TEST (2015) 24:734–751
DOI 10.1007/s11749-015-0432-8

ORIGINAL PAPER

An algorithm for estimating survival under a copula-based dependent truncation model

T. Emura · K. Murotani

Received: 9 March 2014 / Accepted: 24 February 2015 / Published online: 10 March 2015

Abstract Traditional analysis with truncated survival data has been developed under the assumption that the lifetime variable of interest is statistically independent of the truncation variable. However, empirical evidence has shown that the truncation variable may depend on the lifetime of interest in many real-world examples. The lack of independence can lead to seriously biased analysis. In this article, we revisit an existing estimation procedure for survival under a copula-based dependent truncation model. Here, the same estimating equation is adopted but a different algorithm to solve the equation is proposed. We compare the new algorithm with the existing one and discuss its theoretical and practical usefulness. Real data examples are analyzed for illustration. We implemented the proposed algorithm in an R "depend.truncation" package, available from CRAN.

Keywords Archimedean copula · Bivariate survival function · Copula-graphic estimator · Kendall's tau · Left truncation · Product-limit estimator

Mathematics Subject Classification 62N01 · 62N02

Electronic supplementary material The online version of this article (doi:10.1007/s11749-015-0432-8) contains supplementary material, which is available to authorized users.

T. Emura (✉)
Graduate Institute of Statistics, National Central University, Jhongli, Taiwan
e-mail: takeshiemura@gmail.com; emura@stat.ncu.edu.tw

K. Murotani
Center for Advanced Medicine and Clinical Research, Nagoya University Hospital, Nagoya, Japan
e-mail: kmurotani@med.nagoya-u.ac.jp

1 Introduction

Consider the situation that a pair of variables (X, Y) can be included in the sample only if $X \leq Y$. The variable X is said to be right truncated by Y or Y is left truncated by X. Left truncation refers to the situation that the samples are available only when a variable of interest Y exceeds a threshold X. An example of such data is the survival data collected from the Channing house retirement center in Palo Alto, California (Hyde 1977, 1980). The data record residents' lifetime subject to the criteria that a resident had to live long enough to enter the center. Thus, the entry age stands for the left-truncation variable. This type of left truncation is also known as 'delayed entry' (Andersen and Keiding 2002), since the residents are not under observation until they enter the study. The data also include right censoring due to residents' withdrawal. The aim of the Channing house study is to draw statistical inference about the survival of the residents in which the late entry bias is removed (Klein and Moeschberger 2003).

Traditionally, most literature on the truncated data considers statistical estimation by assuming that X and Y are quasi-independent (Tsai 1990). In the case of delayed entry (left truncation), the entry is assumed to be independent of the lifetime (Andersen and Keiding 2002). The independence assumption however may not hold in practice. For example in the study of transfusion-related AIDS, the incubation time X is right truncated by the lapse time Y measured from the time of infection (Lagakos et al. 1988). Tsai's test rejects the hypothesis of quasi-independence between X and Y. The association might be attributed to the change of medical practice along the study period, which would shed some light on AIDS research. Many authors propose statistical tests for quasi-independence (Chen et al. 1996; Martin and Betensky 2005; Emura and Wang 2010; Rodriguez-Girondo and de Uña-Álvarez 2012; de Uña-Álvarez 2012; Strazalkowska-Kominiak and Stute 2013). A need for research for dependent left truncation under competing risks setups is pointed out by Bakoyannis and Touloumi (2012).

To assess the degree of dependence between lifetime and truncated variables, Chaieb et al. (2006) propose a semi-parametric estimation for a copula model for describing dependent truncation data. Beaudoin and Lakhal-Chaieb (2008) develop a copula selection in terms of the distance between the nonparametric and model-based dependence measures. Under the same copula models, Emura et al. (2011) and Emura and Wang (2012) consider alternative estimators based on conditional likelihood and nonparametric likelihood, respectively. Ding (2012) verifies the identifiability of the Archimedean copulas used in Chaieb et al. (2006). A recent proposal is a copula-based nonparametric association study of Strazalkowska-Kominiak and Stute (2013). All these analyses for dependent truncation parallel the copula-based analyses for dependent censoring (Zheng and Klein 1995; Rivest and Wells 2001; Escarela and Carriere 2003; Braekers and Veraverbeke 2005; Chen 2010; Emura and Chen 2014).

This article revisits the estimation procedure of Chaieb et al. (2006) and proposes a different algorithm of solving their estimating function. The proposed algorithm is easier to understand than the original one that involves algebraically advanced techniques. We derive a condition that the proposed algorithm leads to an equivalent result as Chaieb et al. (2006). This condition provides some useful consequence in real data analysis. In addition, the proposed algorithm offers a new derivation of the well-known nonparametric estimators under quasi-independence.

The article is organized as follows. Section 2 reviews existing research. Section 3 presents the proposed algorithm. Section 4 compares the proposed algorithm with the existing one. Section 5 modifies the proposed procedure to account for censoring and small risk set. Section 6 analyses two real datasets. Section 7 concludes the article.

2 Preliminary

This section revisits the paper of Chaieb et al. (2006) and reviews key results that will be used for subsequent discussions.

2.1 Copula models for dependent truncation

To assess the degree of association between lifetime and truncation variables, Chaieb et al. (2006) suggested imposing a general class of Archimedean copulas:

$$\pi(x, y) = \phi_\alpha^{-1}[\phi_\alpha\{F_X(x)\} + \phi_\alpha\{S_Y(y)\}]/c, \quad (x \le y), \tag{1}$$

where $\pi(x, y) \equiv \Pr(X \le x, Y > y | X \le Y)$, $F_X(\cdot)$ and $S_Y(\cdot)$ are arbitrary continuous distribution and survival functions, respectively, and c is a normalizing constant. Here, $\phi_\alpha : [0, 1] \to [0, \infty)$ is a copula generator (Nelsen 2006) with an unknown parameter α. If $\phi_\alpha(t) = -\log(t)$, Eq. (1) corresponds to the independence on the upper wedge $\{(x, y) : x \le y\}$, which is equivalent to quasi-independence.

The Clayton copula is defined by $\phi_\alpha(t) = (t^{-(\alpha-1)} - 1)/(\alpha - 1)$, $\alpha \ge 0$, which yields dependence on (X, Y) as measured by Kendall's tau $\tau_\alpha = -(\alpha - 1)/(\alpha + 1)$. The case of $\alpha = 0$ corresponds to the Fréchet–Hoeffding lower bound (Nelsen 2006).

The Frank copula is specified by $\phi_\alpha(t) = \log\{(1-\alpha^{-1})/(1-\alpha^{-t})\}$, $\alpha > 0$, which yields Kendall's tau on (X, Y) in the form

$$\tau_\alpha = -\left[1 + \frac{4}{\gamma}\left\{\frac{1}{\gamma}\int_0^\gamma \frac{x}{e^x - 1}\mathrm{d}x - 1\right\}\right],$$

where $\gamma = -\log\alpha$. More details about the Clayton and Frank copulas are found in the online Supplementary Materials.

Due to the semi-survival structure, Kendall's tau on (X, Y) is the minus of Kendall's tau on the copula (Genest and Mackay 1986). For the Clayton and Frank copulas, $0 < \alpha < 1$ corresponds to positive dependence, $\alpha = 1$ corresponds to quasi-independence, and $\alpha > 1$ corresponds to negative dependence on (X, Y), respectively. Hence, the value α controls the degree of dependence.

The model (1) is adopted in many statistical methods for dependent truncation, including Beaudoin and Lakhal-Chaieb (2008), Emura et al. (2011) and Emura and Wang (2012). Ding (2012) gives a sufficient condition for an Archimedean copula family that makes the model (1) identifiable. Commonly used Archimedean copulas, such as the Clayton and Frank copulas, satisfy his condition.

2.2 Estimating procedure of Chaieb et al. (2006)

Let $\{(X_i, Y_i)(j = 1, \ldots, n)\}$, satisfying $X_j \leq Y_j$, be iid samples from the model (1). We assume that the samples have no ties, i.e., all the $2n$ data points are different. Ordered values of X and Y are denoted as $X_{(1)} < \cdots < X_{(n)}$ and $Y_{(1)} < \cdots < Y_{(n)}$, respectively.

The model (1) facilitates estimation of (α, c, F_X, S_Y) by replacing $\pi(x, y)$ with

$$\hat{\pi}(x, y) \equiv \frac{1}{n}\sum_{j=1}^{n} \mathbf{I}(X_j \leq x, Y_j > y),$$

where $\mathbf{I}(\cdot)$ is the indicator function. Hence, in principle, the unknown quantities (α, c, F_X, S_Y) are estimated on a basis of Eq. (1) with $\pi(x, y)$ being replaced by $\hat{\pi}(x, y)$.

Chaieb et al. (2006) utilize some algebraic techniques to find the estimator of (α, c, F_X, S_Y). Letting $x = y = t$ in Eq. (1), they propose estimating equations:

$$\phi_\alpha\{c\hat{\pi}(t, t-)\} = \phi_\alpha\{F_X(t)\} + \phi_\alpha\{S_Y(t-)\} \tag{2}$$

where t is an observed point for X_j or Y_j. Applying the idea of Rivest and Wells (2001) to Eq. (2), the difference equations are obtained as:

$$\phi_\alpha\{\, S_Y(Y_j)\} - \phi_\alpha\{S_Y(Y_j-)\} = \phi_\alpha\left\{c\frac{\tilde{R}(Y_j) - 1}{n}\right\} - \phi_\alpha\left\{c\frac{\tilde{R}(Y_j)}{n}\right\}, \tag{3a}$$

$$\phi_\alpha\{F_X(X_j-)\} - \phi_\alpha\{F_X(X_j)\} = \phi_\alpha\left\{c\frac{\tilde{R}(X_j) - 1}{n}\right\} - \phi_\alpha\left\{c\frac{\tilde{R}(X_j)}{n}\right\}, \tag{3b}$$

where $\tilde{R}(t) \equiv \sum_{j=1}^{n} \mathbf{I}(X_j \leq t, Y_j \geq t)$. Denote the estimators of $F_X(t)$ and $S_Y(t)$ as $\hat{F}_X(t)$ and $\hat{S}_Y(t)$, respectively, which are step functions with jumps only at observed points. Equations (3a) and (3b) yield the following recursive formulae:

$$\phi_\alpha\left\{\hat{S}_Y(Y_{(j)})\right\} = \phi_\alpha\left\{\hat{S}_Y(Y_{(j-1)})\right\} + \phi_\alpha\left\{c\frac{\tilde{R}(Y_{(j)}) - 1}{n}\right\} - \phi_\alpha\left\{c\frac{\tilde{R}(Y_{(j)})}{n}\right\}, \tag{4a}$$

$$\phi_\alpha\left\{\hat{F}_X(X_{(j)})\right\} = \phi_\alpha\left\{\hat{F}_X(X_{(j+1)})\right\} + \phi_\alpha\left\{c\frac{\tilde{R}(X_{(j)}) - 1}{n}\right\} - \phi_\alpha\left\{c\frac{\tilde{R}(X_{(j)})}{n}\right\}. \tag{4b}$$

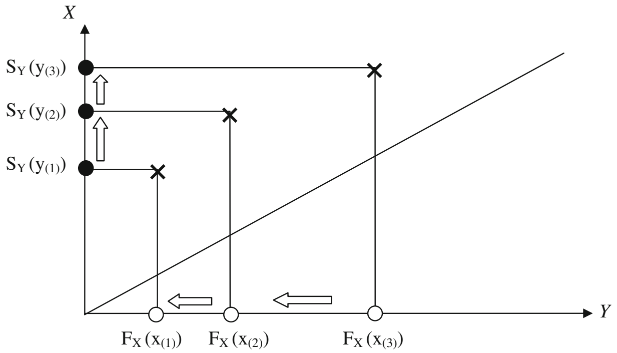

Fig. 1 The algorithm of Chaieb et al. (2006) for solving estimators of $(F_X(\cdot), S_Y(\cdot))$. Here, "×" represents three data points $\{(X_1, Y_1), (X_2, Y_2), (X_3, Y_3)\}$. The algorithm starts from $\hat{S}_Y(Y_{(1)}-) = 1$ and $\hat{F}_X(X_{(3)}) = 1$, and then proceeds as $\hat{S}_Y(Y_{(1)}) \to \hat{S}_Y(Y_{(2)}) \to \hat{S}_Y(Y_{(3)})$, and $\hat{F}_X(X_{(3)}) \to \hat{F}_X(X_{(2)}) \to \hat{F}_X(X_{(1)})$, respectively

Starting with $\hat{S}_Y(Y_{(1)}-) = 1$ and $\hat{F}_X(X_{(n)}) = 1$, the estimators are successively solved as

$$\begin{aligned}1 &= \hat{S}_Y(Y_{(1)}-) \to \hat{S}_Y(Y_{(1)}) \to \hat{S}_Y(Y_{(2)}) \to \cdots \to \hat{S}_Y(Y_{(n)}),\\ 1 &= \hat{F}_X(X_{(n)}) \to \hat{F}_X(X_{(n-1)}) \to \hat{F}_X(X_{(n-2)}) \to \cdots \to \hat{F}_X(X_{(1)}).\end{aligned}$$

Figure 1 depicts the schematic diagram of solving the estimating equations. It is interesting to point out that the direction for solving $\hat{F}_X$ follows a reverse-time scale. The reverse-time representation is standard for constructing the product-limit estimator for right truncated data (Wang et al. 1986; Lagakos et al. 1988). Nevertheless, we raise a question: is it possible to solve for $\hat{F}_X$ in the ordinary direction? The answer is not easy due to $\phi_\alpha(0) = \infty$. In Sect. 3, we will answer this question using our proposed algorithm.

Successively summing together Eqs. (3a) and (3b) up to time t, Chaieb et al. (2006) obtains explicit solutions

$$\phi_\alpha\{\hat{S}_Y(t)\} = -\sum_{j:Y_j \le t}\left[\phi_\alpha\left\{c\frac{\tilde{R}(Y_j)}{n}\right\} - \phi_\alpha\left\{c\frac{\tilde{R}(Y_j)-1}{n}\right\}\right], \tag{5a}$$

$$\phi_\alpha\{\hat{F}_X(t)\} = -\sum_{j:\,X_j > t}\left[\phi_\alpha\left\{c\frac{\tilde{R}(X_j)}{n}\right\} - \phi_\alpha\left\{c\frac{\tilde{R}(X_j)-1}{n}\right\}\right]. \tag{5b}$$

2.3 Estimation of (α, c)

To estimate (α, c), two estimating functions are proposed in Chaieb et al. (2006). They impose additional constraints that, for some $x_0 > y_0 > 0$,

$$F_X(x_0) = 1,\, S_Y(x_0) > 0; \quad F_X(y_0) > 0,\, S_Y(y_0) = 1. \tag{6}$$

By plugging in Eqs. (5a) and (5b) back to Eq. (2) with $t = x_0$, the estimating function becomes

$$U_c(\alpha, c) = \sum_{j:Y_j<x_0} \left[\phi_\alpha \left\{ c\frac{\tilde{R}(Y_j)}{n} \right\} - \phi_\alpha \left\{ c\frac{\tilde{R}(Y_j)-1}{n} \right\} \right] + \phi_\alpha \left\{ c\frac{\tilde{R}(x_0)}{n} \right\} = 0. \tag{7}$$

A consistent and asymptotically normal estimator $(\hat{\alpha}, \hat{c})$ is obtained by solving the estimating Eq. (7) jointly with another estimating function, namely $U_\alpha(\alpha, c) = 0$. The equation $U_\alpha(\alpha, c) = 0$ is obtained by the moment-type equation based on the conditional Kendall's tau (Chaieb et al. 2006) or by the estimating equation based on the conditional likelihood (Emura et al. 2011). While the moment method is simpler to calculate, the conditional likelihood analysis is more efficient by utilizing the distributional information. We will use the former for all the subsequent numerical analyses (Supplementary Material includes numerical results based on the latter).

3 Proposed method

3.1 Proposed algorithm

We propose an new algorithm to solve Eq. (2) for estimating (α, c, F_X, S_Y). We slightly modify Eq. (2) by replacing $\hat{\pi}(t, t-)$ and $S_Y(t-)$ with $\hat{\pi}(t, t)$ and $S_Y(t)$, respectively. This minor change is necessary for the subsequent algorithm to yield proper solutions. We consider modified estimating functions:

$$\phi_\alpha \left\{ c\hat{\pi}(t_j, t_j) \right\} = \phi_\alpha\{F_X(t_j)\} + \phi_\alpha\{S_Y(t_j)\} \quad (j = 1, \ldots, 2n-1), \tag{8}$$

where $t_1 < \cdots < t_{2n-1} < t_{2n}$ are ordered observed points of $(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$. We excluded t_{2n} from Eq. (8) since $\phi_\alpha\{c\hat{\pi}(t_{2n}, t_{2n})\} = \phi_\alpha(0) = \infty$ does not provide a proper estimating equation.

We propose to solve Eq. (8) together with two boundary constraints:

$$\hat{F}_X(t_{2n-1}) = 1 \quad \text{and} \quad \hat{S}_Y(t_1) = 1.$$

For fixed values for (α, c), Eq. (8) can be regarded as an estimating function for $\{F_X(t_j), S_Y(t_j)\}$. In the initial step, notice that $\hat{\pi}(t_1, t_1) = 1/n$. Since nobody has yet died at t_1, we know $\hat{S}_Y(t_1) = 1$. Hence, Eq. (8) gives $\hat{F}_X(t_1) = c/n$, yielding the initial solution $\{\hat{F}_X(t_1), \hat{S}_Y(t_1)\} = (c/n, 1)$. Subsequent calculations for $j = 2, \ldots, 2n-1$ are similarly performed using the solutions from the previous step. The key is that the unknown quantity in Eq. (8) becomes either $F_X(t_j)$ or $S_Y(t_j)$. For instance, if t_j corresponds to an observed value of X, then $S_Y(t_j)$ is known to be $\hat{S}_Y(t_{j-1})$ and $F_X(t_j)$ is unknown. The last solution $\hat{F}_X(t_{2n-1})$ depends on (α, c) and hence does not necessarily equal to 1. We propose to obtain $(\hat{\alpha}, \hat{c})$ that meets $\hat{F}_X(t_{2n-1}) = 1$.

In summary, we propose the following procedure for $j = 2, \dots, 2n-1$.

(Step 0) Set the initial solution $\{\hat{F}_X(t_1), \hat{S}_Y(t_1)\} = (c/n, 1)$.
(Step 1) If t_j corresponds to an observed value of X, set

$$\hat{S}_Y(t_j) = \hat{S}_Y(t_{j-1}) \quad \text{and} \quad \phi_\alpha\{\hat{F}_X(t_j)\} = \phi_\alpha\{c\hat{\pi}(t_j, t_j)\} - \phi_\alpha\{\hat{S}_Y(t_{j-1})\};$$

and if t_j corresponds to an observed value of Y, set

$$\hat{F}_X(t_j) = \hat{F}_X(t_{j-1}) \quad \text{and} \quad \phi_\alpha\{\hat{S}_Y(t_j)\} = \phi_\alpha\{c\hat{\pi}(t_j, t_j)\} - \phi_\alpha\{\hat{F}_X(t_{j-1})\}.$$

(Step 2) Set $U_c(\alpha, c) = \phi_\alpha\{\hat{F}_X(t_{2n-1})\} = 0$ to meet $\hat{F}_X(t_{2n-1}) = 1$. By jointly solving $U_c(\alpha, c) = 0$ and $U_\alpha(\alpha, c) = 0$, the estimators $(\hat{\alpha}, \hat{c})$ can be obtained.
(Step 3) Redo (Step 1) by setting $(\alpha, c) = (\hat{\alpha}, \hat{c})$ and then obtain $\{\hat{F}_X(t_j), \hat{S}_Y(t_j)\}$.

Explicit expressions of (Step 1) can be derived as

$$\phi_\alpha\{\hat{S}_Y(t)\} = -\sum_{j:Y_j \le t} \left[\phi_\alpha\left\{ c\frac{\tilde{R}(Y_j)}{n} \right\} - \phi_\alpha\left\{ c\frac{\tilde{R}(Y_j)-1}{n} \right\} \right], \tag{9a}$$

$$\phi_\alpha\{\hat{F}_X(t)\} = \sum_{j:\, t_1 < X_j \le t} \left[\phi_\alpha\left\{ c\frac{\tilde{R}(X_j)}{n} \right\} - \phi_\alpha\left\{ c\frac{\tilde{R}(X_j)-1}{n} \right\} \right] + \phi_\alpha\left(\frac{c}{n}\right). \tag{9b}$$

Similarly, the equation defined in (Step 2) can be written as

$$U_c(\alpha, c) \equiv \phi_\alpha\left\{\hat{F}_X(t_{2n-1})\right\} = \sum_{j:\, t_1 < X_j} \left[\phi_\alpha\left\{ c\frac{\tilde{R}(X_j)}{n} \right\} - \phi_\alpha\left\{ c\frac{\tilde{R}(X_j)-1}{n} \right\} \right] + \phi_\alpha\left(\frac{c}{n}\right). \tag{10}$$

In the case of quasi-independence with $\phi_\alpha(t) = -\log(t)$, Eq. (9a) reduces to the product-limit estimator (Lynden-Bell 1971; Wang et al. 1986)

$$\hat{S}_Y(t) = \prod_{j;Y_j \le t} \{1 - 1/\tilde{R}(Y_j)\},$$

and Eq. (10) yields the estimator of He and Yang (1998)

$$\hat{c} = n \prod_{j;t_1 < X_j} \{1 - 1/\tilde{R}(X_j)\}.$$

With $c = \hat{c}$, Eq. (9b) produces the product-limit estimator

$$\hat{F}_X(t) = \prod_{j;t < X_j} \{1 - 1/\tilde{R}(X_j)\}.$$

Hence, these well-known nonparametric estimators under quasi-independence are derived from the algorithm of (Step 0)–(Step 3) that solve $\hat{\pi}(t_i, t_i) = F_X(t_i)S_Y(t_i)/c$ with $\hat{F}_X(t_{2n-1}) = 1$ and $\hat{S}_Y(t_1) = 1$. This finding gives us an alternative derivation of the well-known estimators, which appears to be new in the literature.

3.2 Example

We demonstrate the proposed algorithm using small data $(X_1, Y_1) = (1, 3)$, $(X_2, Y_2) = (2, 5)$ and $(X_3, Y_3) = (4, 6)$, which are plotted in Fig. 2.

The initial solution is $\{\hat{F}_X(t_1), \hat{S}_Y(t_1)\} = (c/3, 1)$ for $t_1 = X_1 = 1$ (Step 0). The subsequent calculations for $(\phi_\alpha\{\hat{F}_X(t_j)\}, \phi_\alpha\{\hat{S}_Y(t_j)\})$, $j = 2, \ldots, 5$, are given in Table 1 (Step 1). By setting $\phi_\alpha\{\hat{F}_X(t_5)\} = 0$, we need to solve $U_c(\alpha, c) = 2\phi_\alpha(2c/3) - \phi_\alpha(c/3) = 0$ (Step 2). If one fits the Clayton copula with $\phi_\alpha(t) = (t^{-(\alpha-1)} - 1)/(\alpha - 1)$, it is not difficult to show that the solution to $U_\alpha(\alpha, c) = 0$ and $U_c(\alpha, c) = 0$ becomes $\hat{\alpha} = 0$ (Kendall's tau $= 1$) and $\hat{c} = 1$, respectively. The resultant $\phi_{\alpha=0}(t) = 1 - t$ is the Fréchet–Hoeffding lower bound (Nelsen 2006). Accordingly, (Step 3) leads to the solutions

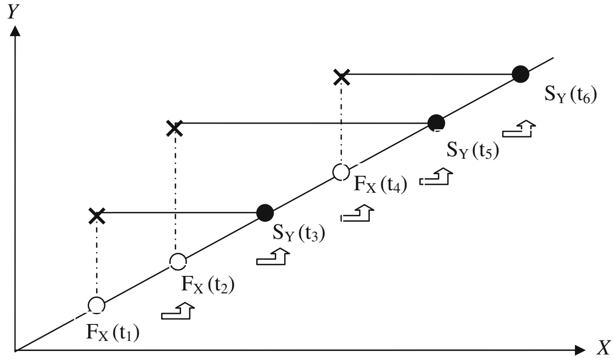

Fig. 2 The proposed algorithm for estimating $(F_X(\cdot), S_Y(\cdot))$. "×" represents the data points $(X_1, Y_1) = (1, 3)$, $(X_2, Y_2) = (2, 5)$ and $(X_3, Y_3) = (4, 6)$

Table 1 Results of performing (Step 1) of the proposed algorithm for a small dataset: $(X_1, Y_1) = (1, 3)$, $(X_2, Y_2) = (2, 5)$ and $(X_3, Y_3) = (4, 6)$

	$\hat{\pi}(t_j, t_j)$	$\phi_\alpha\{c\hat{\pi}(t_j, t_j)\}$	$\phi_\alpha\{\hat{F}_X(t_j)\}$	$\phi_\alpha\{\hat{S}_Y(t_j)\}$
$t_1 = X_1 = 1$	$\frac{1}{3}$	$\phi_\alpha(\frac{c}{3})$	$\phi_\alpha(\frac{c}{3})$	0
$t_2 = X_2 = 2$	$\frac{2}{3}$	$\phi_\alpha(\frac{2c}{3})$	$\phi_\alpha(\frac{2c}{3})$	0
$t_3 = Y_1 = 3$	$\frac{1}{3}$	$\phi_\alpha(\frac{c}{3})$	$\phi_\alpha(\frac{2c}{3})$	$\phi_\alpha(\frac{c}{3}) - \phi_\alpha(\frac{2c}{3})$
$t_4 = X_3 = 4$	$\frac{2}{3}$	$\phi_\alpha(\frac{2c}{3})$	$2\phi_\alpha(\frac{2c}{3}) - \phi_\alpha(\frac{c}{3})$	$\phi_\alpha(\frac{c}{3}) - \phi_\alpha(\frac{2c}{3})$
$t_5 = Y_2 = 5$	$\frac{1}{3}$	$\phi_\alpha(\frac{c}{3})$	$2\phi_\alpha(\frac{2c}{3}) - \phi_\alpha(\frac{c}{3})$	$2\phi_\alpha(\frac{c}{3}) - 2\phi_\alpha(\frac{2c}{3})$
$t_6 = Y_3 = 6$	$\frac{0}{3}$	$\phi_\alpha(0)$	Undetermined	Undetermined

$$\hat{F}_X(t_1) = \hat{F}_X(X_{(1)}) = 1/3, \quad \hat{F}_X(t_2) = \hat{F}_X(X_{(2)}) = 2/3, \quad \hat{F}_X(t_4) = \hat{F}_X(X_{(3)}) = 1,$$
$$\hat{S}_Y(t_3) = \hat{S}_Y(Y_{(1)}) = 2/3, \quad \hat{S}_Y(t_5) = \hat{S}_Y(Y_{(2)}) = 1/3, \hat{S}_Y(t_6) = \hat{S}_Y(Y_{(3)}) = \text{undefined}.$$

Instead of the Clayton copula, we now fit the independence copula with $\phi_\alpha(t) = -\log(t)$. Then, performing (Step 0)–(Step 3), one obtains the solutions

$$\hat{F}_X(t_1) = \hat{F}_X(X_{(1)}) = 1/4, \quad \hat{F}_X(t_2) = \hat{F}_X(X_{(2)}) = 1/2, \quad \hat{F}_X(t_4) = \hat{F}_X(X_{(3)}) = 1,$$
$$\hat{S}_Y(t_3) = \hat{S}_Y(Y_{(1)}) = 1/2, \quad \hat{S}_Y(t_5) = \hat{S}_Y(Y_{(2)}) = 1/4, \hat{S}_Y(t_6) = \hat{S}_Y(Y_{(3)}) = \text{undefined},$$

which are equivalent to the product-limit estimator.

The Clayton copula-based estimator and the product-limit estimator give quite different results. Figure 2 shows that all comparable pairs, (1, 2) and (2, 3), are concordant, which indicates positive dependence. Hence, fitting the present Clayton copula with $\hat{\alpha} = 0$ would be preferable to the independence model.

4 Comparison between two algorithms

Qualitative difference between the two algorithms can be appreciated by comparing Fig. 1 (algorithm of Chaieb et al.) and Fig. 2 (the proposed algorithm). The proposed algorithm constitutes a single sequence $\{\hat{F}_X(t_j), \hat{S}_Y(t_j)\}$ for $j = 1, \ldots, 2n-1$ while the algorithm of Chaieb et al. (2006) runs two separate sequences for $\hat{F}_X$ and $\hat{S}_Y$ (compare Figs. 1, 2). The algorithm of Chaieb et al. (2006) for $\hat{F}_X$ follows the reverse-time scale (Fig. 1) while the proposed algorithm follows an ordinary time scale (Fig. 2). For the proposed algorithm, all ordered data points are projected onto the diagonal line, and then the algorithm runs on the line (Fig. 2). The systems of equations are successively solved along a single sequence:

$$F_X(t_1) \to F_X(t_2) \to S_Y(t_3) \to F_X(t_4) \to S_Y(t_5) \to S_Y(t_6).$$

The above mentioned difference of the two algorithms also yields different formulas for $\hat{F}_X(t)$. In particular, the proposed formula involves the summation for all subjects j with $X_j \le t$ (Eq. 9b) while the formula of Chaieb et al. involves the summation for all subject j with $X_j > t$ (Eq. 5b). Hence, the two formulas appear to use quite different information. Somewhat surprisingly, we will see that these apparently different formulas are identical under some conditions.

For marginal estimation in competing risks data, Zheng and Klein (1995) suggest running the algorithm on the ordered data points. Alternatively, Rivest and Wells (2001) propose to run two separate algorithms. Hence, our proposal is similar to Zheng and Klein (1995) while the algorithm of Chaieb et al. (2006) is similar to Rivest and Wells (2001).

Although the two algorithms give two different principles for solving equations, they are shown to be equivalent under some condition.

Theorem 1 *The proposed estimating equation* $\phi_\alpha\{\hat{F}_X(t_{2n-1})\} = 0$ *is equivalent to the estimating equation* (7) *of* Chaieb et al. (2006) *under* $x_0 \in [X_{(n)}, t_{2n-1}]$.

Proof We will show that $\phi_\alpha\{\hat{F}_X(t_{2n-1})\} = 0$ is identical to Eq. (7). Note that $\phi_\alpha\{\hat{F}_X(t_{2n-1})\} = \phi_\alpha\{\hat{F}_X(x_0)\}$ since there is no jump for X beyond $x_0 \in [X_{(n)}, t_{2n-1}]$. It follows that

$$\begin{aligned}\phi_\alpha\left\{\hat{F}_X(x_0)\right\} &= \phi_\alpha\{c\hat{\pi}(x_0, x_0)\} - \phi_\alpha\left\{\hat{S}_Y(x_0)\right\}\\ &= \phi_\alpha\{c\hat{\pi}(x_0, x_0)\} + \sum_{j:Y_j \le x_0}\left[\phi_\alpha\left\{c\frac{\tilde{R}(Y_j)}{n}\right\} - \phi_\alpha\left\{c\frac{\tilde{R}(Y_j)-1}{n}\right\}\right]\\ &= \phi_\alpha\left\{c\frac{\tilde{R}(x_0)}{n}\right\} + \sum_{j:Y_j < x_0}\left[\phi_\alpha\left\{c\frac{\tilde{R}(Y_j)}{n}\right\} - \phi_\alpha\left\{c\frac{\tilde{R}(Y_j)-1}{n}\right\}\right].\end{aligned}$$

Setting the last equation to zero is equivalent to solving Eq. (7). □

Theorem 1 gives a guideline of choosing x_0 in the estimating equation (7) of Chaieb et al. (2006). Practitioners need to choose the value x_0 defined in Eq. (6) on a basis of observed data. Theorem 1 guarantees that any $x_0 \in [X_{(n)}, t_{2n-1}]$ produces the same estimating equation (yet different formulas for different values of x_0). A convenient choice is $x_0 = t_{2n-1}$, which coincides with the proposed estimating Eq. (10).

Theorem 2 *Let* $x_0 \in [X_{(n)}, t_{2n-1}]$ *in Eq.* (7)*. If* (α, c) *satisfies* $U_c(\alpha, c) = 0$*, the algorithm of* Chaieb et al. (2006) *and the proposed algorithm yield the same estimator for* $\hat{F}_X$*. That is, Eqs.* (5b) *and* (9b) *are identical.*

Proof Suppose $U_c(\alpha, c) = 0$. Then, Eq. (5b) is written as

$$\begin{aligned}&-\sum_{j:\,X_j>t}\left[\phi_\alpha\left\{c\frac{\tilde{R}(X_j)}{n}\right\} - \phi_\alpha\left\{c\frac{\tilde{R}(X_j)-1}{n}\right\}\right]\\ &= U_c(\alpha, c) - \sum_{j:\,X_j>t}\left[\phi_\alpha\left\{c\frac{\tilde{R}(X_j)}{n}\right\} - \phi_\alpha\left\{c\frac{\tilde{R}(X_j)-1}{n}\right\}\right]\\ &= \sum_{j:\,t_1<X_j\le t}\left[\phi_\alpha\left\{c\frac{\tilde{R}(X_j)}{n}\right\} - \phi_\alpha\left\{c\frac{\tilde{R}(X_j)-1}{n}\right\}\right] + \phi_\alpha\left(\frac{c}{n}\right).\end{aligned}$$

The last equation is equivalent to Eq. (9b). □

Remark The condition $U_c(\alpha, c) = 0$ is necessary for Eqs. (5b) and (9b) to be identical. If (α, c) does not satisfy $U_c(\alpha, c) = 0$, they are not equal in general. As seen from (Step 2) of the proposed algorithm, $U_c(\alpha, c) = 0$ is equivalent to the constraint $\hat{F}_X(t_{2n-1}) = 1$. This constraint makes the proposed expression of Eq. (9b) to be identical to the reverse-time expression of Eq. (5b).

5 Extension to left-truncated and right-censored data

So far, we have considered the case that Y is left truncated by X. The proposed algorithm described in Sect. 3.1 can be extended to the case where Y is also right-censored by a censoring variable C. Assume that C is independent of (X, Y). The sample can be written as $\{(X_i, Z_i, \delta_i)(i = 1, \ldots, n)\}$ satisfying $X_i \le Z_i$, where $Z_i = Y_i \wedge C_i, \delta_i = \mathbf{I}(Y_i \le C_i)$ and $(X_i, Y_i, C_i)(i = 1, \ldots, n)$ are random replications of (X, Y, C) given $X \le Z \equiv Y \wedge C$.

Chaieb et al. (2006) expressed the model as

$$\pi^*(x, y) = \Pr(X \le x, Z > y | X \le Z) = S_C(y)\phi_\alpha^{-1}[\phi_\alpha\{F_X(x)\} + \phi_\alpha\{S_Y(y)\}]/c^*,$$

where $S_C(y) = \Pr(C > y)$, $x \le y$ and c^* is a normalizing constant. The objective is to estimate the unknown parameters $(\alpha, c^* F_X, S_Y, S_C)$. Let $t_1 < \cdots < t_{2n-1} < t_{2n}$ be ordered observed points of $(X_1, \ldots, X_n, Z_1, \ldots, Z_n)$, and let

$$\hat{\pi}^*(t, t) \equiv \frac{1}{n}\sum_{j=1}^{n} \mathbf{I}(X_j \le t, Z_j > t).$$

The estimating functions become

$$\phi_\alpha\left\{c^* \frac{\hat{\pi}^*(t_j, t_j)}{S_C(t_j)}\right\} = \phi_\alpha\{F_X(t_j)\} + \phi_\alpha\{S_Y(t_j)\}, \quad (j = 1, \ldots, 2n-1). \tag{11}$$

To solve the above equations, we impose additional constraints that the estimators of F_X, S_Y and S_C are step functions with jumps only at their observed values, and that

$$\hat{F}_X(t_{2n-1}) = 1, \quad \hat{S}_Y(t_1) = 1 \quad \text{and} \quad \hat{S}_C(t_1) = 1. \tag{12}$$

As in Sect. 3.1, notice that $\hat{\pi}^*(t_1, t_1) = 1/n$. Hence, Eqs. (11) and (12) give the initial solution $\{\hat{S}_Y(t_1), \hat{F}_X(t_1), \hat{S}_C(t_1)\} = (1, c^*/n, 1)$. Subsequent solutions successively solve Eq. (11) for $j = 2, \ldots, 2n-1$.

(Step 0) Set the initial solution $\{\hat{S}_Y(t_1), \hat{F}_X(t_1), \hat{S}_C(t_1)\} = (1, c^*/n, 1)$.

(Step 1) If t_j corresponds to an observed value of X_i, set $\hat{S}_Y(t_j) = \hat{S}_Y(t_{j-1})$, $\hat{S}_C(t_j) = \hat{S}_C(t_{j-1})$ and $\phi_\alpha\{\hat{F}_X(t_j)\} = \phi_\alpha\left\{c^* \frac{\hat{\pi}^*(t_j, t_j)}{\hat{S}_C(t_{j-1})}\right\} - \phi_\alpha\{\hat{S}_Y(t_{j-1})\}$;
if t_j corresponds to an observed value of Z_i with $\delta_i = 1$, set $\hat{F}_X(t_j) = \hat{F}_X(t_{j-1})$, $\hat{S}_C(t_j) = \hat{S}_C(t_{j-1})$, and $\phi_\alpha\{\hat{S}_Y(t_j)\} = \phi_\alpha\left\{c^* \frac{\hat{\pi}^*(t_j, t_j)}{\hat{S}_C(t_{j-1})}\right\} - \phi_\alpha\{\hat{F}_X(t_{j-1})\}$;
if t_j corresponds to an observed value of Z_i with $\delta_i = 0$, set $\hat{F}_X(t_j) = \hat{F}_X(t_{j-1})$, $\hat{S}_Y(t_j) = \hat{S}_Y(t_{j-1})$, and $\hat{S}_C(t_j) = \hat{S}_C(t_{j-1})\hat{\pi}^*(t_j, t_j)/\hat{\pi}^*(t_{j-1}, t_{j-1})$.

(Step 2) Set $U_c(\alpha, c^*) = \phi_\alpha\{\hat{F}_X(t_{2n-1})\} = 0$ to meet the constraint $\hat{F}_X(t_{2n-1}) = 1$. Jointly solving this and $U_\alpha(\alpha, c^*) = 0$ produces the estimators $(\hat{\alpha}, \hat{c}^*)$, where $U_\alpha(\alpha, c^*) = 0$ is available in Chaieb et al. (2006) or Emura et al. (2011).

(Step 3) Redo (Step 1) by setting $(\alpha, c^*) = (\hat{\alpha}, \hat{c}^*)$ obtained in (Step 2) and then obtain $\{\hat{F}_X(t_j), \hat{S}_Y(t_j), \hat{S}_C(t_j)\}$.

Explicit formulae for (Step 1) of the above algorithms are given by

$$\phi_\alpha\left\{\hat{S}_Y(t)\right\} = -\sum_{j;Z_j\le t,\delta_j=1}\left[\phi_\alpha\left\{c^*\frac{\tilde{R}(Z_j)}{n\hat{S}_C(Z_j)}\right\} - \phi_\alpha\left\{c^*\frac{\tilde{R}(Z_j)-1}{n\hat{S}_C(Z_j)}\right\}\right], \tag{13a}$$

$$\phi_\alpha\left\{\hat{F}_X(t)\right\} = \sum_{j;\,t_1<X_j\le t}\left[\phi_\alpha\left\{c^*\frac{\tilde{R}(X_j)}{n\hat{S}_C(X_j)}\right\} - \phi_\alpha\left\{c^*\frac{\tilde{R}(X_j)-1}{n\hat{S}_C(X_j)}\right\}\right] + \phi_\alpha\left(\frac{c^*}{n}\right), \tag{13b}$$

$$\hat{S}_C(y) = \prod_{j:Z_j\le y,\,\delta_j=0}\{1-1/\tilde{R}(Z_j)\}. \tag{13c}$$

The estimating function in (Step 2) is equivalent to

$$U_c(\alpha, c^*) = \sum_{j;\,t_1<X_j}\left[\phi_\alpha\left\{c^*\frac{\tilde{R}(X_j)}{n\hat{S}_C(X_j)}\right\} - \phi_\alpha\left\{c^*\frac{\tilde{R}(X_j)-1}{n\hat{S}_C(X_j)}\right\}\right] + \phi_\alpha\left(\frac{c^*}{n}\right). \tag{14}$$

For $\phi_\alpha(t) = -\log(t)$, Eq. (13a) reduces to the product-limit estimator

$$\hat{S}_Y(t) = \prod_{j;Z_j\le t,\delta_j=1}\{1-1/\tilde{R}(Z_j)\},$$

and Eq. (14) reduces to the estimator of He and Yang (1998).

Small values of $\tilde{R}$ in Eqs. (13a–13c) and (14) often produce unreasonable estimates. This problem of small risk sets even occurs in the product-limit estimator under quasi-independence (Klein and Moeschberger 2003). We follow Lai and Ying (1991) and Emura et al. (2011) who suggest simply discarding the calculations corresponding to very small $\tilde{R}$. In particular, we suggest calculating the terms in the summation of Eqs. (13a–13c) and (14) only when $\tilde{R} \ge bn^a$ holds, where $0 < a < 1$ and $b > 0$ are arbitrary tuning parameters. In the following data analysis, we will use $b = 1$ and $a = 1/4$ as considered in Lai and Ying (1991), or $b = 1$ and $a = 1/10$ which produces less biased results (Emura et al. 2011).

6 Data analysis

6.1 Channing house data

We analyze the survival data for elderly residents in the Channing house as introduced in Sect. 1. The data are available in Hyde (1977), where 97 men are included in the

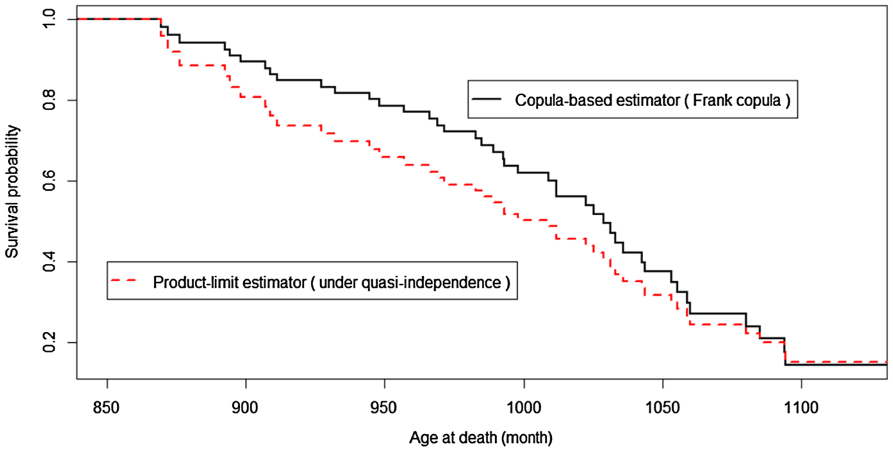

Fig. 3 The copula-based estimator and the product-limit estimator of the survival function $S_Y(y)$ for lifetime of elderly residents based on the Channing house data (Hyde 1977)

sample. After entry to the Channing house, 46 men die and the remaining 51 men are censored due to the withdrawal from the Channing house. Hence, the data is left truncated by the entry age and right censored by the withdrawal. The data are a well-known example of left-truncated and right-censored survival data (Klein and Moeschberger 2003).

Beaudoin and Lakhal-Chaieb (2008) and Emura and Wang (2010) analyzed the same data by taking into account for dependent truncation. Both papers selected the Frank copula as the best-fitting copula among candidates following their own model selection criteria. In addition, the Frank copula satisfies the identifiability condition given by Theorem 2 of Ding (2012). Thus, we also choose the Frank copula.

We break the ties by adding uniform random variables on $[-0.4, 0.4]$ to the original data, which do not change the original ordering (Emura et al. 2011). We set $a = 1/4$ to prevent the problem of small risk set occurring at a few early deaths, and set $x_0 = t_{2n-1}$ to meet the condition of Theorem 1.

As discussed in Sect. 5, the estimates of (α, c^*) are obtained by jointly solving $U_c(\alpha, c^*) = 0$ in Eq. (14) and $U_\alpha(\alpha, c^*) = 0$ in Chaieb et al. (2006).[1] The association parameter estimate is $\hat{\alpha} = 0.083$ (se $= 0.11$) under the Frank copula. The corresponding Kendall's tau $\hat{\tau}_\alpha = 0.26$ (se $= 0.12$) gives a positive association between the entry age and age at death. This conclusion agrees with the previously reported results in Beaudoin and Lakhal-Chaieb (2008) and Emura and Wang (2010).

We estimate the resident's survival function using the proposed algorithm of Sect. 5. Figure 3 depicts the estimated survival function. Clearly, the copula-based estimates of survival probability are higher than the product-limit estimates over the study period.

[1] One can replace the estimating equation of Chaieb et al. (2006) by the estimating equation of Emura et al. (2011). We refer the detailed results under the estimating equation of Emura et al. (2011) to the Supplemental Materials. Although we found some numerical difference between the two approaches of Chaieb et al. (2006) and Emura et al. (2011), the substantive conclusions on the resident's lifetime distribution are similar. Please refer to the Supplemental Materials for the detailed comparison.

For instance, the survival probability at $t = 970$ months is 72.2 % by the copula-based estimator while it is 60.9 % by the product-limit estimator. Hence, the product-limit estimator seriously underestimates the survival probability of residents in the Channing house. In summary, the dependent truncation model gives more survival benefit for the Channing house than the independent model does.

6.2 Japanese centenarian data

We analyze the survival data for centenarians (those who live beyond the age of 100 years) in Japan. The data is prepared from the National Oldest-old Survivors List and Population Movement Statistics by the Ministry of Health and Labor in Japan, as previously reported by Sibuya and Hanayama (2004) and Murotani et al. (2014). The objective is to estimate the lifetime distribution of centenarians. For purpose of illustration, we restrict our samples of $n = 662$ Japanese male centenarians ascertained before year 1980, as summarized in Table 2. Since those centenarians who still survived in 1980 are not counted in the table, the lifetime is right truncated. Specifically, each subject has age at death (X), constrained by $X \leq Y \equiv 1980.5 - T$, where T is the birth year. This scenario is similar to the right truncation of AIDS transfusion data (p. 19 of Klein and Moeschberger 2003).

We choose $a = 1/10$ to avoid the problem of small risk set, and set $x_0 = t_{2n-1}$ to meet the condition of Theorem 1. We break the ties by adding small random noises, which shows little change in the result.

As discussed in Sect. 3.1, the estimates of (α, c) are obtained by jointly solving $U_c(\alpha, c) = 0$ in Eq. (7) and the moment-type equation $U_\alpha(\alpha, c) = 0$ of Chaieb et al. (2006).[2] Then, we estimate the cumulative distribution function F_X of male centenarians using the proposed algorithm.

Figure 4 depicts the estimated cumulative distribution functions. The cumulative distribution function under the Frank copula is nearly identical to the product-limit estimate under quasi-independence. The reason is that the association parameter estimate $\hat{\alpha} = 0.604$ (se $= 0.186$) yields the Kendall's tau nearly equal to zero ($\hat{\tau}_\alpha = 0.056$, se $= 0.035$).

The results under the Clayton copula show little difference from those under the Frank copula in Fig. 4. The association parameter estimate is $\hat{\alpha} = 0.921$ (se $= 0.047$) and the corresponding Kendall's tau is $\hat{\tau}_\alpha = 0.041$ (se $= 0.026$). Though we did not find significant amount of dependence between truncation and lifetime variables, the results of fitting copulas can confirm the traditional analysis by taking into account the potential dependence.

7 Conclusion and discussion

This article revisits the estimating equation proposed by Chaieb et al. (2006), and then proposes a different algorithm to solve it. Unlike the reverse-time representation of

[2] One can replace the estimating equation of Chaieb et al. (2006) by the estimating equation of Emura et al. (2011). Although the two estimating equations are different, there is virtually no numerical difference between the two estimates. This phenomenon occurs in the absence of censoring (Emura et al. 2011).

Table 2 The number of deaths at each year (1963–1980) for Japanese male centenarians ($n = 662$)

	Year at death (1963–1980)																	
	63–64	64–65	65–66	66–67	67–68	68–69	69–70	70–71	71–72	72–73	73–74	74–75	75–76	76–77	77–78	78–79	79–80	80–81
100.5	1	12	9	12	12	12	11	11	16	14	28	21	21	26	25	23	49	23
101.5	7	0	3	3	4	6	10	5	5	7	11	10	11	9	15	10	19	20
102.5	0	4	0	2	5	0	4	5	5	2	7	5	9	6	8	12	9	10
103.5	0	3	0	0	0	4	2	0	4	3	2	0	3	3	2	3	3	8
104.5	0	0	0	1	1	0	2	2	3	1	0	0	0	3	6	1	2	0
105.5	0	0	0	2	0	0	1	0	0	1	3	1	1	0	1	2	1	0
106.5	0	0	0	0	0	0	0	0	0	3	0	0	1	0	1	1	0	0
107.5	0	0	0	0	0	0	0	0	1	0	0	0	2	0	0	0	2	0
108.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
109.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
110.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
111.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
112.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
113.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
114.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
115.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
116.5	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

The number of deaths is categorized by the age at death (100.5–116.5)

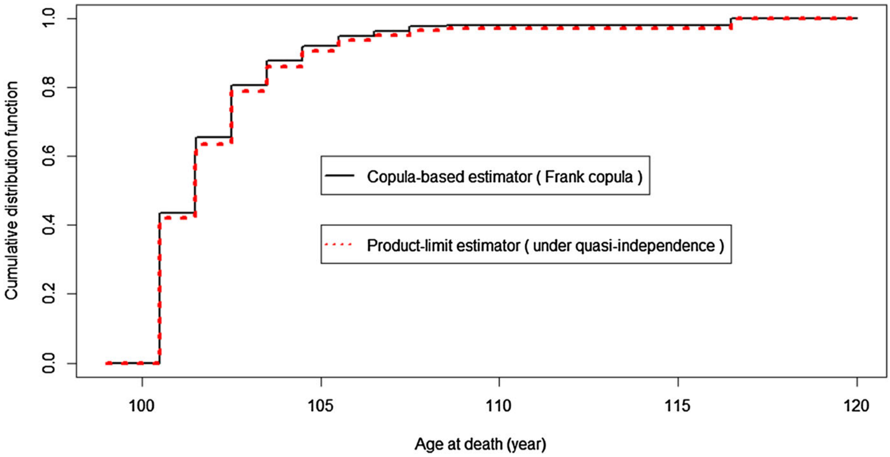

Fig. 4 The copula-based estimator and the product-limit estimator of the cumulative distribution function $F_X(x)$ for lifetime of Japanese male centenarians based on the National Oldest-old Survivors List and Population Movement Statistics by the Ministry of Health and Labor in Japan (Sibuya and Hanayama 2004; Murotani et al. 2014)

Chaieb et al. (2006), the proposed method follows an ordinary time scale, which gives easier understanding and clearer mathematical presentation. In Theorem 1, we give sufficient conditions that the proposed algorithm becomes equivalent to the algorithm of Chaieb et al. (2006). Theorem 1 also offers some practical guideline for the initial conditions (namely, $x_0 = t_{2n-1}$) under which the two algorithms are identical. We implemented automatic routines in R "depend.truncation" package (Emura 2014), available from CRAN. All the given numerical results in the article are reproduced by the package.

This article systematically discusses the qualitative difference between the algorithm of Chaieb et al. (2006) and the proposed algorithm (see Sect. 4). Especially, we highlight the remarkable difference graphically (Figs. 1 vs. 2) and mathematically (Eqs. 5b vs. 9b). Nevertheless, we show that these apparently different algorithms can yield identical results under the conditions of Theorem 1.

Copulas are increasingly popular tools for dependence modeling that is fundamental in bivariate survival analysis. While a firm mathematical understanding for the dependent censoring is obtained (e.g., Zheng and Klein 1995; Rivest and Wells 2001), the models for dependent truncation are relatively new and technically more challenging.

This situation is similar in the presence of covariates. Regression analyses under copula-based dependent censoring models have been well developed (e.g., Escarela and Carriere 2003; Braekers and Veraverbeke 2005; Chen 2010; Emura and Chen 2014). However, to the best of our knowledge, only Ding (2012) discusses the identifiability of the covariate models under a copula-based dependent truncation model. A great deal of work will be necessary to the regression models before they can be used for statistical inference.

The extension might be straightforward with a parametric approach. One possibility is to develop marginal Weibull regression under copula-based dependent truncation models and to perform standard likelihood inference. This parallels the work of Escarela and Carriere (2003) under competing risks models. One technical challenge under dependent truncation models is the complexity of the truncation probability, which would appear in the likelihood function. We refer Emura and Konno (2012) that points out this problem.

One would consider semi-parametric marginal regression approaches to copula-based dependence models along the line of Braekers and Veraverbeke (2005), Chen (2010) and Emura and Chen (2014), all studied under the competing risks setting. A similar approach is to introduce the proportional hazard structures for marginal distributions and perform the nonparametric maximum likelihood estimation under a copula-based dependent truncation model as in Emura and Wang (2012). These approaches would face the computational challenge due to the joint estimation of the two infinite dimensional marginal functions.

Acknowledgments We are grateful to Professor Weijing Wang for raising the main idea of the method. We also thank the associate editor and the reviewer for their careful reading of our manuscript and their helpful comments that substantially improve the manuscript. This work is supported by the research grant funded by the National Science Council of Taiwan (NSC 101-2118-M-008-002-MY2), the Ministry of Science and Technology of Taiwan (MOST 103-2118-M-008-MY2) and MEXT KAKENHI, Grant-in-Aid for Scientific Research (C) 13203524 of Japan.

References

Andersen PK, Keiding N (2002) Multi-state models for event history analysis. Stat Methods Med Res 11:91–115

Bakoyannis G, Touloumi G (2012) Practical methods for competing risks data: a review. Stat Method Med Res 21:257–272

Beaudoin D, Lakhal-Chaieb L (2008) Archimedean copula model selection under dependent truncation. Stat Med 27:4440–4454

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

Chaieb LL, Rivest LP, Abdous B (2006) Estimating survival under a dependent truncation. Biometrika 93:655–69

Chen CH, Tsai WY, Chao WH (1996) The product-moment correlation coefficient and linear regression for truncated data. J Am Stat Assoc 91:1181–1186

Chen YH (2010) Semiparametric marginal regression analysis for dependent competing risks under an assumed copula. J R Stat Soc Ser B 72:235–251

de Uña-Álvarez J (2012) On the Markov three-state progressive model; recent advances in system reliability. Springer, New York

Ding AA (2012) Copula identifiability conditions for dependent truncated data model. Lifetime Data Anal 18(4):397–407

Emura T (2014) R depend.truncation: statistical inference for parametric and semiparametric models based on dependently truncated data. Version 2.1, CRAN

Emura T, Wang W (2010) Testing quasi-independence for truncation data. J Multivar Anal 101:223–239

Emura T, Wang W, Hung HN (2011) Semi-parametric inference for copula models for truncated data. Stat Sin 21:349–367

Emura T, Wang W (2012) Nonparametric maximum likelihood estimation for dependent truncation data based on copulas. J Multivar Anal 110:171–188

Emura T, Konno Y (2012) Multivariate normal distribution approaches for dependently truncated data. Stat Pap 53:133–149

Emura T, Chen YH (2014) Gene selection for survival data under dependent censoring: a copula-based approach. Stat Methods Med Res. doi:10.1177/0962280214533378

Escarela G, Carriere JF (2003) Fitting competing risks with an assumed copula. Stat Methods Med Res 12:333–349

Genest C, Mackay RJ (1986) The joy of copulas: bivariate distributions with uniform marginals. Am Stat 40:280–283

He S, Yang GL (1998) Estimation of the truncation probability in the random truncation model. Ann Stat 26:1011–1027

Hyde J (1977) Testing survival under right censoring and left truncation. Biometrika 64:225–230

Hyde J (1980) Survival analysis with incomplete observations. In: Miller RG, Efron B, Brown BW, Moses LE (eds) Biostatistics casebook. Wiley, New York, pp 31–46

Klein JP, Moeschberger ML (2003) Survival analysis: techniques for censored and truncated data, 2nd edn. Springer, New York

Lagakos SW, Barraj LM, De Gruttola V (1988) Non-parametric analysis of truncated survival data with application to AIDS. Biometrika 75:515–23

Lai TL, Ying Z (1991) Estimating a distribution function with truncated and censored data. Ann Stat 19:417–442

Lynden-Bell D (1971) A method of allowing for known observational selection in small samples applied to 3RC quasars. Mon Nat R Astron Soc Lett 155:95–118

Martin EC, Betensky RA (2005) Testing quasi-independence of failure and truncation via conditional Kendall’s tau. J Am Stat Assoc 100:484–492

Murotani K, Zhou B, Kaneda H, Nakatani E, Kojima S, Nagai Y, Fukushima M (2014) Survival of centenarians in Japan. J Biosoc Sci. doi:10.1017/S0021932014000388

Nelsen RB (2006) An introduction to copulas. Springer, New York

Rivest LP, Wells MT (2001) A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. J Multivar Anal 79:138–155

Rodriguez-Girondo M, de Uña-Álvarez J (2012) Testing Markovian in the three-state progressive model via future-past association. Biom J 54(2):163–180

Sibuya M, Hanayama N (2004) Estimation of human longevity distribution based on tabulated statistics. Proc Inst Stat Math 52:117–134

Strazalkowska-Kominiak E, Stute W (2013) Empirical copulas for consequtive survival data: copulas in survival analysis. TEST 22:688–714

Tsai WY (1990) Testing the association of independence of truncation time and failure time. Biometrika 77:169–177

Wang MC, Jewell NP, Tsai WY (1986) Asymptotic properties of the product-limit estimate under random truncation. Ann Stat 13:1597–1605

Zheng M, Klein J (1995) Estimates of marginal survival for dependent competing risks based on an assumed copula. Biometrika 82:127–38