton

   a) Générer $u \sim \mathcal{U}(0,1)$ et $t \sim \mathcal{U}(0,1)$ de façon indépendante.
   b) Poser
   $$v = (1 - u^{-\alpha} + t^{-\alpha/(\alpha+1)} u^{-\alpha})^{-1/\alpha}.$$
   c) Le couple $(u, v) \sim \text{Clayton}(\alpha)$.

2. Algorithme de simulation de la famille de Frank

   a) Générer $u \sim \mathcal{U}(0,1)$ et $t \sim \mathcal{U}(0,1)$ de façon indépendante.
   b) Poser
   $$v = -\frac{1}{\alpha}\ln\left\{\frac{e^{-\alpha u}(1-t) + te^{-\alpha}}{e^{-\alpha u}(1-t) + t}\right\}.$$
   c) Le couple $(u, v) \sim \text{Frank}(\alpha)$.

3. Algorithme de simulation de la famille de Gumbel–Hougaard

   a) Générer $u \sim \mathcal{U}(0,1)$ et $t \sim \mathcal{U}(0,1)$ de façon indépendante.
   b) Poser $a = \{-\ln(u)\}^{\alpha}$ et $b = \{-\ln(v)\}^{\alpha}$.
   c) À l'aide de la méthode de la bissection, trouver la valeur de v permettant d'obtenir l'égalité suivante :
   $$t = \frac{-\exp\left\{-(a+b)^{1/\alpha}\right\} a(a+b)^{(1-\alpha)/\alpha}}{u\ln(u)}.$$
   c) Le couple $(u, v) \sim \text{Gumbel-Hougaard}(\alpha)$.

4. Algorithme de simulation de la famille de Plackett

   a) Générer $u \sim \mathcal{U}(0,1)$ et $t \sim \mathcal{U}(0,1)$ de façon indépendante.
   b) Poser
   $$\begin{aligned} a &= 2t(1-t)(u\alpha^2 + 1 - u) + \alpha\{1 - 2t(1-t)\} \\ b &= (1-2t)\sqrt{\alpha}\sqrt{\alpha + 4t(1-t)u(1-u)(1-\alpha)^2} \\ c &= 2\left\{\alpha + t(1-t)(\alpha-1)^2\right\} \end{aligned}$$
   c) Calculer $v = (a - b)/c$.
   d) Le couple $(u, v) \sim \text{Plackett}(\alpha)$.

5. Algorithme de simulation de la famille gaussienne

   a) Générer $x \sim \mathcal{N}(0,1)$ et $y \sim \mathcal{N}(0,1)$ de façon indépendante.

   b) Poser
   $$u = \Phi(x), \quad v = \Phi(x\rho + y\sqrt{1-\rho^2}),$$
   où Φ est la fonction de répartition d'une variable aléatoire $\mathcal{N}(0,1)$.

   c) Le couple $(u,v) \sim$ Normale(ρ).

6. Algorithme de simulation de la famille de Student

   a) Générer $x \sim \mathcal{N}(0,1)$, $y \sim \mathcal{N}(0,1)$ et $z \sim \chi^2_\nu$ de façon indépendante.

   b) Poser
   $$a = \frac{x}{\sqrt{z/\nu}}, \quad b = \frac{x\rho + y\sqrt{1-\rho^2}}{\sqrt{z/\nu}}.$$

   c) Calculer $u = t_\nu(a)$ et $v = t_\nu(b)$, où t_ν est la fonction de répartition d'une variable aléatoire de Student à ν degrés de liberté.

   d) Le couple $(u,v) \sim$ Student(ρ,ν).

7. Algorithme de simulation de la famille de Pearson type II

   a) Générer $u \sim \mathcal{U}(0,1)$ et $t \sim \mathcal{U}(0,1)$ de façon indépendante.

   b) Poser
   $$x = \cos(2\pi u), \quad y = \sin(2\pi u).$$

   c) Poser $R = \sqrt{1-(1-t)^{1/(\nu+1)}}$.

   d) Calculer
   $$u = F_{PII,\nu}(Rx), \quad v = F_{PII,\nu}\{R(x\rho + y\sqrt{1-\rho^2})\},$$
   où $F_{PII,\nu}$ est la fonction de répartition d'une variable aléatoire suivant la loi marginale de Pearson de type II à ν degrés de liberté.

   e) Le couple $(u,v) \sim$ Pearson type II(ρ,ν).

8. Algorithme de simulation de la famille de copules # 3 du quatrième article

   a) Générer $u \sim \mathcal{U}(0,1)$ et $t \sim \mathcal{U}(0,1)$ de façon indépendante.

   b) Poser
   $$C_{10}(u,v,\alpha) = u^{(1/\alpha)-1}(1-u^{1/\alpha})^{\alpha-1}A^{(1/\alpha)-1}(1-A^{1/\alpha})^{\alpha-1},$$
   où
   $$A = (1-u^{1/\alpha})^\alpha + (1-v^{1/\alpha})^\alpha.$$

c) Si $C_{10}(u,0,\alpha) > t$, alors

$$v = \phi^{-1}\{\phi_\alpha(0) - \phi_\alpha(u)\},$$

où ϕ_α est le générateur de la copule 3 (voir Tableau A.1).

d) Si $C_{10}(u,0,\alpha) \leq t$, alors trouver la valeur de v telle que $C_{10}(u,v,\alpha) = t$.

e) Le couple $(u,v) \sim$ Copule # 3(α).

9. Algorithme de simulation de la famille de copules # 4 du quatrième article

a) Générer $u \sim \mathcal{U}(0,1)$ et $t \sim \mathcal{U}(0,1)$ de façon indépendante.

b) Poser

$$v = \max\left[\frac{t-\alpha}{1-\alpha}, \phi_\alpha^{-1}\{\phi_\alpha(0) - \phi_\alpha(u)\}\right],$$

où ϕ_α est le générateur de la copule 4 (voir Tableau A.1).

c) Le couple $(u,v) \sim$ Copule # 4(α).

10. Algorithme de simulation de la famille de copules # 5 du quatrième article

a) Générer $u \sim \mathcal{U}(0,1)$ et $t \sim \mathcal{U}(0,1)$ de façon indépendante.

b) Poser $C_{10}(u,v,\alpha) = (AB - CD)/A^2$, où

$$\begin{aligned}
A &= \alpha^2 - (\alpha-1)^2(1-v) + u(\alpha-1)^2(1-v),\\
B &= \alpha^2 v + 1 - v,\\
C &= \alpha^2 uv - (1-v) + u(1-v),\\
D &= (\alpha-1)^2(1-v).
\end{aligned}$$

c) Si $C_{10}(u,0,\alpha) > t$, alors $v = \phi^{-1}\{\phi_\alpha(0) - \phi_\alpha(u)\}$, où ϕ_α est le générateur de la copule 5 (voir Tableau A.1).

d) Si $C_{10}(u,0,\alpha) \leq t$, alors trouver la valeur de v telle que $C_{10}(u,v,\alpha) = t$.

e) Le couple $(u,v) \sim$ Copule # 5(α).

Annexe B

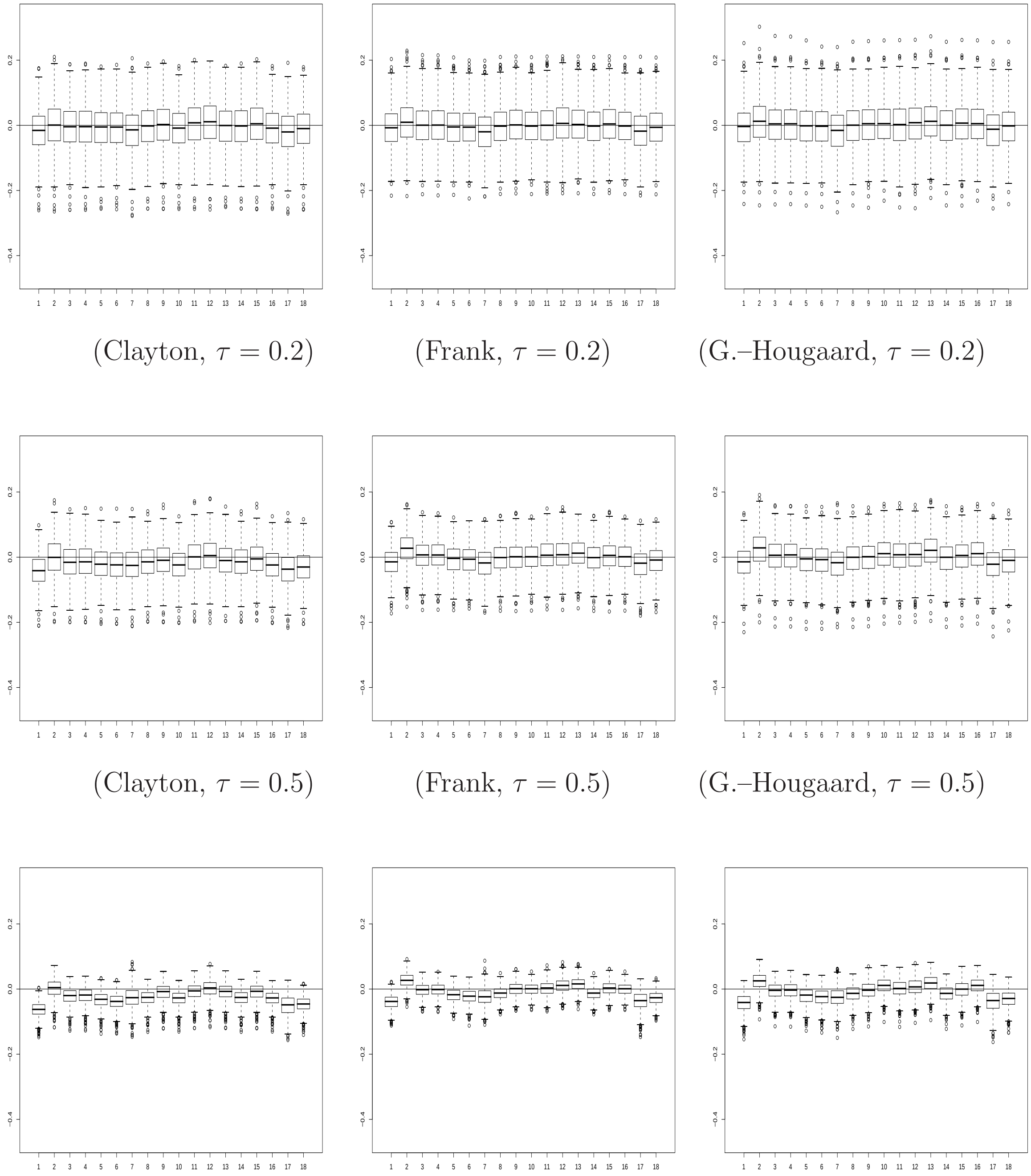
Diagrammes en boîte du chapitre 4

Le deuxième article de la thèse (chapitre 4) présente 12 diagrammes en boîte. Chacun d'eux compare la performance des 18 estimateurs du tau de Kendall considérés dans le projet. Ces diagrammes correspondent aux conditions de simulation suivantes :

- 3 choix de copules (Clayton, Frank, Gumbel–Hougaard) ;
- 2 fractions de censure (20%, 40%) ;
- 2 valeurs du tau de Kendall ($\tau = 0.5, 0.8$) ;
- une taille d'échantillon ($n = 100$).

Or, l'étude de Monte-Carlo originale considérait également les cas $\tau = 0.2$ et $n = 200$. Ceci mène plutôt à 36 diagrammes en boîte, qui sont consignés dans cette annexe. Il est important de noter que la même échelle a été utilisée pour les 18 diagrammes en boîte correspondant au cas $n = 100$. La même restriction a été appliquée aux 18 graphiques associés au cas $n = 200$.

FIG. B.1 – Comparison of 18 estimators of τ subject to $\mathbf{C} = \mathbf{20}\%$ censoring in 1000 samples of size $\mathbf{n} = \mathbf{100}$ from distributions with different dependence structures and degrees of dependence. Both X and Y are assumed to be log-normal with mean 30 and variance 50.

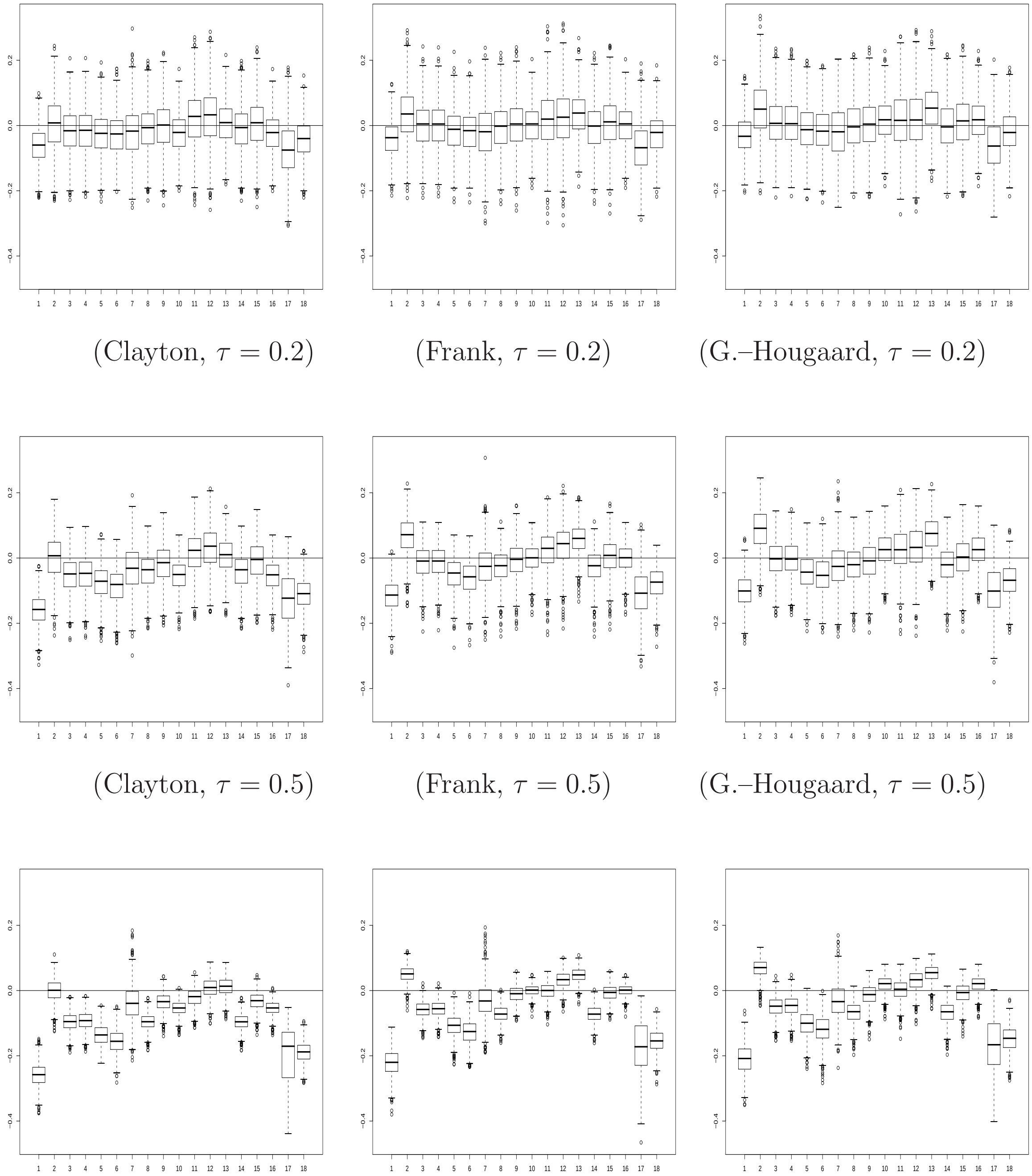


(Clayton, $\tau = 0.2$) (Frank, $\tau = 0.2$) (G.–Hougaard, $\tau = 0.2$)

(Clayton, $\tau = 0.5$) (Frank, $\tau = 0.5$) (G.–Hougaard, $\tau = 0.5$)

(Clayton, $\tau = 0.8$) (Frank, $\tau = 0.8$) (G.–Hougaard, $\tau = 0.8$)

Fig. B.2 – Comparison of 18 estimators of τ subject to $\mathbf{C = 40}$% censoring in 1000 samples of size $\mathbf{n = 100}$ from distributions with different dependence structures and degrees of dependence. Both X and Y are assumed to be log-normal with mean 30 and variance 50.

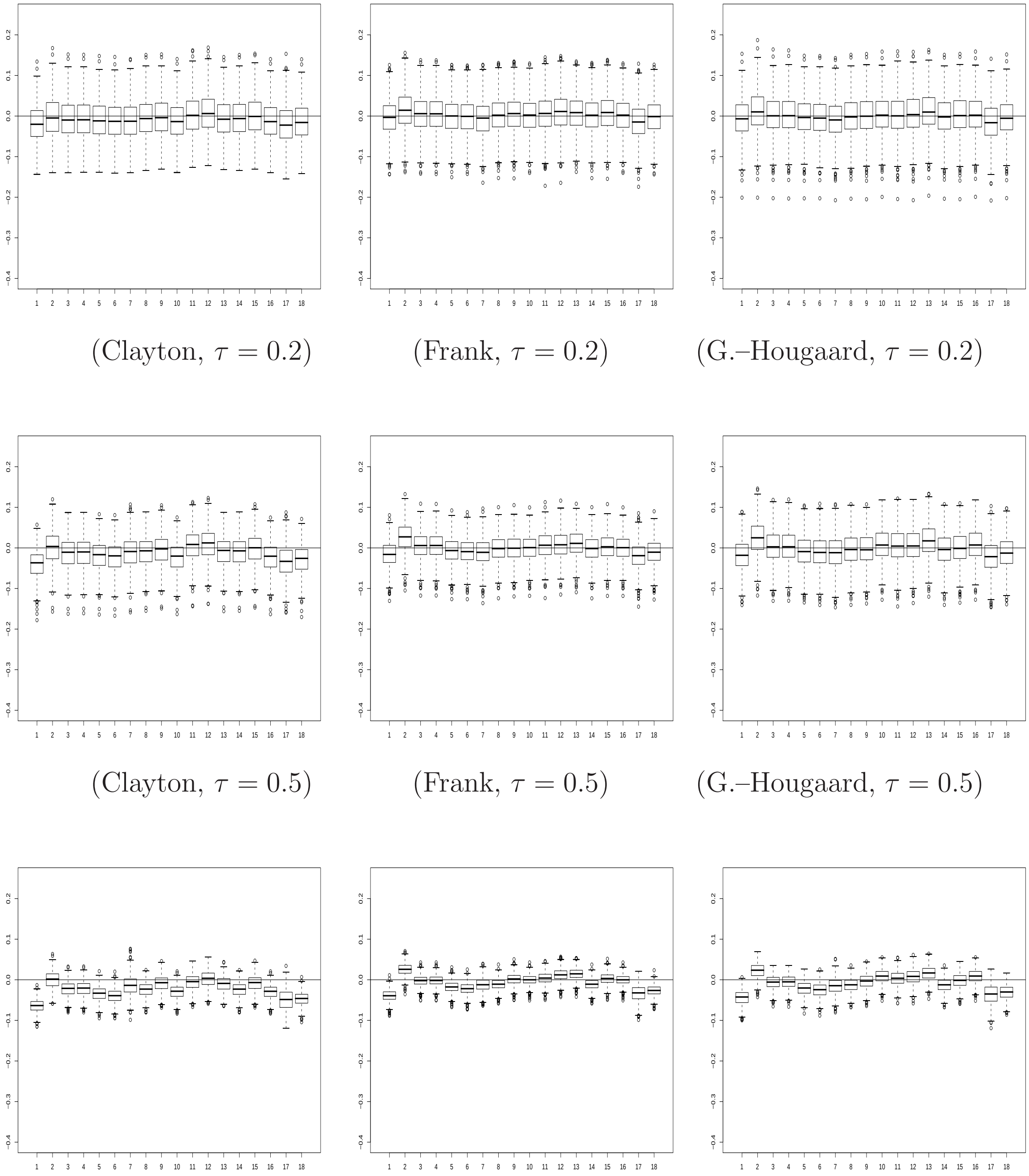


(Clayton, $\tau = 0.2$) (Frank, $\tau = 0.2$) (G.–Hougaard, $\tau = 0.2$)

(Clayton, $\tau = 0.5$) (Frank, $\tau = 0.5$) (G.–Hougaard, $\tau = 0.5$)

(Clayton, $\tau = 0.8$) (Frank, $\tau = 0.8$) (G.–Hougaard, $\tau = 0.8$)

Fig. B.3 – Comparison of 18 estimators of τ subject to $\mathbf{C} = \mathbf{20}\%$ censoring in 1000 samples of size $\mathbf{n} = \mathbf{200}$ from distributions with different dependence structures and degrees of dependence. Both X and Y are assumed to be log-normal with mean 30 and variance 50.

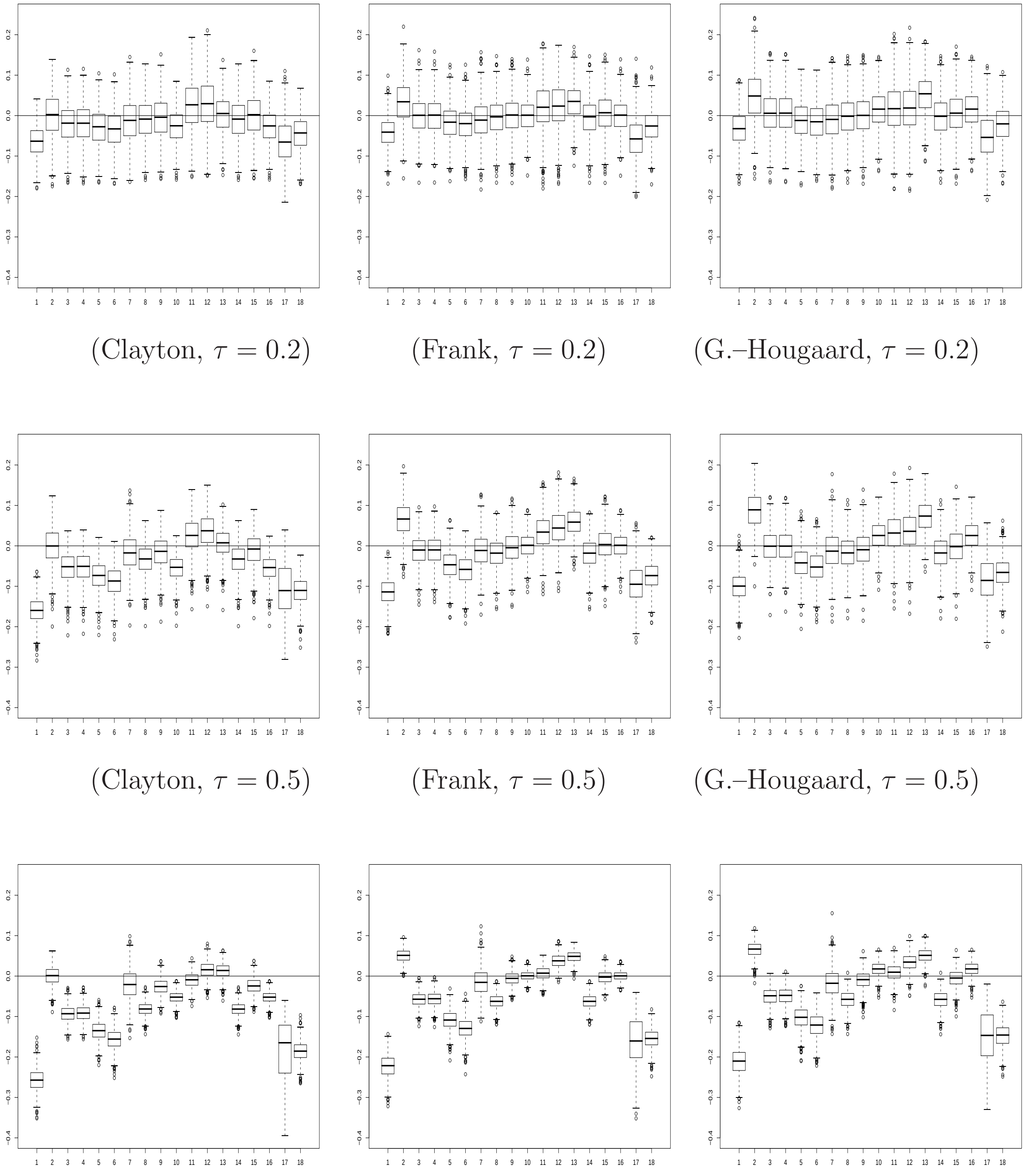


(Clayton, $\tau = 0.2$) (Frank, $\tau = 0.2$) (G.–Hougaard, $\tau = 0.2$)

(Clayton, $\tau = 0.5$) (Frank, $\tau = 0.5$) (G.–Hougaard, $\tau = 0.5$)

(Clayton, $\tau = 0.8$) (Frank, $\tau = 0.8$) (G.–Hougaard, $\tau = 0.8$)

Fig. B.4 – Comparison of 18 estimators of τ subject to $\mathbf{C} = \mathbf{40}\%$ censoring in 1000 samples of size $\mathbf{n} = \mathbf{200}$ from distributions with different dependence structures and degrees of dependence. Both X and Y are assumed to be log-normal with mean 30 and variance 50.



(Clayton, $\tau = 0.2$) (Frank, $\tau = 0.2$) (G.–Hougaard, $\tau = 0.2$)

(Clayton, $\tau = 0.5$) (Frank, $\tau = 0.5$) (G.–Hougaard, $\tau = 0.5$)

(Clayton, $\tau = 0.8$) (Frank, $\tau = 0.8$) (G.–Hougaard, $\tau = 0.8$)

Annexe C

Histogrammes de normalité du chapitre 4

Dans le cadre du second article de la thèse (chapitre 4), la normalité de 1000 estimations du tau de Kendall déduites de la méthode WePa-VKA (9) a été étudiée sous divers scénarios. Par manque d'espace, la présentation des résultats a dû être limitée à la copule de Frank et à une taille d'échantillon, n, fixée à 100.

Tel que mentionné dans l'article, une analyse semblable a toutefois été faite pour les copules de Clayton et de Gumbel–Hougaard avec $n = 200$ pour les deux autres estimateurs les plus performants, à savoir CoWeBa-LPT (11) et Icdf-VKA (15). Les histogrammes liés à cette analyse sont présentés ici.

Un tableau résumant les seuils observés du test de Shapiro–Wilk sous toutes ces conditions est également fourni à la fin.

Fig. C.1 – Histogrammes montrant la répartition des estimations de τ selon la **méthode WePa–VKA (9)**. Chaque graphique est basé sur 1000 échantillons de taille $\mathbf{n = 100}$ issus de la copule de **Clayton**. Les colonnes de gauche et de droite correspondent respectivement à $C = 20\%$ et $C = 40\%$. Les valeurs théoriques de τ sont respectivement de 0.2, 0.5 et 0.8 dans les lignes du haut, du milieu et du bas.

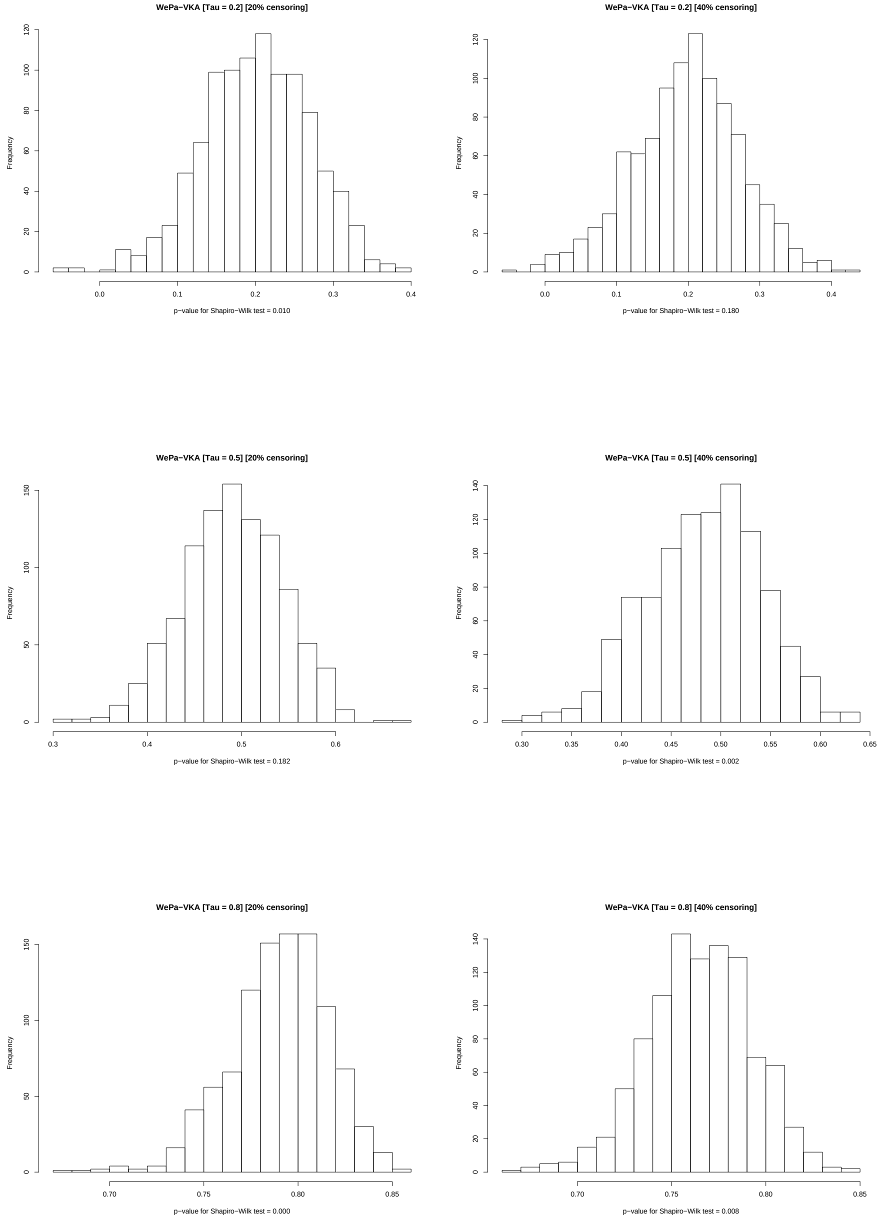


FIG. C.2 – Histogrammes montrant la répartition des estimations de τ selon la **méthode CoWeBa–LPT (11)**. Chaque graphique est basé sur 1000 échantillons de taille $\mathbf{n = 100}$ issus de la copule de **Clayton**. Les colonnes de gauche et de droite correspondent respectivement à $C = 20\%$ et $C = 40\%$. Les valeurs théoriques de τ sont respectivement de 0.2, 0.5 et 0.8 dans les lignes du haut, du milieu et du bas.

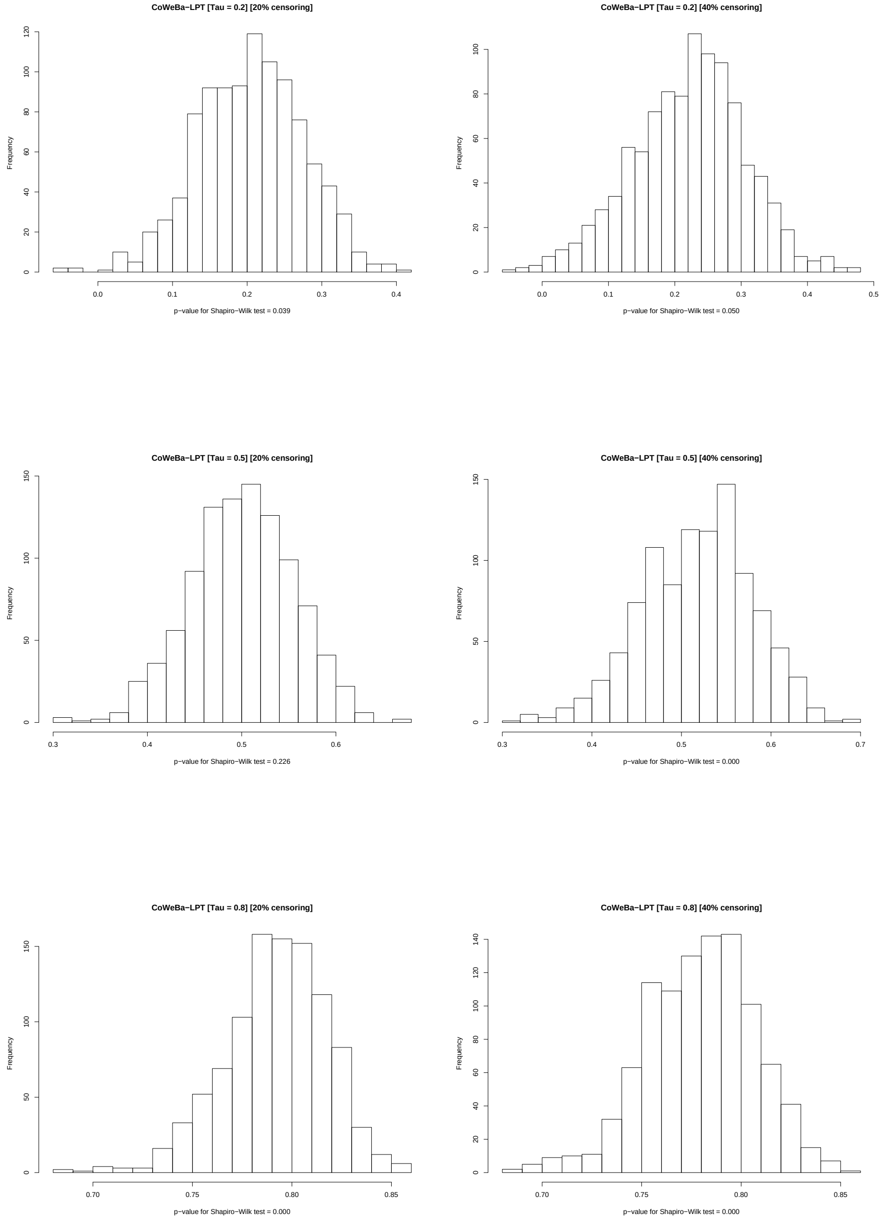


Fig. C.3 – Histogrammes montrant la répartition des estimations de τ selon la **méthode Icdf–VKA (15)**. Chaque graphique est basé sur 1000 échantillons de taille $\mathbf{n = 100}$ issus de la copule de **Clayton**. Les colonnes de gauche et de droite correspondent respectivement à $C = 20\%$ et $C = 40\%$. Les valeurs théoriques de τ sont respectivement de 0.2, 0.5 et 0.8 dans les lignes du haut, du milieu et du bas.

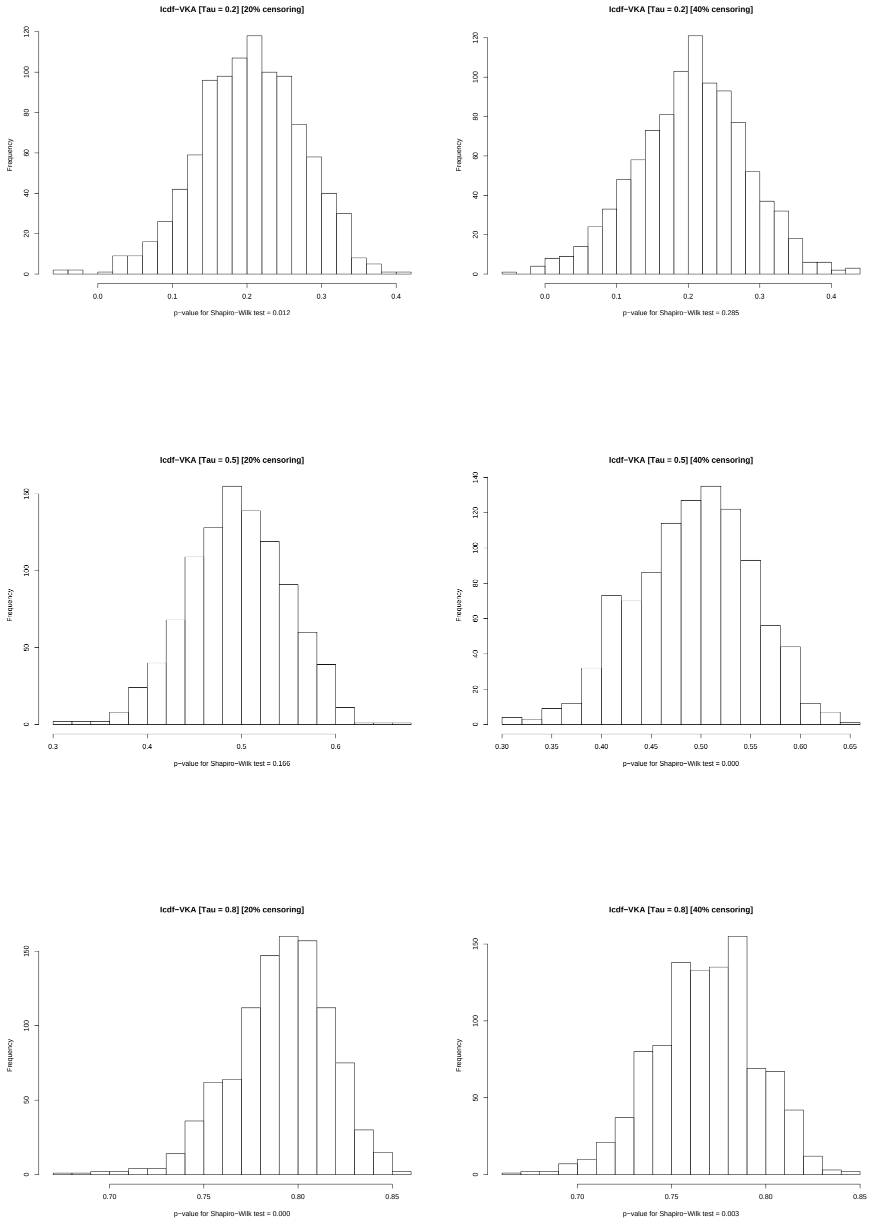


Fig. C.4 – Histogrammes montrant la répartition des estimations de τ selon la **méthode WePa–VKA (9)**. Chaque graphique est basé sur 1000 échantillons de taille $\mathbf{n = 100}$ issus de la copule de **Frank**. Les colonnes de gauche et de droite correspondent respectivement à $C = 20\%$ et $C = 40\%$. Les valeurs théoriques de τ sont respectivement de 0.2, 0.5 et 0.8 dans les lignes du haut, du milieu et du bas.

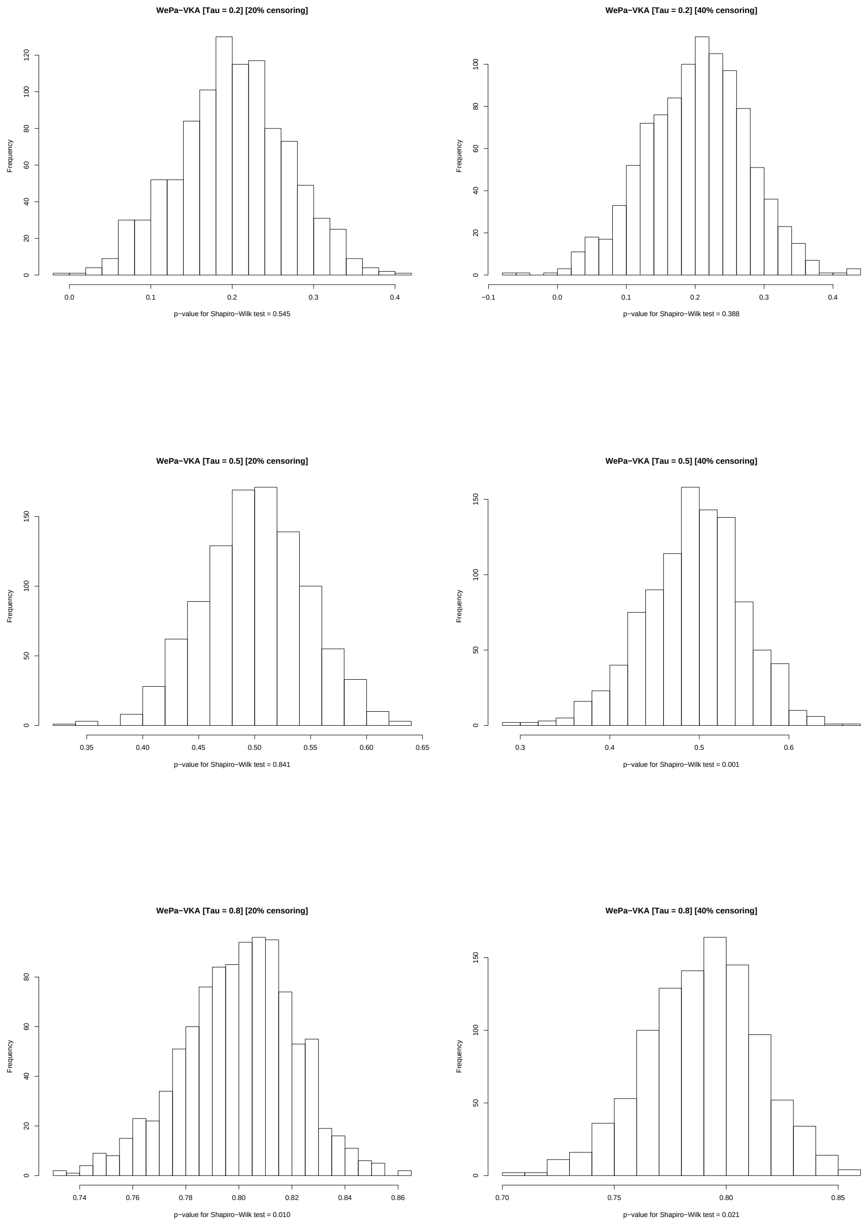


Fig. C.5 – Histogrammes montrant la répartition des estimations de τ selon la **méthode CoWeBa–LPT (11)**. Chaque graphique est basé sur 1000 échantillons de taille $\mathbf{n = 100}$ issus de la copule de **Frank**. Les colonnes de gauche et de droite correspondent respectivement à $C = 20\%$ et $C = 40\%$. Les valeurs théoriques de τ sont respectivement de 0.2, 0.5 et 0.8 dans les lignes du haut, du milieu et du bas.

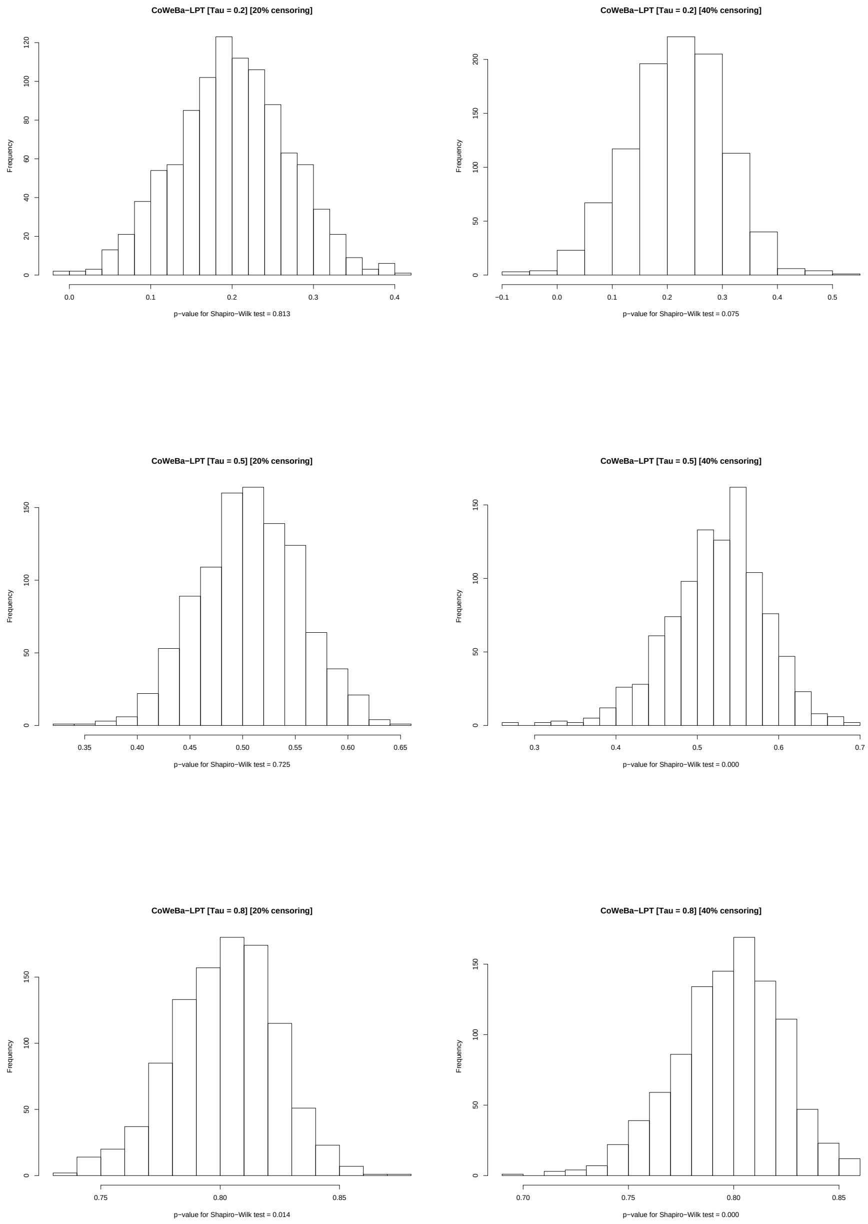


Fig. C.6 – Histogrammes montrant la répartition des estimations de τ selon la **méthode Icdf–VKA (15)**. Chaque graphique est basé sur 1000 échantillons de taille $\mathbf{n = 100}$ issus de la copule de **Frank**. Les colonnes de gauche et de droite correspondent respectivement à $C = 20\%$ et $C = 40\%$. Les valeurs théoriques de τ sont respectivement de 0.2, 0.5 et 0.8 dans les lignes du haut, du milieu et du bas.

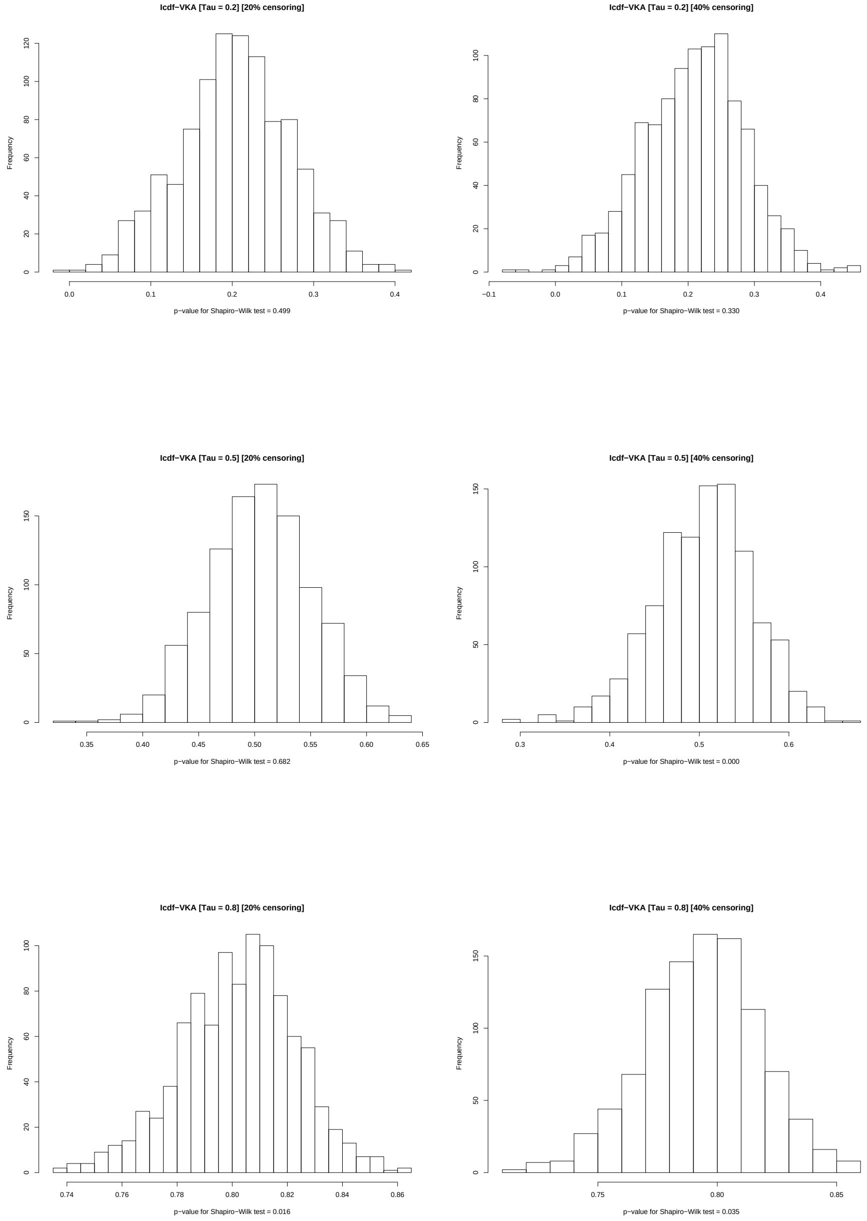


Fig. C.7 – Histogrammes montrant la répartition des estimations de τ selon la **méthode WePa–VKA (9)**. Chaque graphique est basé sur 1000 échantillons de taille $\mathbf{n = 100}$ issus de la copule de **Gumbel–Hougaard**. Les colonnes de gauche et de droite correspondent respectivement à $C = 20\%$ et $C = 40\%$. Les valeurs théoriques de τ sont respectivement de 0.2, 0.5 et 0.8 dans les lignes du haut, du milieu et du bas.

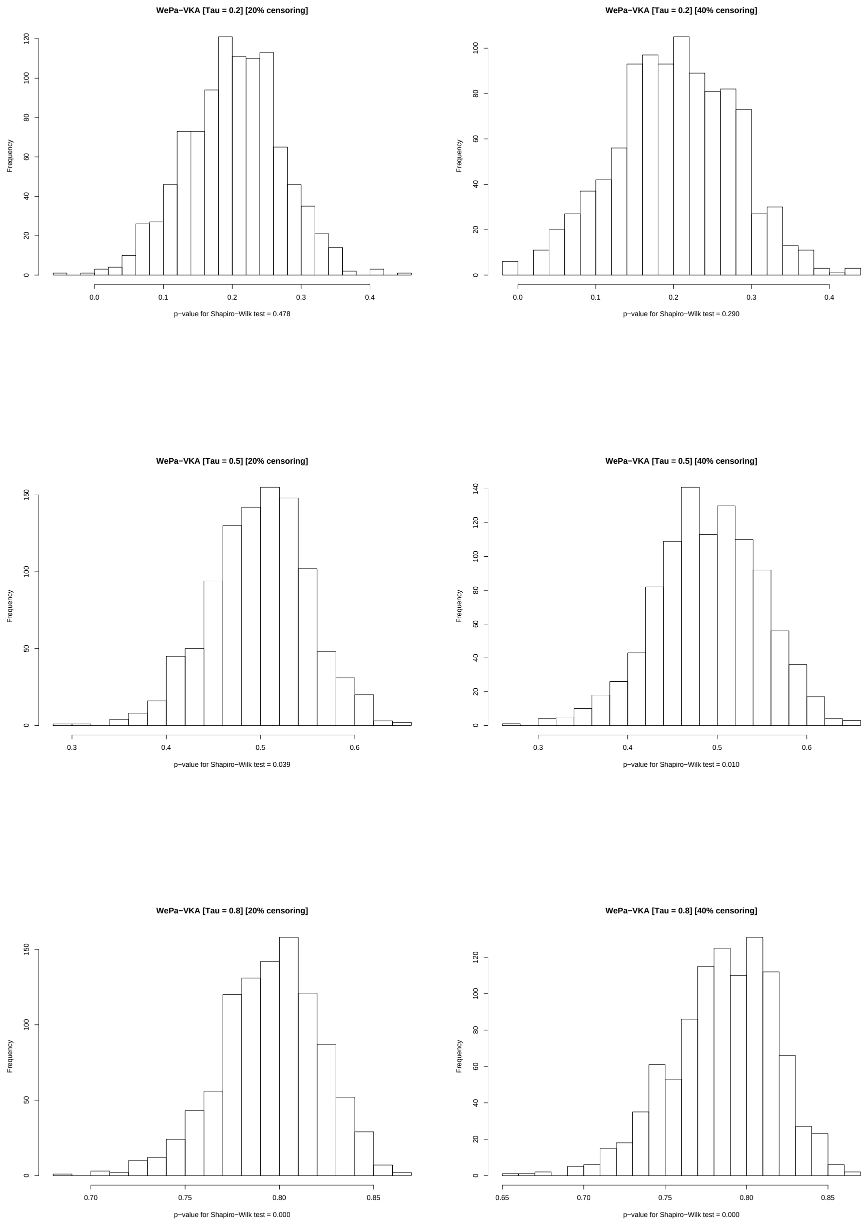


Fig. C.8 – Histogrammes montrant la répartition des estimations de τ selon la **méthode CoWeBa–LPT (11)**. Chaque graphique est basé sur 1000 échantillons de taille $\mathbf{n = 100}$ issus de la copule de **Gumbel–Hougaard**. Les colonnes de gauche et de droite correspondent respectivement à $C = 20\%$ et $C = 40\%$. Les valeurs théoriques de τ sont respectivement de 0.2, 0.5 et 0.8 dans les lignes du haut, du milieu et du bas.

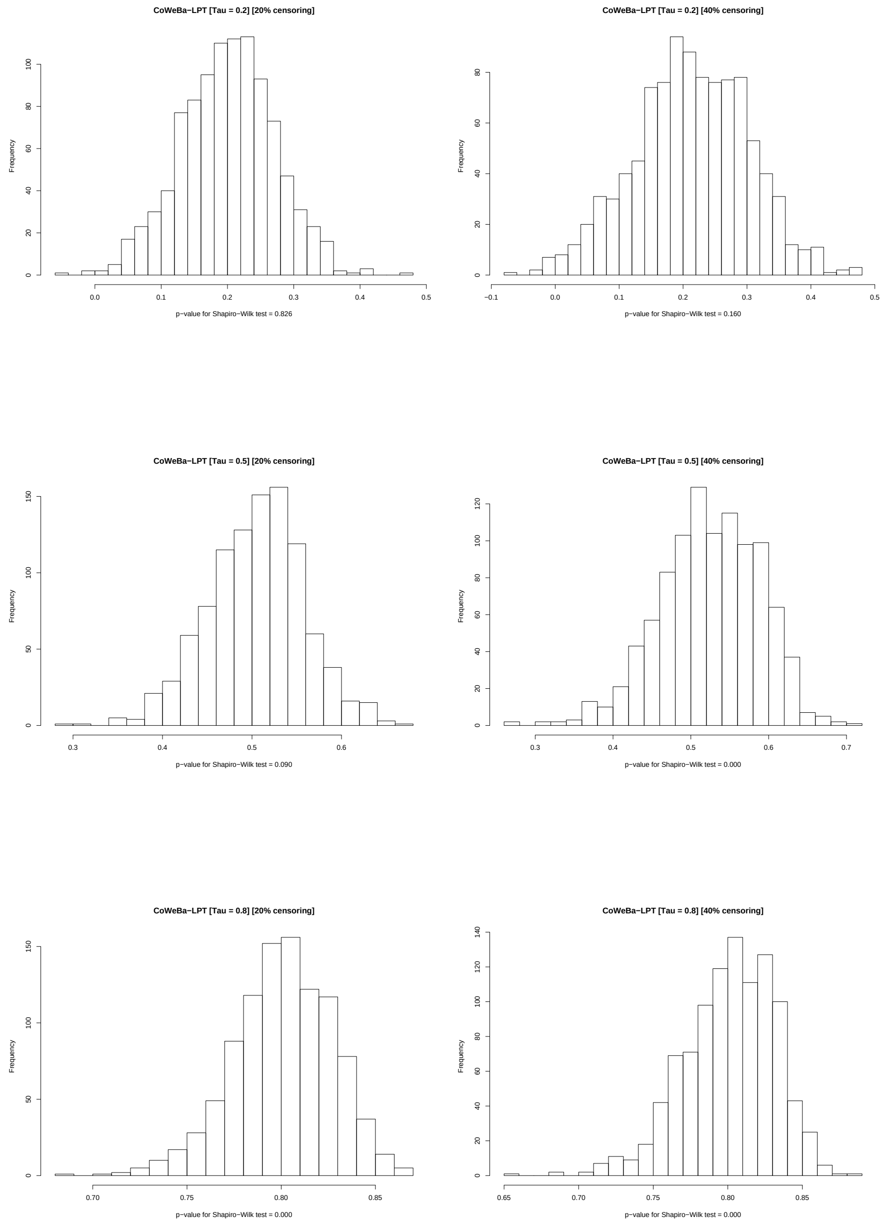


Fig. C.9 – Histogrammes montrant la répartition des estimations de τ selon la **méthode Icdf–VKA (15)**. Chaque graphique est basé sur 1000 échantillons de taille $\mathbf{n = 100}$ issus de la copule de **Gumbel–Hougaard**. Les colonnes de gauche et de droite correspondent respectivement à $C = 20\%$ et $C = 40\%$. Les valeurs théoriques de τ sont respectivement de 0.2, 0.5 et 0.8 dans les lignes du haut, du milieu et du bas.

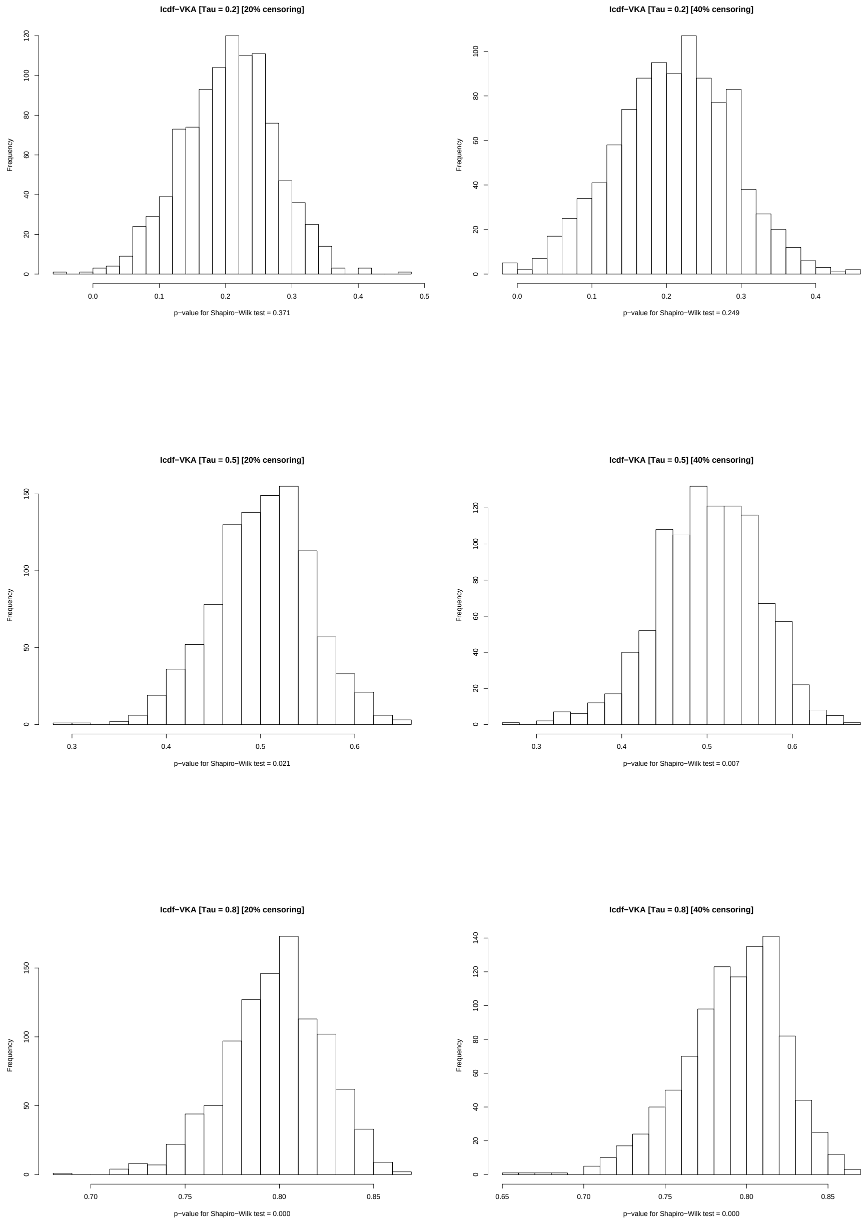


Fig. C.10 – Histogrammes montrant la répartition des estimations de τ selon la **méthode WePa–VKA (9)**. Chaque graphique est basé sur 1000 échantillons de taille $\mathbf{n = 200}$ issus de la copule de **Clayton**. Les colonnes de gauche et de droite correspondent respectivement à $C = 20\%$ et $C = 40\%$. Les valeurs théoriques de τ sont respectivement de 0.2, 0.5 et 0.8 dans les lignes du haut, du milieu et du bas.

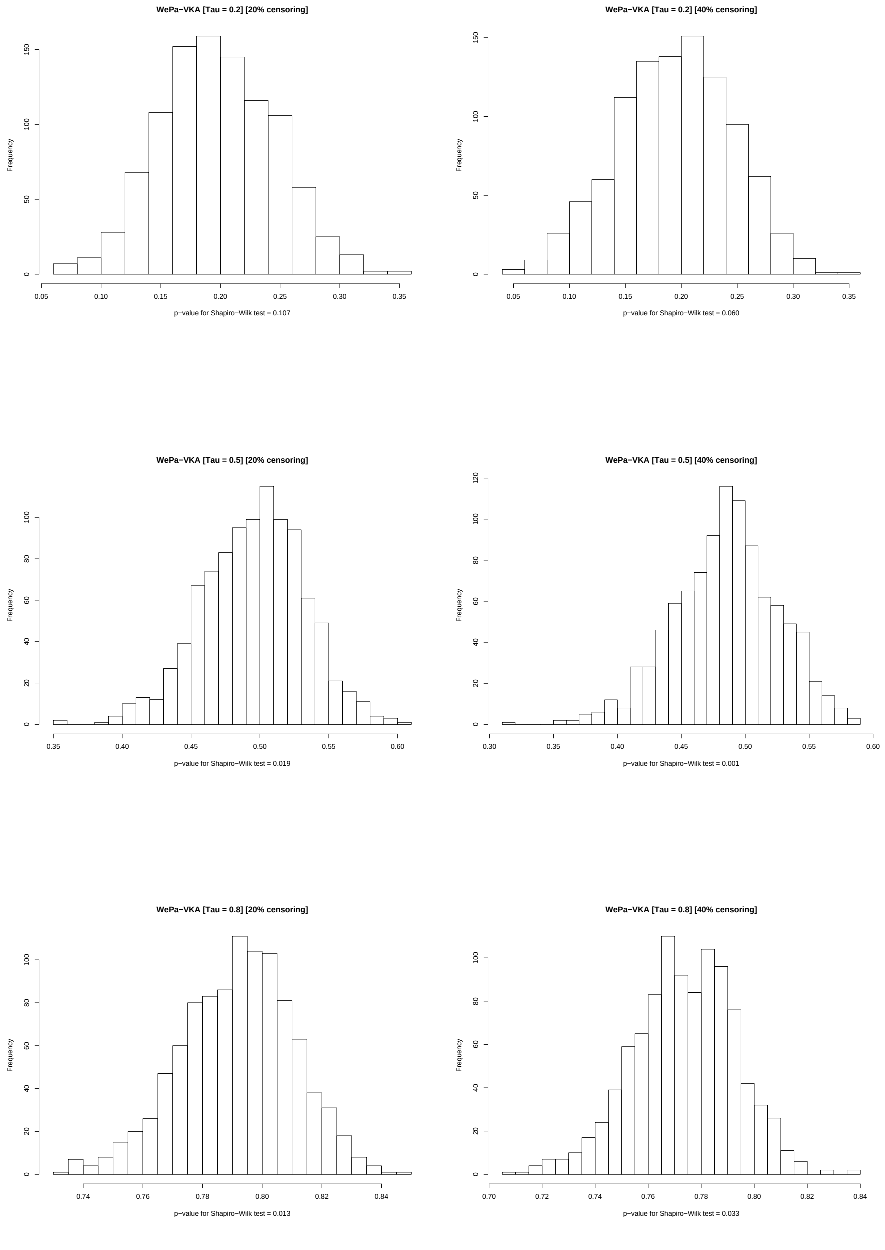


Fig. C.11 – Histogrammes montrant la répartition des estimations de τ selon la **méthode CoWeBa–LPT (11)**. Chaque graphique est basé sur 1000 échantillons de taille $\mathbf{n = 200}$ issus de la copule de **Clayton**. Les colonnes de gauche et de droite correspondent respectivement à $C = 20\%$ et $C = 40\%$. Les valeurs théoriques de τ sont respectivement de 0.2, 0.5 et 0.8 dans les lignes du haut, du milieu et du bas.

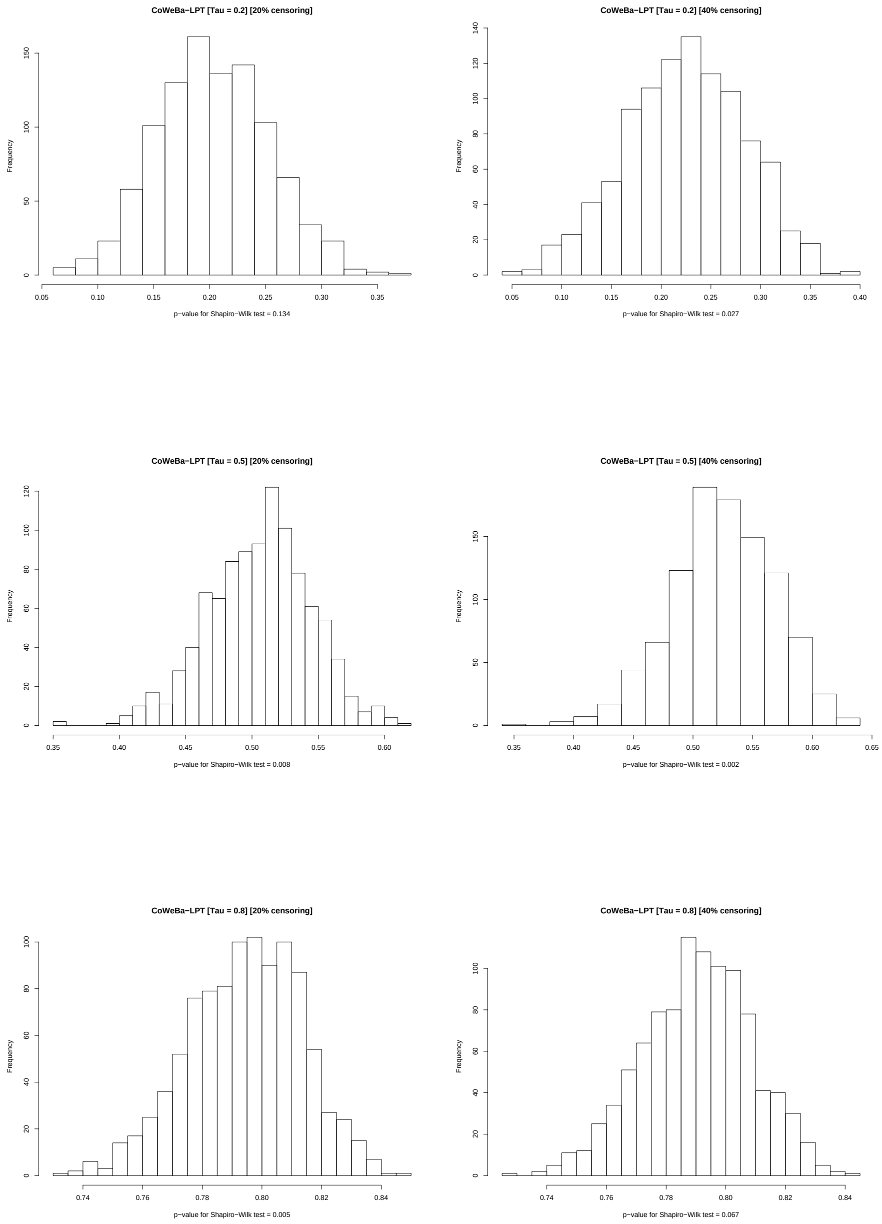


Fig. C.12 – Histogrammes montrant la répartition des estimations de τ selon la **méthode Icdf–VKA (15)**. Chaque graphique est basé sur 1000 échantillons de taille $\mathbf{n = 200}$ issus de la copule de **Clayton**. Les colonnes de gauche et de droite correspondent respectivement à $C = 20\%$ et $C = 40\%$. Les valeurs théoriques de τ sont respectivement de 0.2, 0.5 et 0.8 dans les lignes du haut, du milieu et du bas.

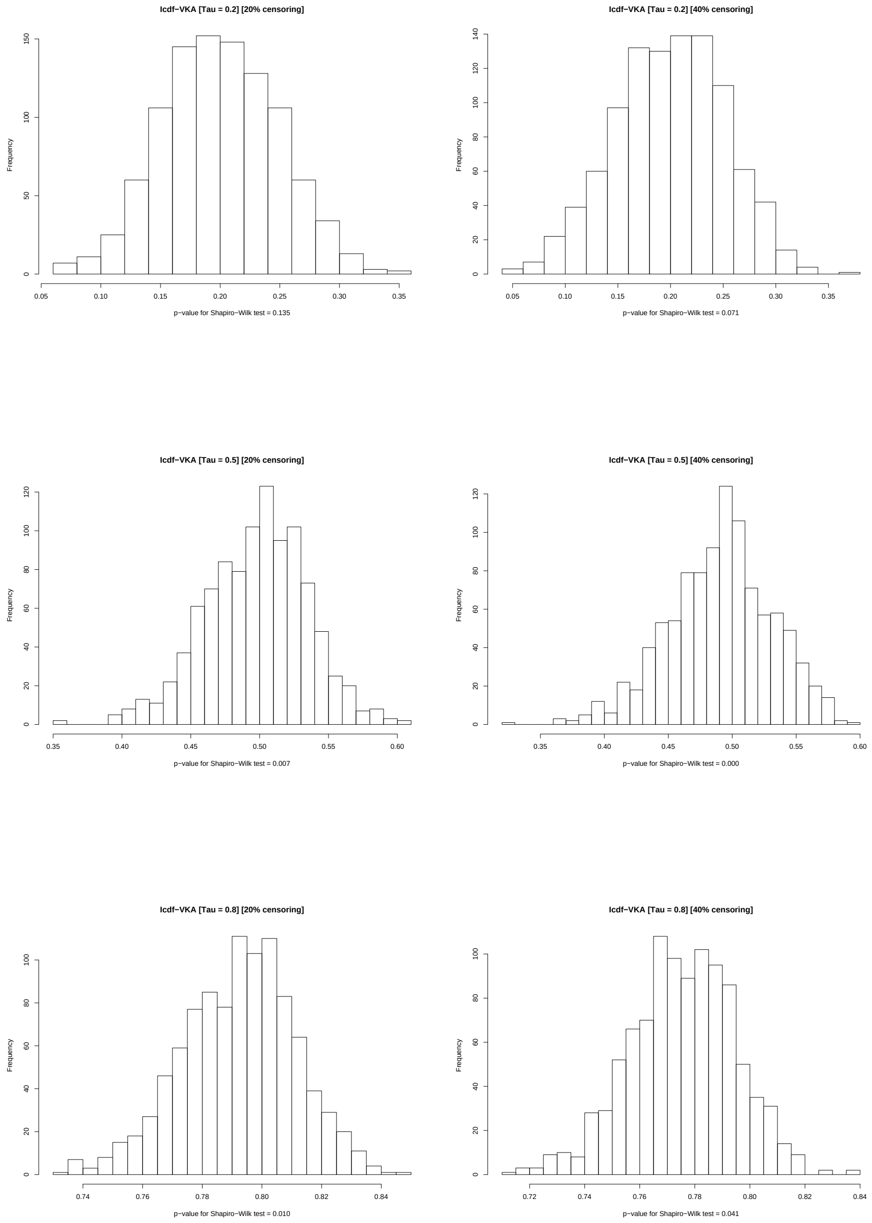


Fig. C.13 – Histogrammes montrant la répartition des estimations de τ selon la **méthode WePa–VKA (9)**. Chaque graphique est basé sur 1000 échantillons de taille $\mathbf{n} = \mathbf{200}$ issus de la copule de **Frank**. Les colonnes de gauche et de droite correspondent respectivement à $C = 20\%$ et $C = 40\%$. Les valeurs théoriques de τ sont respectivement de 0.2, 0.5 et 0.8 dans les lignes du haut, du milieu et du bas.

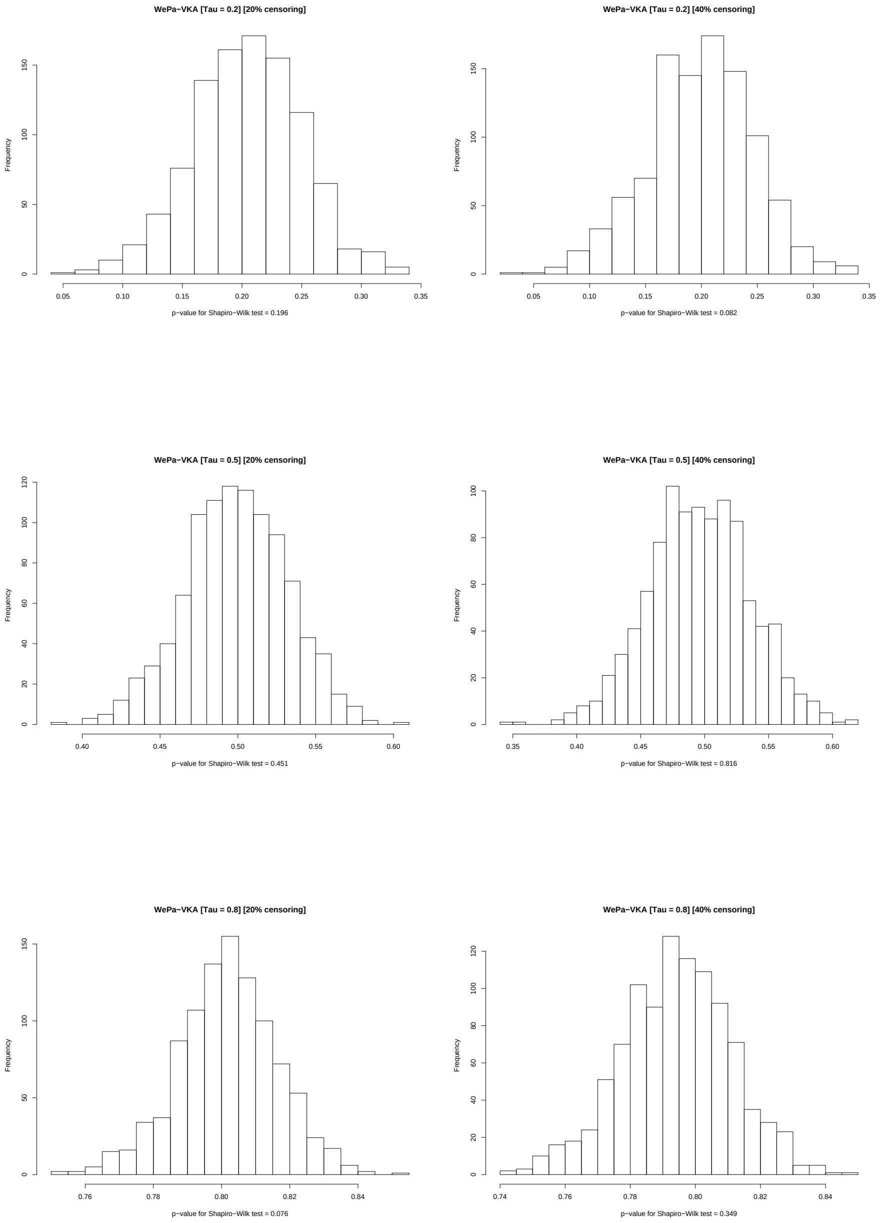


FIG. C.14 – Histogrammes montrant la répartition des estimations de τ selon la **méthode CoWeBa–LPT (11)**. Chaque graphique est basé sur 1000 échantillons de taille $\mathbf{n} = \mathbf{200}$ issus de la copule de **Frank**. Les colonnes de gauche et de droite correspondent respectivement à $C = 20\%$ et $C = 40\%$. Les valeurs théoriques de τ sont respectivement de 0.2, 0.5 et 0.8 dans les lignes du haut, du milieu et du bas.

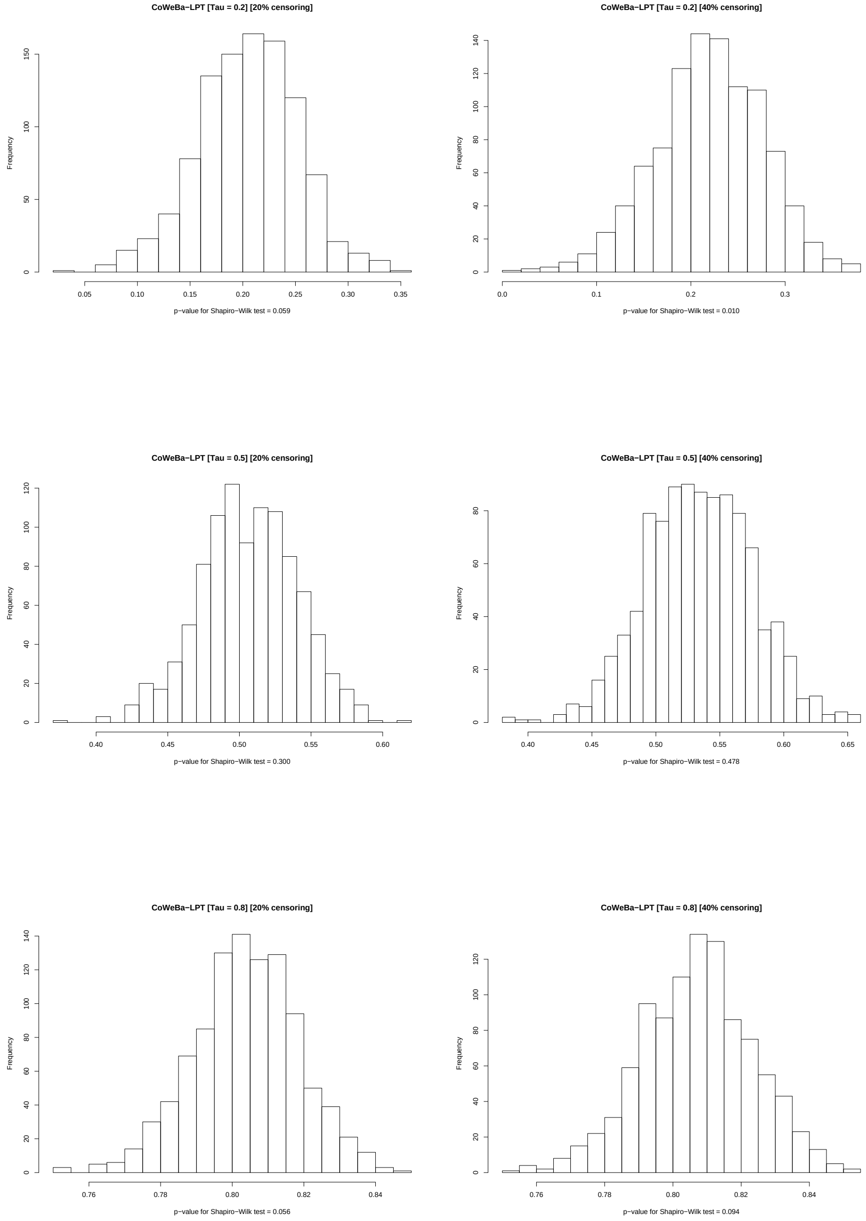


Fig. C.15 – Histogrammes montrant la répartition des estimations de τ selon la **méthode Icdf–VKA (15)**. Chaque graphique est basé sur 1000 échantillons de taille $\mathbf{n = 200}$ issus de la copule de **Frank**. Les colonnes de gauche et de droite correspondent respectivement à $C = 20\%$ et $C = 40\%$. Les valeurs théoriques de τ sont respectivement de 0.2, 0.5 et 0.8 dans les lignes du haut, du milieu et du bas.

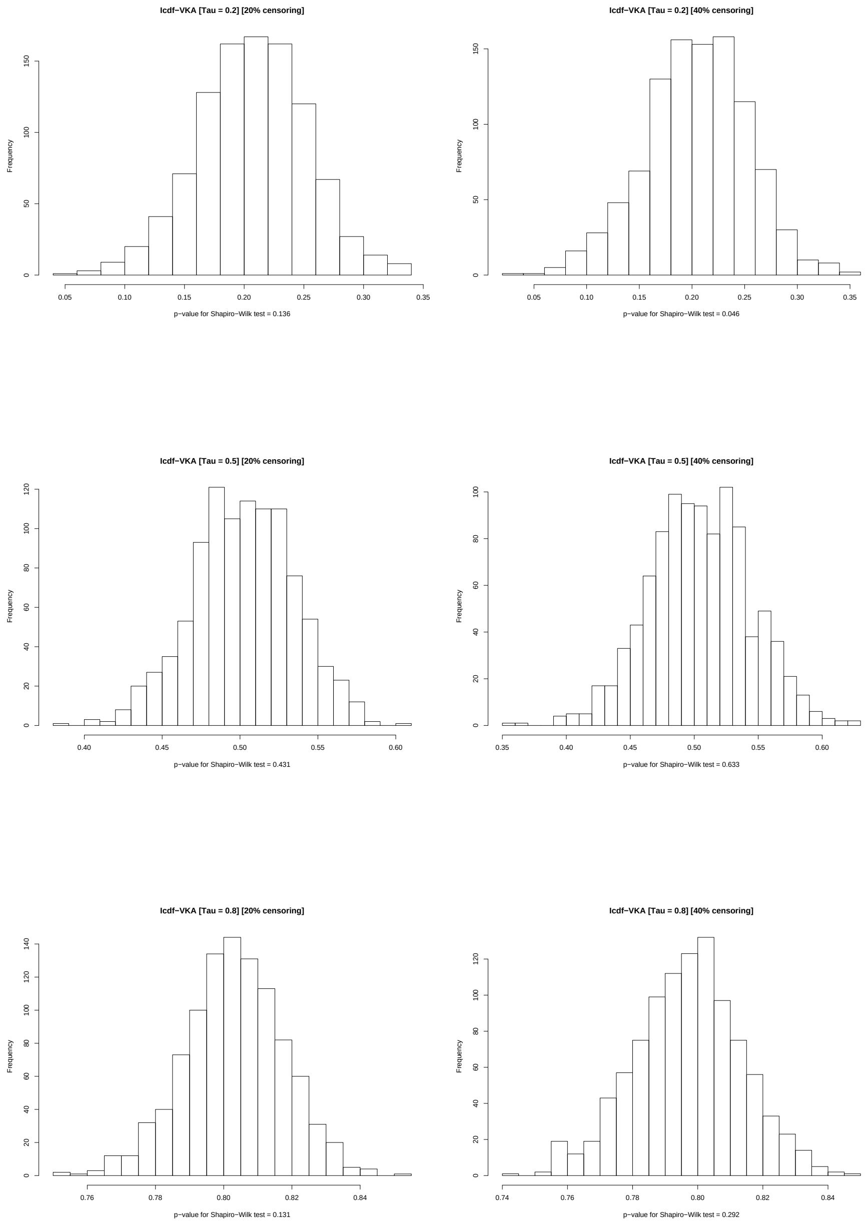


Fig. C.16 – Histogrammes montrant la répartition des estimations de τ selon la **méthode WePa–VKA (9)**. Chaque graphique est basé sur 1000 échantillons de taille $\mathbf{n = 200}$ issus de la copule de **Gumbel–Hougaard**. Les colonnes de gauche et de droite correspondent respectivement à $C = 20\%$ et $C = 40\%$. Les valeurs théoriques de τ sont respectivement de 0.2, 0.5 et 0.8 dans les lignes du haut, du milieu et du bas.

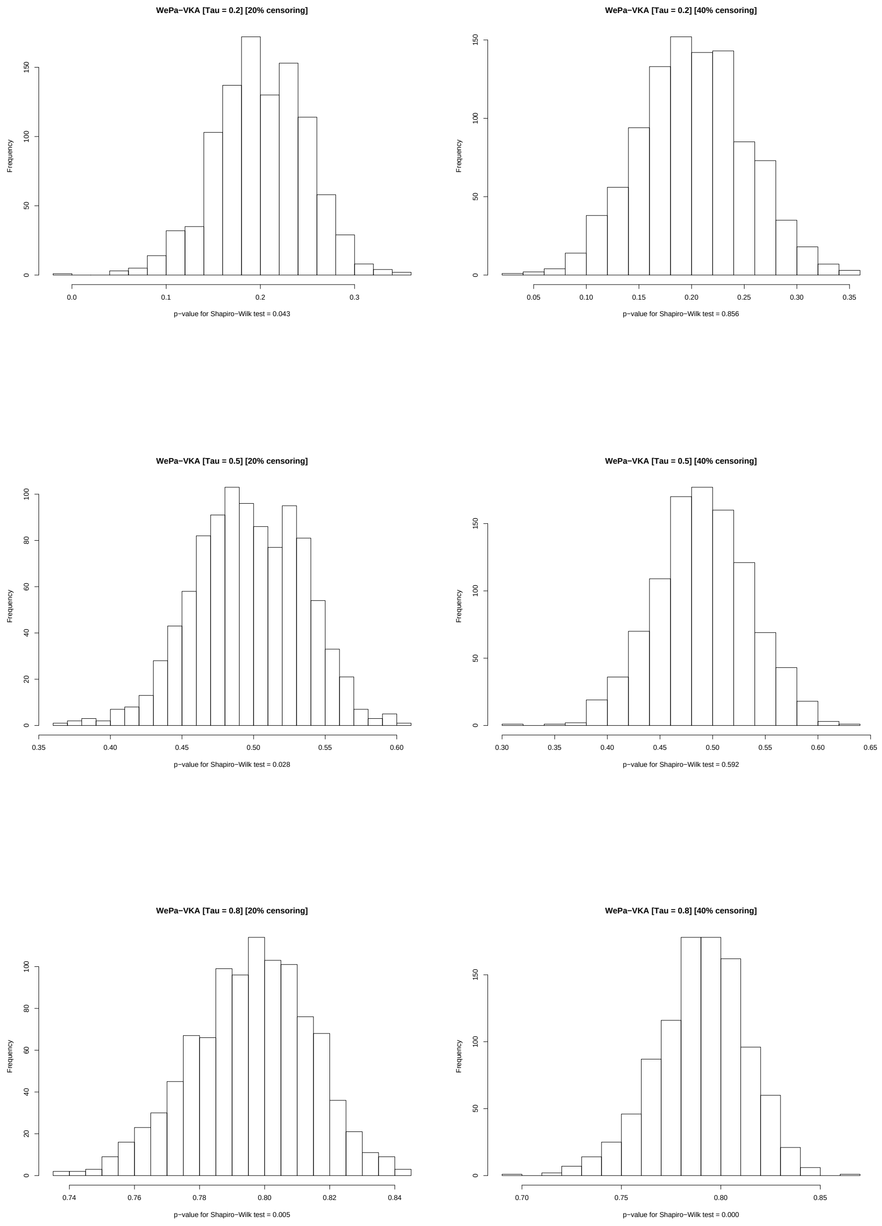


Fig. C.17 – Histogrammes montrant la répartition des estimations de τ selon la **méthode CoWeBa–LPT (11)**. Chaque graphique est basé sur 1000 échantillons de taille $\mathbf{n = 200}$ issus de la copule de **Gumbel–Hougaard**. Les colonnes de gauche et de droite correspondent respectivement à $C = 20\%$ et $C = 40\%$. Les valeurs théoriques de τ sont respectivement de 0.2, 0.5 et 0.8 dans les lignes du haut, du milieu et du bas.

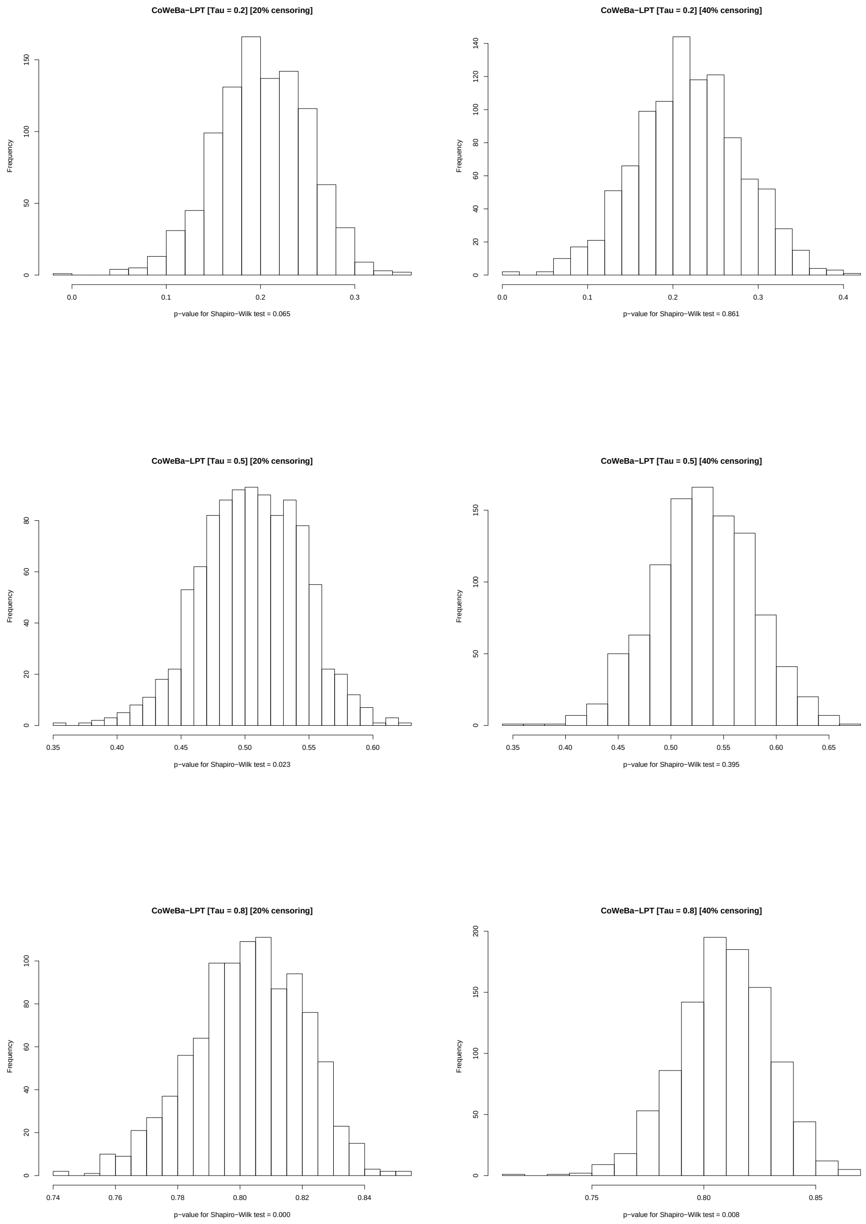
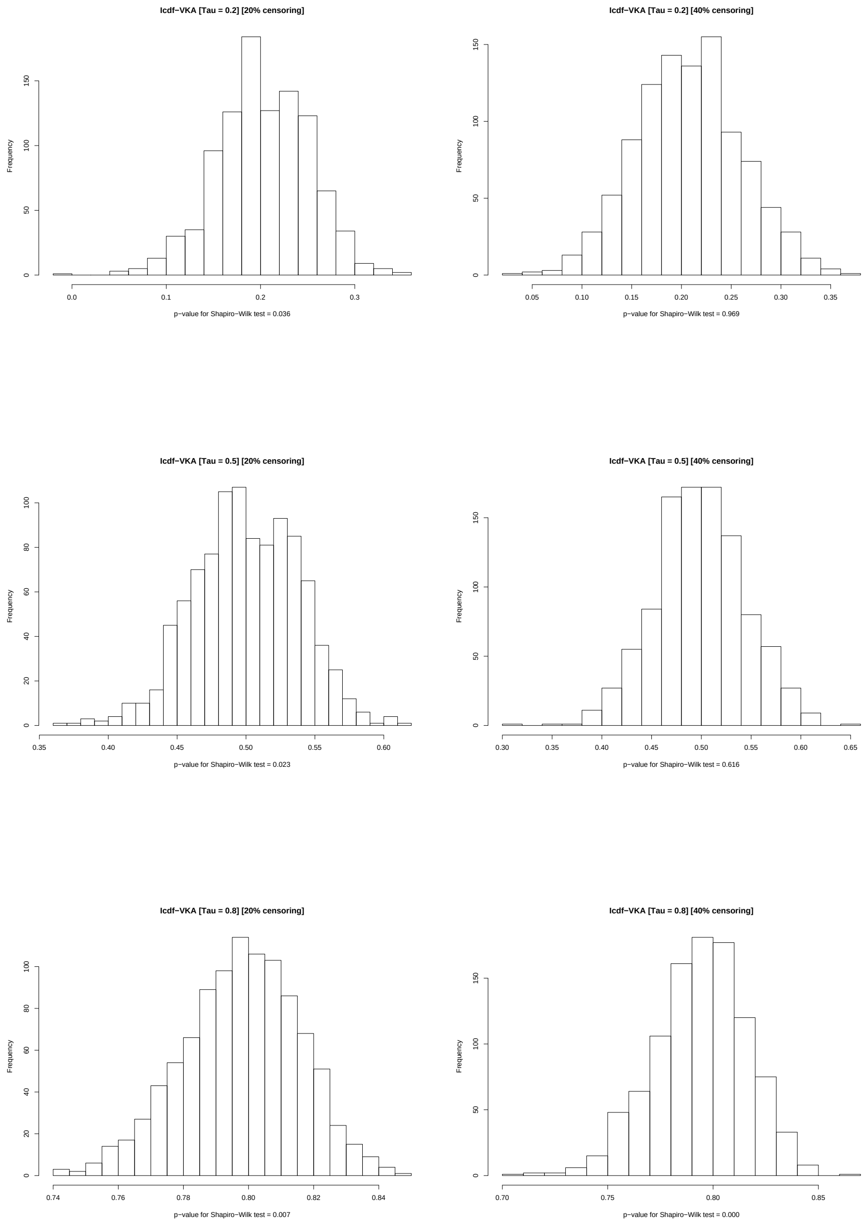


FIG. C.18 – Histogrammes montrant la répartition des estimations de τ selon la **méthode Icdf–VKA (15)**. Chaque graphique est basé sur 1000 échantillons de taille $\mathbf{n = 200}$ issus de la copule de **Gumbel–Hougaard**. Les colonnes de gauche et de droite correspondent respectivement à $C = 20\%$ et $C = 40\%$. Les valeurs théoriques de τ sont respectivement de 0.2, 0.5 et 0.8 dans les lignes du haut, du milieu et du bas.



Tab. C.1 – Seuils observés du test de Shapiro–Wilk (basé sur 1000 observations) pour les trois meilleurs estimateurs de τ pour diverses combinaisons de copule, de tau de Kendall (τ), de fraction de censure (C) et de taille d'échantillon (n). Les seuils inférieurs à 1% (menant au rejet de l'hypothèse de normalité) sont en caractère gras.

Copule	n	τ	C	WePa-VKA	CoWeBa-LPT	Icdf-VKA
Clayton	100	0.2	20%	0.010	0.039	0.012
			40%	0.180	0.050	0.285
		0.5	20%	0.182	0.226	0.166
			40%	**0.002**	**0.000**	**0.000**
		0.8	20%	**0.000**	**0.000**	**0.000**
			40%	**0.008**	**0.000**	**0.003**
Frank	100	0.2	20%	0.545	0.813	0.499
			40%	0.388	0.075	0.330
		0.5	20%	0.841	0.725	0.682
			40%	**0.001**	**0.000**	**0.000**
		0.8	20%	0.010	0.014	0.016
			40%	0.021	**0.000**	0.035
G.–Hougaard	100	0.2	20%	0.478	0.826	0.371
			40%	0.290	0.160	0.249
		0.5	20%	0.039	0.090	0.021
			40%	0.010	**0.000**	**0.007**
		0.8	20%	**0.000**	**0.000**	**0.000**
			40%	**0.000**	**0.000**	**0.000**
Clayton	200	0.2	20%	0.107	0.134	0.135
			40%	0.060	0.027	0.071
		0.5	20%	0.019	**0.008**	**0.007**
			40%	**0.001**	**0.002**	**0.000**
		0.8	20%	0.013	**0.005**	0.010
			40%	0.033	0.067	0.041
Frank	200	0.2	20%	0.196	0.059	0.136
			40%	0.082	0.010	0.046
		0.5	20%	0.451	0.300	0.431
			40%	0.816	0.478	0.633
		0.8	20%	0.076	0.056	0.131
			40%	0.349	0.094	0.292
G.–Hougaard	200	0.2	20%	0.043	0.065	0.036
			40%	0.856	0.861	0.969
		0.5	20%	0.028	0.023	0.023
			40%	0.592	0.395	0.616
		0.8	20%	**0.005**	**0.000**	**0.007**
			40%	**0.000**	**0.008**	**0.000**