Statistics and Its Interface Volume 13 (2020) 103–117

Copula modeling for data with ties

Yan Li, Yang Li*, Yichen Qin, and Jun Yan

Tied observations in copula modeling may cause serious problems to rank-based inference methods that are intended for data with no ties. Simple methods such as breaking the ties at random or using midrank could lead to bias in estimation and invalidity in naive bootstrap inferences. We propose to treat the ranks of tied observations as being interval censored and estimate the copula parameters by maximizing a pseudo-likelihood based on interval censored pseudo-observations. A parametric bootstrap procedure that preserves the tied ranks in the observed data is adapted to do interval estimation and goodness-of-fit test. The proposed approach is shown to be very competitive in comparison to the simple treatments in a large scale simulation study. The utility of the method is illustrated in real data examples.

Keywords and phrases: Interval censored data, Multivariate distribution, Pseudo-observations, Rank-based method.

1. INTRODUCTION

Ties, which are not expected for continuous data but can be present for various reasons, have serious consequences in practical copula modeling. Multivariate modeling based on copulas has been widely applied in many fields such as finance (e.g., Mackenzie and Spears, 2014), actuarial science (e.g., You and Li, 2014), hydrology (e.g., Parent et al., 2014), public heath (e.g., Hu and Liang, 2014), and so on. An important advantage of such models is that the dependence structure of a multivariate distribution is separated from its marginal distributions. Many approaches for copula modeling are rank-based, which do not specify the parametric forms of the marginal distributions (e.g., Genest, Ghoudi and Rivest, 1995; Genest, Ghoudi and Rémillard, 2007). Under the assumption of continuous marginal distributions, there should be no ties in the observed data so the ranks are unique. In many practical settings, however, ties are present in one or multiple margins due to rounding or precision limit of the measurements in observed data. For example, consider analyzing the relationship between two epidemics, hypertension and obesity, using the data from China Health and Nutrition Survey (CHNS; see http://www.cpc.unc.edu/projects/china) of Beijing, China, in 2011. A total of 1,214 observations were available for body mass index (BMI), diastolic blood pressure (DBP), and systolic blood pressure (SBP). The BMI values were rounded to two decimal points while DBP and SBP were collected as integers. Severe ties are present in this dataset, with 911 unique values in BMI, 63 unique values in SBP, and 43 in DBP; see the pseudo-observations scatters in Figure 1. Presence of ties means loss of information on the dependence structure. Further, a naive parametric bootstrap procedure would generate data without ties. Due to the rank-based nature of many copula modeling methods, ties inevitably spoil accuracy and efficiency in estimation and hypotheses tests (Kojadinovic and Yan, 2010; Genest, Neslehova and Ruppert, 2011; Kojadinovic, 2017).

Handling ties appropriately has not received much attention until recently. In rank-based methods, quick but inferior solutions are to use midrank, to break the ties at random multiple times and summarize the multi-data results, or to estimate the parameters via inversion of Kendall's τ estimator. Kojadinovic and Yan (2010) compared the first two naive methods using the bivariate insurance data from Frees and Valdez (1998). Since these naive approaches handle ties in each margin independently and ignore the dependence, they essentially introduce independence into the data. This leads to biased estimation of copula parameters, especially when the dependence is strong. Pappadà, Durante and Salvadori (2016) proposed a randomization strategy where jittering is done by a mixture of the independence copula and the Fréchet-Hoeffding upper bound such that the Kendall's τ matches the empirical Kendall's τ. Such randomization still alters the dependence of the data, albeit less severely than independent or co-monotone randomizations; see illustration in our real data analysis in Section 4. Hypotheses tests in copula modeling based on parametric bootstrap can be severely affected by ties because the bootstrap samples contain no ties, which makes the observed statistics from data with ties look overly large. Consequently, such tests do not hold their sizes by over rejection. Bücher and Kojadinovic (2015) proposed to use the maximum rank in calculating the testing statistics and preserve the observed ties in bootstrap samples. With this approach, tests of exchangeability, radial symmetry, extreme-value dependence, and goodness-of-fit can be adapted in the presence of ties, and have been confirmed in numerical studies (Kojadinovic, 2017).

*Corresponding author.

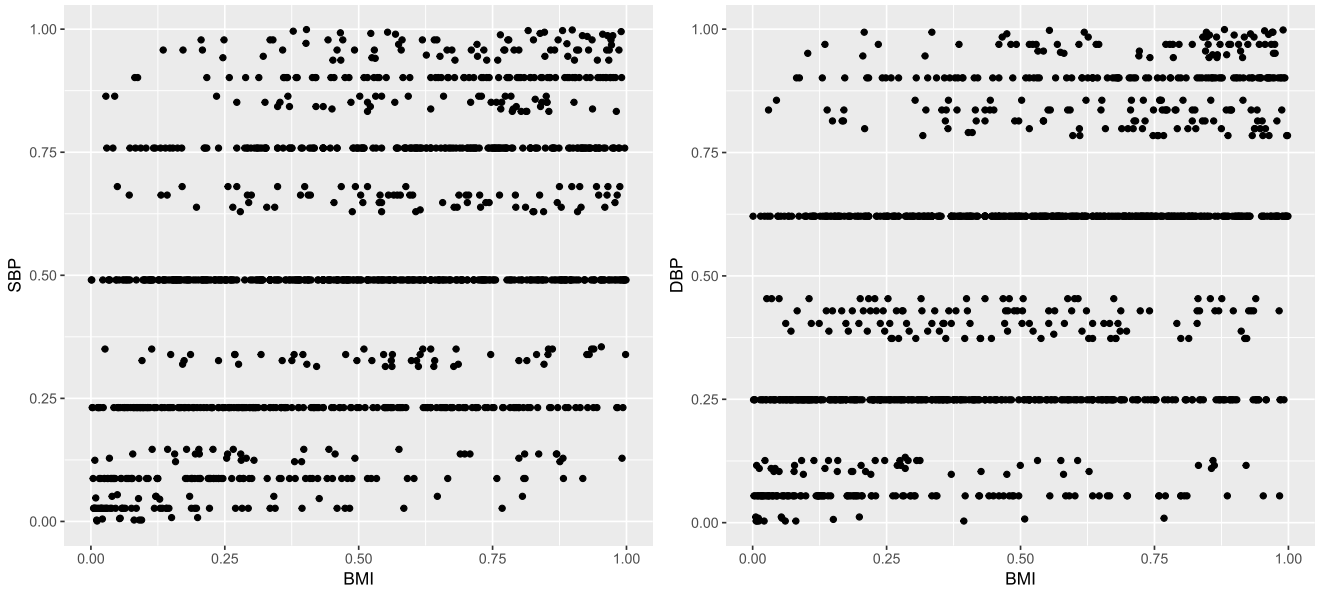

Figure 1. Pseudo-observations (rescaled ranks) of the BMI versus SBP and BMI versus DBP.

We propose to handle tied observations by treating their ranks as being interval censored as in survival analysis (e.g., Sun, 2007; Chen, Sun and Peace, 2012). An observation tied with other observations in one margin means that its rank is known to be in an interval from the minimum rank and maximum rank. Interval censored pseudo-observations can be used to construct a pseudo-likelihood, which can be maximized to obtain point estimates. When an observation is tied in every margin, its contribution to the pseudo-likelihood of the copula is similar to that in copula modeling for discrete data (Nikoloulopoulos and Karlis, 2009; Nikoloulopoulos, 2013; Faugeras, 2017). When an observation is tied but not at all margins, the pseudo-likelihood involves partial derivatives of an appropriate order of the copula distribution function. Further, in parametric bootstrap samples, the tie structure in the observed data needs to be preserved, which is not an issue in full likelihood modeling for discrete data. We apply the tie-preserving parametric bootstrap of Bücher and Kojadinovic (2015) and Kojadinovic (2017) in interval estimation and goodness-of-fit test. In a large scale simulation study, the point estimators are virtually unbiased, the interval estimations provide valid uncertainty measures, and the goodness-of-fit tests maintain their sizes with substantial power. A byproduct of the interval censoring approach is that it works more naturally than the midrank method in nonparametric bootstrap which certainly leads to tied observations from sampling with replacement. In our simulation studies, the midrank method gives appropriate coverage rate only for data without ties, while the censoring method works well in coverage of confidence intervals for data with or without ties.

The rest of this article is organized as follows. In Section 2, we present our interval censoring approach to copula modeling for data with ties, including parameter estimation, interval estimation, and goodness-of-fit test. A large scale numerical study assessing the performance of the proposed methods is reported in Section 3. The methods are illustrated with the CHNS data and the stock price data of the Swiss Market Index (SMI) in Section 4. A discussion concludes in Section 5. An additional simulation study to investigate the performance of the proposed method in nonparametric bootstrap is relegated in the Appendix.

2. METHODOLOGY

2.1 Interval censored pseudo-observations

Let (X, Y) be a continuous random vector with marginal distribution functions F and G, and joint distribution function H. By Sklar's theorem (Sklar, 1959), there is a unique copula $C : [0,1]^2 \to [0,1]$ such that

$$H(x,y) = C\big(F(x), G(y)\big), \quad x, y \in \mathbb{R}.$$

The copula C completely characterizes the dependence structure in H. This representation suggests that the dependence structure can be separated from the marginal distributions in multivariate modeling. Let (X_i, Y_i), $i = 1, \ldots, n$ be a random sample from H. Often, the marginal distributions are modeled by their empirical distributions and the copula is modeled parametrically, leading to a semiparametric inference in multivariate modeling (Genest, Ghoudi and Rivest, 1995). This approach avoids the bias in copula estimation caused by misspecification in marginal distributions (Kim, Silvapulle and Silvapulle, 2007).

Continuous data have no ties and no ambiguity in ranks. Let $\hat{F}_n$ and $\hat{G}_n$ be the empirical distribution functions of F

and G, respectively. Pseudo-observations $U_{n,i}$ and $V_{n,i}$ are simply $\hat{F}_n(X_i)$ and $\hat{G}_n(Y_i)$ rescaled by a constant $n/(n+1)$ to avoid evaluation of the copula density on the edges of the unit square ending at $(1,1)$. That is,

$$(U_{n,i}, V_{n,i}) = \left(\frac{n\hat{F}_n(X_i)}{n+1}, \frac{n\hat{G}_n(Y_i)}{n+1}\right), \tag{1}$$

for $i = 1, \ldots, n$. Without ties, the marginal empirical distribution functions at both margins have jumps of size $1/(n+1)$.

The pseudo-likelihood estimator of θ is constructed from the margin-free pseudo-observations (Genest, Ghoudi and Rivest 1995):

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log c(U_{n,i}, V_{n,i}; \theta),$$

where $c(\cdot,\cdot;\theta)$ is the density of copula function C from a copula family $\{C_\theta : \theta \in \Theta\}$.

When ties are present due to rounding or lack of precision in measurements, the ranks and pseudo-observations are not fully observed but interval censored. An interval censored observation is a data point that is known to be somewhere between two values but the exact value is unknown. For illustration, consider a toy example of 9 observations where the order statistics of the pseudo-observations of X from (1) are

$$(U_{9:1}, \ldots, U_{9:9}) = (1,2,5,5,5,6,8,8,9)/10. \tag{2}$$

In this example, there are ties in the 3rd, 4th, and 5th order statistics and in the 7th and 8th order statistics. If midranks (i.e., averages of ranks) are used, they will be, respectively, 4 and 7.5. Handling ties by their midranks invalidates the parametric bootstrap method because there would be no ties in bootstrap samples. The null distributions of many test statistics cannot be well approximated through a naive parametric bootstrap (Kojadinovic, 2017). Breaking the ties at random gives many possibilities of untied data and introduces independence to the data; the results from the multiple untied data need to be summarized (Kojadinovic and Yan, 2010). As shown in our simulation study, breaking the ties at random can lead to bias in copula estimation when the dependence is high. This is expected because it replaces the dependent interval censored pseudo-observations with independent observations.

We propose to use the concept of interval censored data from survival analysis to handle tied data in copula estimation. In particular, we define upper and lower boundaries of pseudo-observations, respectively, as

$$(\overline{U}_{n,i}, \overline{V}_{n,i}) = \left(\frac{n\hat{F}_n(X_i)}{n+1}, \frac{n\hat{G}_n(Y_i)}{n+1}\right),$$

$$(\underline{U}_{n,i}, \underline{V}_{n,i}) = \left(\frac{n\hat{F}_n(X_i-)+1}{n+1}, \frac{n\hat{G}_n(Y_i-)+1}{n+1}\right).$$

where $\hat{F}_n(x-)$ and $\hat{G}_n(y-)$ are the left limit of $\hat{F}_n$ and $\hat{G}_n$ at x and y, respectively. Note that the upper bounds are the same as $(U_{n,i}, V_{n,i})$. If X_i (or Y_i) is a tied observation, then its pseudo observation $U_{n,i}$ (or $V_{n,i}$) is interval censored by $[\underline{U}_{n,i}, \overline{U}_{n,i}]$ (or $[\underline{V}_{n,i}, \overline{V}_{n,i}]$). If X_i (or Y_i) is not a tied observation, the interval reduces to a single value, i.e., $\underline{U}_{n,i} = \overline{U}_{n,i} = U_{n,i}$ (or $\underline{V}_{n,i} = \overline{V}_{n,i} = V_{n,i}$).

2.2 Pseudo-likelihood estimator

The observation $(U_{n,i}, V_{n,i})$'s contribution to the pseudo-likelihood, $L_i(\theta)$, depends on the censoring pattern on the two margins. There are four cases.

1. If $\underline{U}_{n,i} < \overline{U}_{n,i}$ and $\underline{V}_{n,i} < \overline{V}_{n,i}$ (i.e., the observation is tied observation in both margins), then $L_i(\theta)$ is
$$C_\theta(\overline{U}_{n,i}, \overline{V}_{n,i}) - C_\theta(\overline{U}_{n,i}, \underline{V}_{n,i}) - C_\theta(\underline{U}_{n,i}, \overline{V}_{n,i}) + C_\theta(\underline{U}_{n,i}, \underline{V}_{n,i}).$$
2. If $\underline{U}_{n,i} < \overline{U}_{n,i}$ and $\overline{V}_{n,i} = \underline{V}_{n,i} = V_{n,i}$ (i.e., the observation is a tied observation only in X), then $L_i(\theta)$ is
$$\left.\frac{\partial C_\theta(u,v)}{\partial v}\right|_{u=\overline{U}_{n,i}, v=V_{n,i}} - \left.\frac{\partial C_\theta(u,v)}{\partial v}\right|_{u=\underline{U}_{n,i}, v=V_{n,i}}.$$
3. If $\underline{U}_{n,i} = \overline{U}_{n,i} = U_{n,i}$ and $\overline{V}_{n,i} < \underline{V}_{n,i}$ (i.e., the observation is a tied observation only in Y), then $L_i(\theta)$ is
$$\left.\frac{\partial C_\theta(u,v)}{\partial u}\right|_{u=U_{n,i}, v=\overline{V}_{n,i}} - \left.\frac{\partial C_\theta(u,v)}{\partial u}\right|_{u=U_{n,i}, v=\underline{V}_{n,i}}.$$
4. If $\underline{U}_{n,i} = \overline{U}_{n,i} = U_{n,i}$ and $\overline{V}_{n,i} = \underline{V}_{n,i} = V_{n,i}$ (i.e., the observation is not tied in either margin), then $L_i(\theta) = c(U_{n,i}, V_{n,i}; \theta)$.

The adjusted pseudo-loglikelihood function under interval censoring is

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \log L_i(\theta).$$

The maximum pseudo-likelihood estimator (MPLE) of θ is then

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} \mathcal{L}(\theta). \tag{3}$$

This estimator reduces to the traditional MPLE when neither margin has tied observations. For implementation, we need partial derivatives of the copula in addition to the distribution and density functions. Expressions of these partial derivatives for commonly used copulas are available from the R package `copula` (Hofert et al., 2016).

2.3 Interval estimation

The asymptotic properties of the pseudo-likelihood estimator are challenging to establish due to the inclusion of interval censored pseudo-observations. We resort to bootstrap for confidence intervals, but a plain vanilla parametric bootstrap procedure would not work in this case because no ties would be present if bootstrap samples are generated from the fitted copulas. The parametric bootstrap procedure needs to be modified so that the ties in the observed data are preserved in each of the bootstrap samples in order to sufficiently capture the uncertainty in parameter estimation.

Given a sample generated from the fitted copula, which contains no ties, we introduce ties into the sample such that at each margin the ties in the observed data are reproduced in the bootstrap sample (Bücher and Kojadinovic, 2015). Let $\tilde{F}_n$ and $\tilde{G}_n$ be the empirical distribution of the observed pseudo-observations $U_{n,i}$'s and $V_{n,i}$'s, respectively, i.e., $\tilde{F}_n(u) = \sum_{i=1}^n \mathbb{1}(U_{n,i} \le u)/n$ and $\tilde{G}_n(v) = \sum_{i=1}^n \mathbb{1}(V_{n,i} \le v)/n$. When ties are present, $\tilde{F}_n$ and $\tilde{G}_n$ have jumps of sizes greater than $1/n$. Let $U_{n,i}^{(b^*)}$'s and $V_{n,i}^{(b^*)}$'s be the pseudo-observations from a bootstrap sample, which have no ties, generated from the fitted copula. Ties are introduced into to $U_{n,i}^{(b^*)}$'s and $V_{n,i}^{(b^*)}$'s by applying the corresponding quantile functions $\tilde{F}_n^{-1}$ and $\tilde{G}_n^{-1}$ of $\tilde{F}_n$ and $\tilde{G}_n$ to $U_{n,i}^{(b^*)}$'s and $V_{n,i}^{(b^*)}$'s, respectively:

$$\left(U_{n,i}^{(b)}, V_{n,i}^{(b)}\right) = \left(\tilde{F}_n^{-1}(U_{n,i}^{(b^*)}), \tilde{G}_n^{-1}(V_{n,i}^{(b^*)})\right), \quad i = 1, \dots, n, \tag{4}$$

where $\tilde{F}_n^{-1}(y) = \inf\{u : \tilde{F}_n(u) \ge y, u \in [0,1]\}$ and $\tilde{G}_n^{-1}(y) = \inf\{u : \tilde{G}_n(u) \ge y, u \in [0,1]\}$. After this transformation, $U_{n,i}^{(b)}$'s and $V_{n,i}^{(b)}$'s are tie-adjusted bootstrap pseudo-observations whose marginal empirical distributions are the same as those of $U_{n,i}$'s and $V_{n,i}$'s, respectively. Note that the joint empirical distribution of $(U_{n,i}^{(b)}, V_{n,i}^{(b)})$, however, is not the same as that of $(U_{n,i}, V_{n,i})$, which is the source of variation of the bootstrap sample.

After ties are introduced, we can further obtain the upper and lower boundaries of the adjusted pseudo-observations of $U_{n,i}^{(b)}$'s and $V_{n,i}^{(b)}$'s,

$$\left(\overline{U}_{n,i}^{(b)}, \overline{V}_{n,i}^{(b)}\right) = \left(U_{n,i}^{(b)}, V_{n,i}^{(b)}\right),$$

$$\left(\underline{U}_{n,i}^{(b)}, \underline{V}_{n,i}^{(b)}\right) = \left(\tilde{F}_n^{-1}(U_{n,i}^{(b)}-) + \frac{1}{n+1}, \tilde{G}_n^{-1}(V_{n,i}^{(b)}-) + \frac{1}{n+1}\right).$$

where $\tilde{F}_n$ and $\tilde{G}_n$ are the empirical distribution functions of $U_{n,i}^{(b)}$ and $V_{n,i}^{(b)}$ (and also of $U_{n,i}$ and $V_{n,i}$). Note that

$$\overline{U}_{n:i}^{(b)} = \overline{U}_{n:i}, \qquad \underline{U}_{n:i}^{(b)} = \underline{U}_{n:i},$$

$$\overline{V}_{n:i}^{(b)} = \overline{V}_{n:i}, \qquad \underline{V}_{n:i}^{(b)} = \underline{V}_{n:i}.$$

where the subscript of $A_{n:i}$ represents the ith order statistics (i.e., ith smallest number) of the sequence $\{A_{n,i}\}_{i=1}^n$.

We illustrate the tie-preserving procedure using the same toy example with pseudo-observations (2) in Section 2.1. The bootstrap pseudo-observations (without ties) after being sorted are always

$$(U_{9,1}^{(b^*)}, \dots, U_{9,9}^{(b^*)}) = (1,2,3,4,5,6,7,8,9)/10.$$

By applying (4), we obtain the tie-adjusted bootstrap pseudo-observations

$$(U_{9,1}^{(b)}, \dots, U_{9,9}^{(b)}) = (1,2,5,5,5,6,8,8,9)\,/10,$$

where we have changed 3/10 and 4/10 to 5/10, and 7/10 to 8/10 to match the ties in the observed pseudo-observations. Consequently, the lower and upper boundaries of adjusted pseudo-observations of $(U_{9,1}^{(b)}, \dots, U_{9,9}^{(b)})$ are

$$(\overline{U}_{9,1}^{(b)}, \dots, \overline{U}_{9,9}^{(b)}) = (1,2,5,5,5,6,8,8,9)\,/10,$$
$$(\underline{U}_{9,1}^{(b)}, \dots, \underline{U}_{9,9}^{(b)}) = (1,2,3,3,3,6,7,7,9)\,/10.$$

The same procedure can be applied to the other marginal component $V_{n,i}$.

In summary, the tie-preserving parametric bootstrap procedure given the MPLE $\hat{\theta}_n$ to construct a $1-\alpha$ confidence interval runs as follows. For some large integer B, repeat the following steps 1–3 for every $b \in \{1, \dots, B\}$:

1. Generate bootstrap pseudo-observations with no ties from the fitted copula $C_{\hat{\theta}_n}$.
2. Obtain tie-adjusted pseudo-observations via (4).
3. Obtain the MPLE $\hat{\theta}_n^{(b)}$ using the tie-adjusted pseudo-observations.

A bootstrap sample $(\hat{\theta}_n^{(1)}, \dots, \hat{\theta}_n^{(B)})$ is formed to approximate the sampling distribution of $\hat{\theta}_n$. The sample $\alpha/2$ and $1-\alpha/2$ quantiles can then be used to form a confidence interval of level $1-\alpha$.

The computing cost of the tie-preserving parametric bootstrap procedure is similar to that of the standard parametric bootstrap procedure. The only extra part is the tie-preserving step, which is minimal compared to the optimization in the fitting for each bootstrap sample.

2.4 Goodness-of-fit test

Goodness-of-fit tests with standard parametric bootstrap are known to be vulnerable to ties in keeping their sizes (Kojadinovic and Yan, 2010). This is because goodness-of-fit test statistics (usually distance-based) tend to be bigger for data with ties than for data without ties. Consequently, the tests would not hold their sizes with over rejection. From our numerical studies, the empirical size of a 5%-level test could reach 100% when a moderate amount

of ties are present. Therefore, preserving ties in parametric bootstrap is crucial (Kojadinovic, 2017).

We propose to adapt the standard bootstrap procedure for goodness-of-fit (Genest and Rémillard, 2008) with observed ties-preserved (Bücher and Kojadinovic, 2015; Kojadinovic, 2017). The hypothesis is

$$H_0 : C \in \mathcal{C} = \{C_\theta : \theta \in \Theta\} \quad \text{versus} \quad H_1 : C \notin \mathcal{C}.$$

Consider goodness-of-fit tests based on the goodness-of-fit empirical process

$$\mathbb{C}_n(u, v) = \sqrt{n}(C_n(u, v) - C_{\hat{\theta}_n}(u, v))^2, (u, v) \in [0, 1]^2,$$

where C_n is the empirical copula defined as

$$C_n(u, v) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(U_{n,i} \leq u, V_{n,i} \leq v),$$

and $\hat{\theta}_n$ is a parametric estimator of θ (which could be the MPLE from (3) or other rank-based estimator) under the null hypothesis H_0. Statistics of goodness-of-fit tests can be formed as $\mathcal{F}(\mathcal{C}_n)$, where $\mathcal{F}$ is a functional of $\mathbb{C}_n$. We use the Cramér–von Mises (CvM) statistic, which is known to have a good power (Genest, Rémillard and Beaudoin, 2009; Kojadinovic and Yan, 2010), to illustrate the procedure.

Consider a CvM statistic defined as

$$D_n = \sum_{i=1}^{n} \left(C_n(\overline{U}_{n,i}, \overline{V}_{n,i}) - C_{\hat{\theta}_n}(\overline{U}_{n,i}, \overline{V}_{n,i}) \right)^2. \tag{5}$$

After D_n is obtained, we use the following bootstrap procedure to draw samples from the distribution of D_n under H_0. For some large integer B, repeat the following steps for each $b \in \{1, ..., B\}$:

1. Generate bootstrap pseudo-observations with no ties from the fitted copula $C_{\hat{\theta}_n}$.
2. Obtain tie-adjusted pseudo-observations via (4).
3. Obtain the MPLE $\hat{\theta}_n^{(b)}$ using the tie-adjusted pseudo-observations.
4. Obtain the empirical copula $C_n^{(b)}$ based on the tied-adjusted pseudo-observations.
5. Obtain the bootstrap test statistic $D_n^{(b)}$ using (5).

An approximated p-value of the observed test statistic is then $\sum_{b=1}^{B} \mathbb{1}(D_n^{(b)} \geq D_n)/B$.

Again, this tie-preserving bootstrap procedure has similar computing cost compared to the standard parametric bootstrap procedure. The difference from the procedure of Kojadinovic (2017) is that, after each tie-preserving bootstrap sample is obtained, we use the interval censoring approach for estimation instead of the midrank method.

3. NUMERICAL STUDIES

A large-scale simulation study was carried out to assess the performance of proposed methods in point estimation, interval estimation, and goodness-of-fit.

3.1 Point estimation

We first study the accuracy of the point estimation of the proposed method (denoted as "censoring") in comparison with three naive methods: breaking ties at random (denoted as "random"), using midrank, and inverting the Kendall's tau (denoted as "itau"). For the random method, the mean result from 100 randomizations was used. Data were generated from three one-parameter copulas parameterized by Kendall's τ, Clayton (C), Gumbel (G), and normal (N), with $\tau \in \{0.1, \ldots, 0.9\}$ to control the dependence level. Ties were introduced by rounding the first margin to the first decimal place. Three sample sizes $n \in \{100, 200, 400\}$ were considered. For each setting, 1,000 datasets were generated.

The boxplots of the estimation error of the four estimators from 1000 replicates are summarized in Figure 2. The Gumbel copula is different from the other two copulas in that it only allows positive dependence. In the case of $\tau = 0.1$, the estimate of τ is bounded below by 0, so the error is never below -0.1. This is most obvious when sample size is small ($n = 100$), in which case, all four estimators appear positively biased. The decrease in the variation of the estimates for each sample size as τ goes beyond 0.3 has to do with the scale of τ; the variation on the usual parametrization scale of the Gumbel copula, for example, increases drastically. Other than these, observations from Figure 2 are all as expected. The estimates from the three naive methods have little bias when the dependence is weak (lower τ); as τ increases, however, their bias becomes more obvious, with the random method more severely under-estimate the dependence than the midrank method and the itau method slightly over-estimate the dependence. In contrast, the estimate from the censoring method remains unbiased in all settings. The variances of all four estimators appear comparable for cases with lower τ, but the censoring method seems to have slightly higher variance for cases with higher τ and when τ varies in 0.3–0.7, the itau method seems to have higher variance. Because of its advantage in bias, the censoring method has smallest mean squared error (MSE), especially for higher τ.

We then study the effect of the severity of ties on estimation accuracy. From Figure 2, we see the differences among the methods are not obvious in those cases where Kendall's $\tau = 0.4$ or lower. Thus data were generated from the three copulas with $\tau = 0.75$ and $n = 200$. The first margin was rounded to the first decimal place if its value was smaller than a threshold, which controls the severity of ties; the bigger the threshold, the smaller the percentage of unique observations. The square root of MSE (RMSE) of the four estimators are plotted against the percentage of ties in the first margin in Figure 3. The censoring method has the smallest RMSE, and, unlike the other three methods, its RMSE is stable regardless of the severity level of the ties. The RMSEs of the itau method, the midrank method

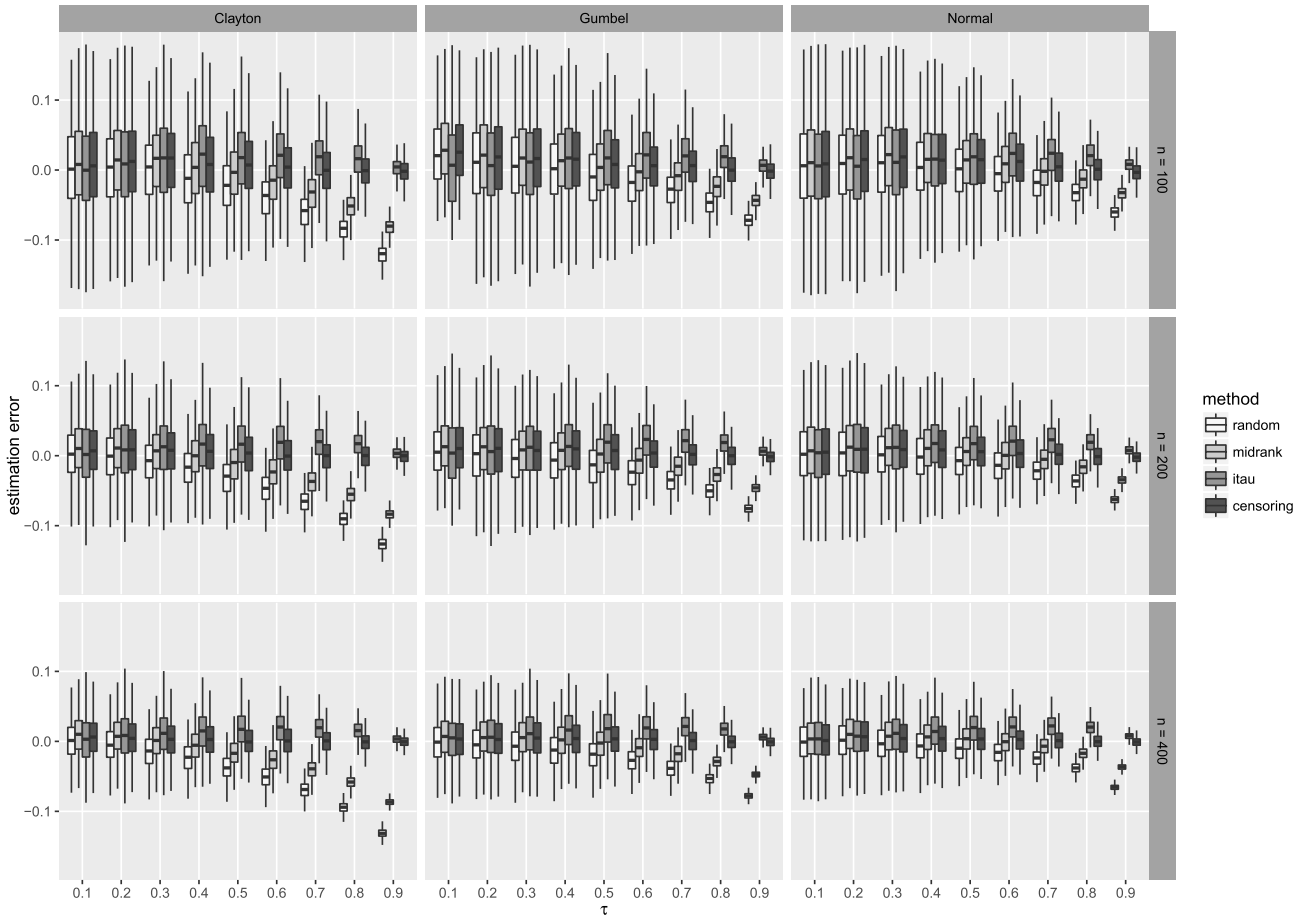

Figure 2. Boxplots of the empirical error in estimating Kendall's τ using four methods (random, midrank, itau, and censoring) for three one-parameter copula families (Clayton, Gumbel, and normal) from 1,000 replicates. Ties were introduced by rounding the first margin to the first decimal place.

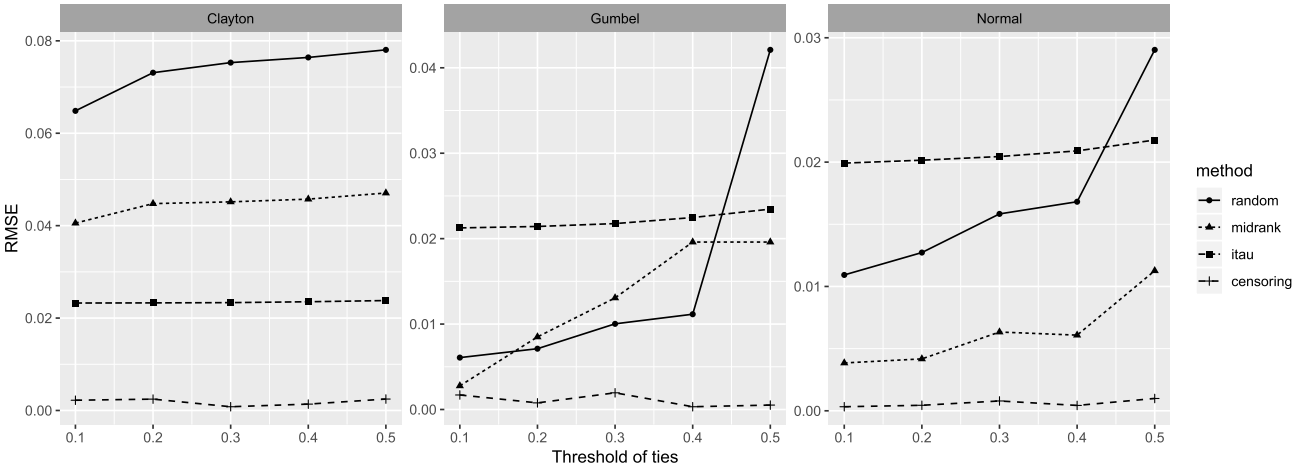

Figure 3. Comparison of RMSE in estimating Kendall's τ using four methods (random, midrank, itau, and censoring) for three one-parameter copula families (Clayton, Gumbel and normal) with different percentages of ties in one margin and sample size $n = 200$. Ties were introduced by rounding the first margin to the first decimal place for observations below the "threshold" on the horizontal axis, which controls the percentage of unique observations.

Table 1. Empirical coverage rate of the 95% bootstrap confidence intervals of the censoring method and the midrank method for three copula families from 1,000 replicates. The bootstrap sample size was $B = 1000$. The values in the parentheses are the average width of the confidence intervals from 1000 replications.

		Clayton		Gumbel		Normal	
n	τ	Censoring	Midrank	Censoring	Midrank	Censoring	Midrank
50	0.25	89.8 (0.324)	87.5 (0.300)	90.6 (0.346)	89.3 (0.335)	88.6 (0.344)	87.4 (0.342)
	0.50	95.4 (0.265)	96.8 (0.250)	92.0 (0.271)	93.6 (0.265)	88.0 (0.245)	89.7 (0.240)
	0.75	96.1 (0.180)	58.5 (0.166)	96.3 (0.164)	94.9 (0.159)	95.3 (0.145)	96.8 (0.143)
100	0.25	92.7 (0.226)	87.5 (0.211)	91.8 (0.247)	89.5 (0.242)	89.3 (0.239)	86.9 (0.237)
	0.50	95.7 (0.189)	97.3 (0.177)	92.5 (0.194)	93.8 (0.188)	91.6 (0.176)	92.2 (0.172)
	0.75	94.4 (0.124)	21.5 (0.114)	93.2 (0.115)	83.8 (0.111)	94.9 (0.101)	92.9 (0.099)
200	0.25	90.8 (0.159)	80.8 (0.149)	92.7 (0.174)	90.6 (0.171)	90.6 (0.167)	87.5 (0.166)
	0.50	94.7 (0.136)	94.6 (0.126)	92.4 (0.140)	94.8 (0.134)	91.8 (0.125)	92.1 (0.122)
	0.75	91.9 (0.087)	2.9 (0.079)	94.4 (0.082)	60.1 (0.078)	93.9 (0.072)	80.8 (0.070)

and the random method all have some increasing patterns as the percentage of ties increases. The increasing rate is the fastest for data generated from the Gumbel copula. The method is much less sensitive to the severity of ties than the other two methods.

3.2 Interval estimation

To assess the coverage properties of the bootstrap confidence intervals, we generated data from the three copulas (C, G, and N) with Kendall's $\tau \in \{0.25, 0.50, 0.75\}$ and sample size $n \in \{50, 100, 200\}$. Ties were introduced by rounding the first margin to the first decimal place. The 95% confidence intervals of the censoring method were constructed with the tie-preserving bootstrap procedure with bootstrap sample size $B = 1000$. The censoring method and midrank method were used to handle ties in data. We compare the performance of these two methods via the coverage rate and the average width of the 95% bootstrap confidence intervals.

Based on the finding from point estimation in Section 3.1 that the proposed method obviously outperforms the others when dependence level is moderate or high, more numerical studies for interval estimation with choices of stronger level of dependence (Kendall's $\tau \in \{0.5, 0.6, \ldots, 0.9\}$) are conducted and presented in Appendix A.1.

The empirical coverage rates of the confidence intervals and their average length based on 1000 replicates are summarized in Table 1. For the censoring method, most of the empirical coverage rates are close to the nominal level in the higher dependence case ($\tau \geq 0.50$). Under low dependence level ($\tau = 0.25$), especially with small sample size ($n = 50$), some empirical coverage rates are just below 90%. As the sample size increases, the width of the confidence interval decreases. As the Kendall's τ increases, the width of the confidence interval decreases as determined by the scale of τ; see Figure 2. Compared with the midrank method, the censoring method performs much better in terms of coverage rates. Although the intervals of the midrank method are slightly narrower, their coverage rates are much lower (with a striking 2.9% for the Clayton copula with $n = 200$ and $\tau = 0.75$) than the censoring method. Recall that, in Figure 2, the bias of midrank method increases rapidly as τ increases, these low coverage rates are expected. Consequently, as the Kendall's τ increase, the advantage of the censoring method in coverage rates becomes more obvious.

We also considered constructing the confidence intervals with the nonparametric bootstrap procedure. The censoring method allows natural processing for the nonparametric bootstrap samples which for sure contain ties even for data with no ties. The censoring method provides appropriate coverage of confidence intervals for both data with no ties and data with ties. See details in Appendix A.2.

3.3 Goodness-of-fit test

The performance of the goodness-of-fit test based on tie-preserving bootstrap using the censoring method in estimation was assessed with data of sample size $n = 100$ generated from the three copulas (C, G, and N) with Kendall's $\tau \in \{0.25, 0.50, 0.75\}$. Three patterns of ties were considered: no ties, ties introduced by rounding one margin, or both margins, to the first decimal place. For each configuration, 500 datasets were generated. For each dataset, goodness-of-fit tests were performed with each of the three copula families (C, G, and N) as the hypothesized copula. The parametric bootstrap sample size was $B = 200$. In the bootstrap procedure, two methods of preserving ties were considered: matching the observed ranks as proposed in Section 2.4, and rounding the margins with ties to the first decimal place. Note that the rounding approach was under the assumption of known tie-introducing mechanism, which is unavailable in general. It was included in the comparison as a benchmark to check whether knowing the tie-introducing mechanism helps to improve the performance of the tests.

Table 2. Empirical rejection percentage of the goodness-of-fit tests at level 5% with sample size $n = 100$ for three copula families (C = Clayton, G = Gumbel, and N = Normal) based on 1000 replicates, each with bootstrap sample size $B = 200$. Ties were introduced by rounding data from the corresponding margin to first decimal place. "No ties" indicate no ties in any sample; "One side" means rounding the first marginal component to the first decimal place; "Two sides" means rounding both components to the first decimal place.

Ties pattern	Kendall's τ	True copula	Hypothesized copula: C Match	C Round	G Match	G Round	N Match	N Round
No ties	0.25	C	**5.5**		62.0		15.9	
		G	79.1		**5.3**		16.8	
		N	47.7		13.6		**5.2**	
	0.50	C	**7.0**		96.8		60.0	
		G	99.0		**6.2**		28.6	
		N	88.2		24.0		**5.0**	
	0.75	C	**4.0**		100.0		85.8	
		G	100.0		**3.0**		31.0	
		N	98.2		21.6		**3.0**	
One side	0.25	C	**4.4**	**4.2**	57.6	58.9	3.8	18.2
		G	76.5	75.9	**4.2**	**3.4**	18.3	18.6
		N	44.9	45.5	12.7	12.1	**3.5**	**6.4**
	0.50	C	**4.6**	**5.4**	95.4	95.6	52.2	55.2
		G	99.8	99.6	**6.2**	**6.6**	32.4	33.6
		N	87.0	88.2	22.2	21.4	**3.6**	**4.2**
	0.75	C	**1.6**	**3.8**	99.6	99.6	79.4	79.6
		G	100	100	**4.0**	**4.2**	25.6	26.6
		N	96.6	97.0	14.4	15.4	**3.4**	**3.8**
Two sides	0.25	C	**6.4**	**4.4**	51.4	55.6	13.8	15.4
		G	71.7	73.5	**4.7**	**3.7**	19.4	18.8
		N	40.2	38.6	8.8	11.8	**5.0**	**4.2**
	0.50	C	**4.6**	**3.8**	96.0	96.8	53.2	54.6
		G	98.6	99.0	**4.2**	**5.8**	28.6	31.8
		N	82.8	85.0	19.2	18.6	**5.6**	**4.8**
	0.75	C	**0.4**	**4.2**	97.8	98.0	75.0	83.2
		G	99.6	100.0	**4.4**	**5.4**	26.6	30.6
		N	93.6	94.8	10.6	15.0	**4.6**	**4.4**

The empirical rejection percentages of the goodness-of-fit tests with significance level 5% are summarized in Table 2. When the hypothesized copula is the same as the data generating copula, the reported percentages are put in bold, representing the empirical sizes. The empirical sizes are close to the nominal size of 5% in most cases. The two methods of preserving ties show little difference, except that the test is conservative for the Clayton copula with $\tau = 0.75$, with empirical rejection percentage 1.6 and 0.4 for one- and two-side ties, respectively. When the hypothesized copula is not the data generating copula, the empirical powers of the tests are lower than those obtained when no ties are present, which is as expected due to the information loss in ties. Between the two tie-preserving methods, the rounding approach seems to have slightly higher power, but the advantage seems minimal. Note that the rounding approach is not available in practice.

Since the difference between the two tie-preserving methods is little, we focus on the matching ties method and investigate sample sizes 50 and 200. The empirical rejection percentages are summarized in Table 3. For sample size 50, the test appears to be a little conservative when $\tau = 0.75$. Nonetheless, the test holds its size closely at sample size 200, and the power increases as the sample size increases in all settings as expected. The power in the case of one side ties

Table 3. Empirical rejection percentages of the goodness-of-fit tests with sample size $n \in \{50, 200\}$ for three types of copulas (C = Clayton, G = Gumbel, and N = Normal) based on 1000 replicates, each with bootstrap sample size $B = 200$. Ties in the original data were introduced by rounding data from the first margin to first decimal place. Matching rank was used to preserve ties in bootstrap samples.

Ties pattern	Kendall's τ	True copula	Hypothesized copula: C, $n=50$	C, $n=200$	G, $n=50$	G, $n=200$	N, $n=50$	N, $n=200$
One side	0.25	C	**4.3**	**6.4**	32.5	88.8	6.0	41.8
		G	50.7	94.2	**6.9**	**5.4**	9.4	30.4
		N	24.1	69.4	10.7	25.6	**2.6**	**6.8**
	0.50	C	**3.2**	**4.0**	72.9	100.0	21.9	93.8
		G	89.2	100.0	**4.7**	**5.2**	21.8	42.2
		N	59.2	99.8	11.6	38.6	**5.0**	**4.6**
	0.75	C	**1.1**	**5.2**	83.7	100.0	35.1	99.0
		G	94.6	100.0	**2.9**	**5.8**	17.1	47.6
		N	73.3	100.0	7.4	32.2	**3.9**	**4.0**
Two sides	0.25	C	**4.6**	**4.4**	11.9	55.6	1.5	18.0
		G	46.5	73.5	**3.3**	**3.7**	10.8	19.2
		N	24.2	38.6	6.8	11.8	**3.9**	**4.8**
	0.50	C	**2.9**	**3.8**	53.2	96.8	8.4	55.0
		G	86.0	99.0	**2.5**	**5.8**	16.1	26.4
		N	57.2	85.0	10.3	18.6	**3.4**	**4.2**
	0.75	C	**0.7**	**4.2**	75.9	98.0	34.6	78.4
		G	89.8	100.0	**1.0**	**5.4**	13.9	30.6
		N	60.2	94.8	2.8	15.0	**1.8**	**5.0**

is substantially higher than at in the case of two-side ties, a result of less information loss in the former.

4. REAL DATA EXAMPLES

4.1 Hypertension and obesity

For the aforementioned CHNS data, we consider all of the three pairs: (SBP, BMI), (DBP, BMI), and (DBP, SBP). As shown in Figure 1, many ties are present in the data. Out of the 1,214 observations, DBP and SBP have only 43 and 63 unique values, respectively. Ties are much fewer in BMI, with 911 unique values. So the (SBP, BMI) and (DBP, SBP) pairs have a large number of ties in one margin, while the (SBP, DBP) pair has both margins with heavy ties. For each pair, we fitted four one-parameter copulas (Clayton, Gumbel, normal, and Frank) parameterized by Kendall's τ, estimated the standard errors of the parameter estimates, and checked goodness-of-fit using the proposed methods in Section 2. The tie-preserving bootstrap procedure was run with $B = 10000$. Results from the midrank method and itau method were also obtained for comparison.

Table 4 summarizes the estimation and goodness-of-fit results from the three methods. All methods give very similar results. At level 5%, only the normal copula passes the goodness-of-fit test for the (DBP, BMI) pair, with p-values well around or above 15%. While the p-value from itau method is obviously larger than the results from the other two methods. The point estimates from the midrank method are very close to those from the censoring method in all cases for the (SBP, BMI) pair and the (DBP, BMI) pair. This is not surprising as our simulation study suggests more visible bias of the midrank method for higher τ. The bootstrap standard errors from the two methods are also similar, with those from the censoring method more likely to be slightly higher. The point estimates and the corresponding standard errors from the itau method are obviously different from those from the other two methods in all the case. This is expected since from our simulation studies the itau method gives more visible bias for moderate dependence level. As for the (SBP, DBP) pair, since the dependence level is higher, the estimates of τ from the censoring method are higher than those from the midrank method and are much lower than the empirical Kendall's τ. This is consistent with the results in the simulation studies: the itau method tends to overestimate while the midrank method tends to underestimate τ when the dependence is strong. The bootstrap standard errors from the censoring method are slightly larger than those from the other two methods.

4.2 Stock prices

Consider the closing prices of two SMI constituents, Swatch Group (UHR) and Cie Financiere Richemont

Table 4. Estimate ($\hat{\tau}_n$) of τ and standard errors (se) obtained from three methods (midrank, itau, censoring) for three pairs (SBP, BMI), (DBP, BMI), and (DBP, SBP) from the CHNS data fitted to four copulas, along with the p-values of the goodness-of-fit (GoF) test. The standard errors are in parentheses. Standard errors and p values are computed by taking average over $B = 10000$ replicates. All values reported are in percentage.

	Clayton		Gumbel		Normal		Frank	
Method	$\hat{\tau}_n$ (se)	GoF	$\hat{\tau}_n$ (se)	GoF	$\hat{\tau}_n$ (se)	GoF	$\hat{\tau}_n$ (se)	GoF
Pair #1: (SBP, BMI)								
	Percentage of unique values: (5.2, 75.0); empirical τ: 32.2							
midrank	26.9 (1.66)	0.02	25.9 (1.82)	0.00	30.2 (1.62)	0.04	30.8 (1.69)	0.00
itau	32.2 (1.89)	1.72	32.2 (1.88)	0.00	32.2 (1.79)	0.02	32.2 (1.76)	0.03
censoring	27.0 (1.69)	0.04	25.9 (1.83)	0.00	30.2 (1.66)	0.04	31.0 (1.70)	0.04
Pair #2: (DBP, BMI)								
	Percentage of unique values: (3.5, 75.0); empirical τ: 32.8							
midrank	26.2 (1.62)	0.00	27.1 (1.81)	0.00	30.6 (1.63)	15.59	31.4 (1.70)	3.57
itau	32.8 (1.90)	0.00	32.8 (1.95)	0.00	32.8 (1.80)	33.97	32.8 (1.75)	2.40
censoring	26.3 (1.70)	0.00	27.2 (1.84)	0.00	30.6 (1.67)	15.76	31.7 (1.71)	3.15
Pair #3: (DBP, SBP)								
	Percentage of unique values: (3.5, 5.2); empirical τ: 61.6							
midrank	48.0 (1.33)	0.00	53.5 (1.35)	0.00	56.1 (1.14)	0.00	55.8 (1.14)	0.00
itau	61.6 (1.34)	0.00	61.6 (1.29)	0.00	61.6 (1.19)	0.59	61.6 (1.12)	0.00
censoring	51.0 (1.49)	0.00	54.4 (1.43)	0.00	57.0 (1.22)	0.00	57.7 (1.26)	0.00

(CFR), from 2011-09-09 to 2012-03-28. The prices of the two stocks are highly correlated. The data are available from the R package `copula` (Hofert et al., 2016). The 141 daily observations contained a few ties in both margins, with 135 unique values in UHR and 130 unique values in CFR. For illustration purpose, we introduced artificial ties to the original data to see how different methods perform in response. Ties were introduced by rounding, which could happen in many financial applications (Frees and Valdez, 1998). Two rounding mechanisms were used: rounding on the log scale of original data to the second decimal place or rounding the pseudo-observations of the original data to the first decimal place. Rounding was conducted in the first margin or both margins of pair for different ties scenarios. Compared with the single margin rounding, rounding data in both margins will cause more severe problem, as is shown in the following Figure 4. We repeated the same estimation and goodness-of-fit test in each scenario as we do in above Section 4.1. The tie-preserving bootstrap procedure was also run with $B = 10000$.

Table 5 summarizes the estimation and goodness-of-fit results from the three methods with four one-parameter copulas (Clayton, Gumbel, normal and Frank) under the different rounding mechanisms.

Firstly, for the point estimations of Kendall's τ from three methods, we do not see many differences in most cases except the cases when Clayton copula is fitted to the pair. For the Clayton copula, the estimate $\hat{\tau}$ from itau method is obviously larger than those from other methods under each of the five severity levels of ties. While for the midrank and censoring method, the results are still similar. Another thing we should note is that as the severity of ties increases, the empirical Kendall's τ or estimate from itau method has a noticeable increasing trend. The mixed randomization strategy of Pappadà, Durante and Salvadori (2016) would therefore introduce distortion to the dependence structure of the original data. In contrast, the $\hat{\tau}$s from the midrank and censoring methods remain stable regardless of the severity level of ties. Then as for the estimated standard error, we do not see much significant difference between the results from the midrank and censoring methods. While compared to the results from above two methods, the estimated standard errors from itau method are much smaller in most cases except for the case when Gumbel, Normal and Frank copulas are fitted to the original pair in which there is very few ties. As the severity of ties increases, we see a noticeable decreasing trend in the estimated standard errors from the itau method, while the results from other two methods are stable to the changes of severity of ties. Finally, with respect to the goodness-of-fit test, the Clayton, Normal and Frank copulas are rejected at 5% level regardless of the severity of ties. While for the Gumbel copula, we get different conclusions. When fitted to the original pair, the Gumbel copula is rejected. Nonetheless, as ties are introduced to the pair, the p values from midrank method lead to the conclusion that the Gumbel copula cannot be rejected, with the p value changing from 1.88% to 13.36% when rounding only the first margin of pair and from 1.88% to 7.58% in the cases that both margins are rounded. The above results are consistent with the results in the simulation studies: the censoring method

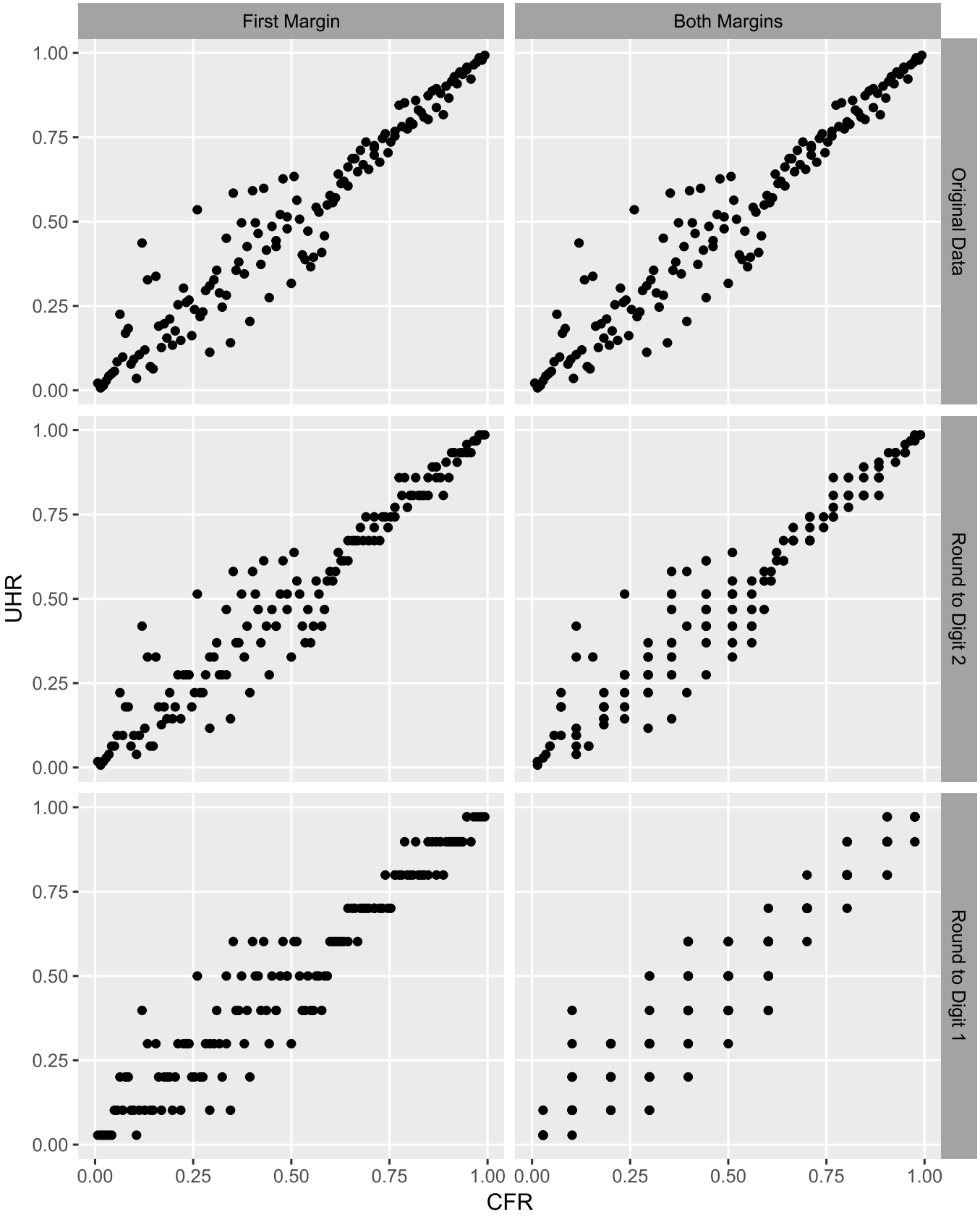

Figure 4. Pseudo-observations of the stock prices of (UHR, CRR) under different rounding mechanism.

Table 5. Estimate ($\hat{\tau}_n$) of τ and standard errors (se) obtained from three methods (midrank, itau, censoring) for the closing prices of (UHR, CFR) fitted to four copulas, along with the p-values of the goodness-of-fit tests, under different rounding mechanisms, including rounding first margin or both margins and combined with rounding log scale and pseudo-observation. The standard errors are in parentheses. Standard errors and p values are computed by taking average over $B = 10000$ replicates. All values reported are in percentage.

	Clayton		Gumbel		Normal		Frank	
Method	$\hat{\tau}_n$ (se)	GoF	$\hat{\tau}_n$ (se)	GoF	$\hat{\tau}_n$ (se)	GoF	$\hat{\tau}_n$ (se)	GoF
Original data								
Percentage of unique values: (95.7, 92.2); empirical τ: 84.9								
midrank	71.9 (2.77)	0.00	85.2 (1.53)	1.88	83.2 (1.49)	0.03	83.9 (1.33)	0.00
itau	84.9 (1.82)	0.00	84.9 (1.67)	0.75	84.9 (1.48)	0.00	84.9 (1.26)	0.00
censoring	71.9 (2.80)	0.00	85.2 (1.54)	1.93	83.2 (1.51)	0.05	83.9 (1.32)	0.00
Round first margin								
Round log scale of pair to digit 2; percentage of unique values: (23.4, 92.2); empirical τ: 86.0								
midrank	71.9 (2.72)	0.00	84.8 (1.55)	6.52	83.4 (1.48)	0.09	83.9 (1.32)	0.00
itau	86.0 (1.71)	0.00	86.0 (1.52)	0.89	86.0 (1.37)	0.02	86.0 (1.18)	0.00
censoring	72.4 (2.73)	0.00	85.5 (1.52)	3.59	83.5 (1.50)	0.07	84.0 (1.38)	0.00
Round pobs of pair to digit 1; percentage of unique values: (7.8, 92.2); empirical τ: 85.6								
midrank	69.2 (2.74)	0.00	81.6 (1.74)	13.36	81.5 (1.57)	0.15	82.3 (1.46)	0.03
itau	85.6 (1.67)	0.00	85.6 (1.52)	0.74	85.6 (1.38)	0.00	85.6 (1.19)	0.00
censoring	73.9 (2.76)	0.00	84.3 (1.73)	2.81	82.8 (1.63)	0.15	83.0 (1.56)	0.03
Round both margins								
Round log scale of pair to digit 2; percentage of unique values: (23.4, 24.1); empirical τ: 86.9								
midrank	72.0 (2.66)	0.00	84.7 (1.51)	8.71	83.0 (1.50)	0.11	83.6 (1.31)	0.01
itau	86.9 (1.60)	0.00	86.9 (1.45)	0.56	86.9 (1.29)	0.03	86.9 (1.12)	0.01
censoring	73.3 (2.70)	0.00	85.4 (1.53)	4.42	83.4 (1.50)	0.10	83.9 (1.47)	0.00
Round pobs of pair to digit 1; percentage of unique values: (7.8, 7.8); empirical τ: 88.0								
midrank	72.7 (3.12)	0.00	82.5 (1.98)	7.58	81.5 (1.66)	0.17	82.1 (1.53)	0.04
itau	88.0 (1.45)	0.00	88.0 (1.32)	0.38	88.0 (1.21)	0.00	88.0 (1.05)	0.00
censoring	77.3 (2.59)	0.00	84.5 (1.79)	2.62	83.6 (1.71)	0.10	83.4 (1.56)	0.03

is more stable than the other two methods with respect to handling ties.

5. DISCUSSION

The interval censoring method to handle ties does not distort the features of the observed data. This is in contrast to the midrank method, the independence randomization method (Kojadinovic and Yan, 2010), or the comonotone/mixed randomization method (Pappadà, Durante and Salvadori 2016). Consequently, it does not have the bias that other approaches may have introduced, especially when the dependence is strong. When the dependence is weak, although the point estimates from the censoring method may not be very different from those from the midrank method, the small difference might still propagate to a noticeable impact when estimation is repetitively needed as in the case of parametric bootstrap procedures.

Our study provides a proof of concept of the interval censoring method to handle ties in the bivariate case. Extending it to higher dimensional cases is straightforward in principle, but the implementation may not be trivial as it requires higher-order cross partial derivatives of the distribution function of the copula. The limiting distribution of the MPLE under interval censoring is a challenging problem because the interval censored data used in the estimation are pseudo-observations resulting from the probability integral transformation with marginal empirical distribution functions, instead of the real observations. Establishing the asymptotic properties of the MPLE from interval-censored pseudo-observations like Genest, Ghoudi and Rivest (1995) did for the case of no ties would be of interest.

The tie-preserving parametric bootstrap procedure provides valid finite sample inferences for the estimator from the interval censoring method. The procedure has been applied to several inference problems for copula modeling for data with ties in Kojadinovic (2017), such as tests for

Table A.1. Empirical coverage rate of the 95% bootstrap confidence intervals of the censoring method and the midrank method for three copula families from 1,000 replicates. Kendall's $\tau \in \{0.5, 0.6, \ldots, 0.9\}$ The bootstrap sample size was $B = 1000$. The values in the parentheses are the average width of the confidence intervals from 1000 replications.

		Clayton		Gumbel		Normal	
n	τ	Censoring	Midrank	Censoring	Midrank	Censoring	Midrank
50	0.5	95.9 (0.26)	98.6 (0.25)	93.1 (0.27)	94.8 (0.26)	91.5 (0.25)	92.2 (0.24)
	0.6	95.9 (0.24)	96.5 (0.22)	95.0 (0.23)	98.1 (0.23)	92.8 (0.21)	95.5 (0.20)
	0.7	95.6 (0.20)	81.0 (0.19)	96.0 (0.19)	96.8 (0.18)	96.7 (0.17)	97.7 (0.16)
	0.8	93.7 (0.16)	3.7 (0.16)	96.8 (0.14)	78.9 (0.14)	97.4 (0.13)	89.4 (0.12)
	0.9	88.0 (0.11)	0.0 (0.12)	90.8 (0.09)	0.0 (0.10)	87.1 (0.09)	0.0 (0.08)
100	0.5	96.7 (0.19)	98.2 (0.18)	92.8 (0.19)	96.4 (0.19)	90.9 (0.18)	90.4 (0.17)
	0.6	96.0 (0.17)	90.1 (0.16)	95.4 (0.17)	97.4 (0.16)	93.7 (0.15)	96.0 (0.14)
	0.7	95.7 (0.14)	38.9 (0.13)	94.6 (0.13)	93.4 (0.13)	95.8 (0.12)	97.0 (0.11)
	0.8	93.1 (0.11)	0.0 (0.11)	94.7 (0.10)	33.0 (0.10)	97.3 (0.09)	75.9 (0.08)
	0.9	85.5 (0.07)	0.0 (0.08)	89.5 (0.06)	0.0 (0.06)	84.9 (0.06)	0.0 (0.06)
200	0.5	94.7 (0.14)	92.8 (0.13)	94.4 (0.14)	96.8 (0.14)	93.0 (0.13)	93.0 (0.12)
	0.6	95.2 (0.12)	64.9 (0.11)	93.9 (0.12)	93.6 (0.12)	94.3 (0.11)	96.2 (0.10)
	0.7	95.0 (0.10)	1.6 (0.09)	95.3 (0.09)	79.0 (0.09)	95.7 (0.08)	93.8 (0.08)
	0.8	93.8 (0.07)	0.0 (0.08)	95.4 (0.07)	1.6 (0.07)	95.3 (0.06)	36.0 (0.06)
	0.9	90.6 (0.04)	0.0 (0.06)	90.0 (0.04)	0.0 (0.05)	87.2 (0.04)	0.0 (0.04)

Table A.2. Empirical coverage rate of the 95% bootstrap confidence intervals of the censoring method and the midrank method for three copula families from 1,000 replicates. Two conditions are considered: the original pair does not contain any ties and the first margin of original pair has many ties. Nonparametric bootstrap with replacement is used. The bootstrap sample size was $B = 200$. The values in the parentheses are the average width of the confidence intervals from 1000 replications.

		Clayton		Gumbel		Normal	
n	τ	Censoring	Midrank	Censoring	Midrank	Censoring	Midrank
Original pair does not contain any ties							
50	0.25	92.0 (0.328)	91.8 (0.330)	93.5 (0.334)	93.5 (0.335)	92.5 (0.343)	92.3 (0.343)
	0.5	93.6 (0.272)	94.2 (0.277)	95.0 (0.280)	95.1 (0.282)	90.9 (0.248)	91.0 (0.249)
	0.75	95.9 (0.193)	96.9 (0.204)	96.4 (0.178)	96.7 (0.183)	95.7 (0.155)	96.1 (0.157)
100	0.25	92.4 (0.223)	92.4 (0.224)	94.1 (0.242)	94.1 (0.242)	92.5 (0.232)	92.5 (0.232)
	0.5	95.5 (0.190)	95.7 (0.192)	93.4 (0.195)	93.6 (0.196)	91.9 (0.173)	91.9 (0.174)
	0.75	94.7 (0.128)	95.4 (0.133)	94.1 (0.119)	94.3 (0.121)	93.9 (0.102)	94.1 (0.103)
200	0.25	93.4 (0.155)	93.4 (0.155)	95.2 (0.171)	95.2 (0.172)	94.0 (0.162)	94.2 (0.162)
	0.5	94.8 (0.134)	94.8 (0.134)	93.7 (0.138)	93.7 (0.138)	91.8 (0.121)	91.8 (0.121)
	0.75	93.4 (0.085)	94.3 (0.087)	92.8 (0.080)	93.2 (0.081)	94.6 (0.070)	94.8 (0.070)
First margin of original pair is rounded to the first digit							
50	0.25	90.5 (0.335)	90.3 (0.330)	94.8 (0.336)	95.0 (0.333)	91.9 (0.346)	91.6 (0.346)
	0.5	93.7 (0.273)	94.2 (0.253)	93.7 (0.281)	94.2 (0.273)	93.0 (0.252)	93.5 (0.247)
	0.75	96.4 (0.191)	83.9 (0.164)	95.4 (0.180)	95.1 (0.167)	95.1 (0.157)	95.4 (0.149)
100	0.25	91.5 (0.225)	91.3 (0.220)	91.6 (0.244)	91.8 (0.242)	94.0 (0.237)	93.5 (0.237)
	0.5	93.7 (0.190)	93.6 (0.172)	92.1 (0.195)	93.0 (0.188)	92.1 (0.175)	92.5 (0.171)
	0.75	95.0 (0.127)	53.3 (0.102)	93.7 (0.119)	90.8 (0.107)	94.2 (0.106)	93.6 (0.099)
200	0.25	92.1 (0.157)	92.3 (0.152)	92.3 (0.172)	92.4 (0.170)	92.5 (0.163)	91.6 (0.164)
	0.5	94.7 (0.135)	93.2 (0.120)	94.3 (0.139)	94.7 (0.132)	93.5 (0.123)	93.4 (0.121)
	0.75	94.1 (0.088)	8.1 (0.067)	94.2 (0.082)	76.8 (0.072)	94.7 (0.072)	88.1 (0.067)

exchangeability, extreme-value dependence, radial symmetry, and goodness-of-fit. Most of these tests require nonparametric estimation of copula in the presence of ties, which has been done with the midrank method. Development of nonparametric copula estimation in the presence of ties with interval censored pseudo-observations may lead to more efficient test procedures when the dependence is strong.

APPENDIX A

A.1 Additional simulation studies for interval estimation

Based on the conclusions from point estimation, we repeat the numerical studies in Table 1 with choices of moderate to high levels of dependence, in order to show more clearly the differences between our proposed censoring approach and the midrank rank method. The results are presented in the following Table A.1. The Kendall's τ are $\tau \in \{0.5, 0.6, \ldots, 0.9\}$ and ties are still introduced by rounding the first margin to the first digit.

From the results, we can see that in most cases, the coverage rate from our proposed method is much closer to the nominal level 95% with compatible length of interval. As the Kendall's τ increases, the empirical coverage rate from both approaches tends to decrease and diverge from the nominal level. But the proposed censoring method is much less affected by the dependence level than the midrank method. In the extreme cases like the $\tau = 0.9$, the coverage rate from midrank method goes down to 0%. Moreover, as the sample size increases, which means the severity of ties becomes stronger, our proposed method is almost not affected by the ties, but the performance of midrank method becomes worse with the percentage of ties increasing.

A.2 Nonparametric bootstrap with the censoring method

Our censoring method makes the nonparametric bootstrap work better in constructing confidence intervals for data with or without ties. We repeated the simulation studies in Table 1 with the nonparametric bootstrap procedure under two conditions: the original pair has no ties and the first margin of original pair has many ties.

Table A.2 reports the results where original data have no ties and the first margin of original pair is rounded to the first digit. For the case of no ties in original pair, the confidence intervals from both the midrank and the censoring methods give coverage close to the nominal level. While for the condition where ties are present in one margin, we see similar observations as reported in paragraph 2, Section 3.2: the censoring method still performs well, while the coverage rates from midrank method are much lower than the nominal level in many cases. From the additional simulation studies, we can actually conclude that the censoring method performs better and makes nonparametric bootstrap work.

ACKNOWLEDGMENTS

J. Yan's research was partially supported by an NSF grant (DMS 1521730). Y. Li's research was partially supported by the National Natural Science Foundation of China (No. 71771211).

Received 30 January 2019

REFERENCES

Bücher, A. and Kojadinovic, I. (2015). An Overview of Nonparametric Tests of Extreme-Value Dependence and of Some Related Statistical Procedures. In *Extreme Value Modeling and Risk Analysis: Methods and Applications* (D. Dey and J. Yan, eds.) 377–398. CRC Press. MR3644322

Chen, D.-G., Sun, J. and Peace, K. E., eds. (2012). *Interval-Censored Time-to-Event Data: Methods and Applications*. CRC Press. MR3053014

Faugeras, O. P. (2017). Inference for Copula Modeling of Discrete Data: A Cautionary Tale and Some Facts. *Dependence Modeling* **5** 121–132. MR3667090

Frees, E. W. and Valdez, E. A. (1998). Understanding Relationships Using Copulas. *North American Actuarial Journal* **2** 1–25. MR1988432

Genest, C., Ghoudi, K. and Rivest, L.-P. (1995). A Semiparametric Estimation Procedure of Dependence Parameters in Multivariate Families of Distributions. *Biometrika* **82** 543–552. MR1366280

Genest, C., Ghoudi, K. and Rémillard, B. (2007). Rank-Based Extensions of the Brock, Dechert, and Scheinkman Test. *Journal of the American Statistical Association* **102** 1363–1376. MR2372539

Genest, C., Neslehova, J. and Ruppert, M. (2011). Discussion: Statistical Models and Methods for Dependence in Insurance Data. *Journal of the Korean Statistical Society* **40** 141–148. MR2830507

Genest, C. and Rémillard, B. (2008). Validity of the Parametric Bootstrap for Goodness-of-Fit Testing in Semiparametric Models. *Annales De L Institut Henri Poincare Probabilites Et Statistiques* **44** 1096–1127. MR2469337

Genest, C., Rémillard, B. and Beaudoin, D. (2009). Goodness-of-Fit Tests for Copulas: A Review and a Power Study. *Insurance: Mathematics and Economics* **44** 199–213. MR2517885

Hofert, M., Kojadinovic, I., Mächler, M. and Yan, J. (2016). `copula`: Multivariate Dependence with Copulas R package version 0.999-15, URL: https://cran.r-project.org/web/packages/copula/.

Hu, M. and Liang, H. (2014). A Copula Approach to Assessing Granger Causality. *Neuroimage* **100** 125–134.

Kim, G., Silvapulle, M. J. and Silvapulle, P. (2007). Comparison of Semiparametric and Parametric Methods for Estimating Copulas. *Computational Statistics and Data Analysis* **51** 2836–2850. MR2345609

Kojadinovic, I. (2017). Some Copula Inference Procedures Adapted to the Presence of Ties. *Computational Statistics & Data Analysis* **112** 24–41. MR3645586

Kojadinovic, I. and Yan, J. (2010). Modeling Multivariate Distributions with Continuous Margins Using the Copula R Package. *Journal of Statistical Software* **34** 1–20.

Mackenzie, D. and Spears, T. (2014). 'The Formula that Killed Wall Street': The Gaussian Copula and Modelling Practices in Investment Banking. *Social Studies of Science* **44** 393–417.

Nikoloulopoulos, A. K. (2013). Copula-Based Models for Multivariate Discrete Response Data. In *Copulae in Mathematical and Quantitative Finance* (P. Jaworski, F. Durante and W. K. Härdle, eds.) 231- 249. MR3288246

Nikoloulopoulos, A. K. and Karlis, D. (2009). Modeling Multivariate Count Data Using Copulas. *Communications in Statistics: Simulation and Computation* **39** 172–187. MR2784560

Pappadà, R., Durante, F. and Salvadori, G. (2016). Quantification of the Environmental Structural Risk with Spoiling Ties: Is Randomization Worthwhile? *Stochastic Environmental Research and Risk Assessment*. Forthcoming.

Parent, E., Favre, A.-C., Bernier, J. and Perreault, L. (2014). Copula Models for Frequency Analysis What Can Be Learned From a Bayesian Perspective? *Advances in Water Resources* **63** 91–103.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Paris: Publications de l'Institut de Statistique de l'Université de Paris* **8** 229–231. MR0125600

Sun, J. (2007). *The Statistical Analysis of Interval-Censored Failure Time Data*. Springer. MR2287318

You, Y. and Li, X. (2014). Optimal Capital Allocations to Interdependent Actuarial Risks. *Insurance Mathematics and Economics* **57** 104–113. MR3225331

Yan Li
Department of Statistics
University of Connecticut
Mansfield, CT
U.S.A.
E-mail address: yan.4.li@uconn.edu

Yang Li
School of Statistics, Center for Applied Statistics
Renmin University of China
Beijing, P.R. China
E-mail address: yang.li@ruc.edu.cn

Yichen Qin
Department of Operations,
Business Analytics and Information Systems
University of Cincinnati
Cincinnati, OH
U.S.A.
E-mail address: qinyn@ucmail.uc.edu

Jun Yan
Department of Statistics
University of Connecticut
Mansfield, CT
U.S.A.
E-mail address: jun.yan@uconn.edu