Metrika (2024) 87:211–245
https://doi.org/10.1007/s00184-023-00911-7

Check for updates

Nonparametric estimation of univariate and bivariate survival functions under right censoring: a survey

Paul Janssen[1,2] · **Noël Veraverbeke**[1,2]

Received: 22 March 2022 / Accepted: 27 April 2023 / Published online: 5 June 2023

Abstract

Survival analysis studies time to event data, also called survival data in biomedical research. The main challenge in the analysis of survival data is to develop inferential methods that take into account the incomplete information contained in censored observations. The seminal paper of Kaplan and Meier (J Am Stat Assoc 53:457–481,1958) gave a boost to the development of statistical methods for time to event data subject to right censoring; methods that have been applied in a broad variety of scientific fields including health, engineering and economy. A basic quantity in survival analysis is the survival distribution: $S(t) = P(T > t)$, with T the time to event or, in case of a bivariate vector of lifetimes (T_1, T_2), $S(t_1, t_2) = P(T_1 > t_1, T_2 > t_2)$. Nonparametric estimation of these basic quantities received, since Kaplan and Meier (J Am Stat Assoc 53:457–481,1958), considerable attention resulting in many publications scattered over a large period of time and a large field of applications. The purpose of this paper is to review, in a unified way, nonparametric estimation of $S(t)$ and $S(t_1, t_2)$ for time to event data subject to right censoring. Interesting to realize is that, in the multivariate setting, the form of the nonparametric estimator for $S(t_1, t_2)$ is determined by the actual censoring scheme. In this survey we focus, for the proposed (implicitly) existing or new nonparametric estimators, on the asymptotic normality. By doing so we fill some gaps in the literature by introducing some new estimators and by providing explicit expressions for the asymptotic variances often not yet available for some of the existing estimators.

Keywords Asymptotic normality · Asymptotic variance · Bivariate survival function · Identifiability · Right censoring

---

✉ Paul Janssen
paul.janssen@uhasselt.be

Noël Veraverbeke
noel.veraverbeke@uhasselt.be

1 Center for Statistics (Data Science Institute), Hasselt University, Agoralaan, 3590 Diepenbeek, Belgium

2 North-West University, Potchefstroom, South Africa

Mathematics Subject Classification Primary 62G05 · 62G20 · 62N01 · 62N02

1 Introduction

Time to event is the time from a given origin to the occurrence time of the event of interest. In many applied fields time to event data occur. Examples include duration analysis (e.g. time to first job after graduation); reliability analysis (e.g. lifetime of a mechanical component); survival analysis (e.g. time from onset to death). Other applied fields include: economy, insurance, demography, biology, public health, epidemiology, veterinary medicine.

Within the context of survival analysis, 'survival data' is standard terminology for time to event data. The outcome of the event can be 'good' (e.g. time to pain relief, time to recovery, time to cure) or can be 'bad' (e.g. time to first relapse, time to death, time from diagnosis to onset).

Time to event data can be univariate (e.g. time from onset of virus infection to cure) or multivariate (e.g. the lifetime of monozygotic male twins; time to blindness in left and right eye in diabetic retinopathy patients; time to tumor in a litter-matched rats (one treated and two control rats) tumorigenesis experiment).

The multivariate examples we mentioned are so called parallel multivariate data. Parallel data sets follow several items/subjects/animals simultaneously. Other types of multivariate data structures (such as longitudinal data, repeated measures data) are discussed in Hougaard (2000). In this survey we focus on parallel multivariate data.

In this paper we focus on univariate and (parallel) bivariate survival data. Given a survival time $T \geq 0$ or given a bivariate survival vector (T_1, T_2) with $T_1 \geq 0$ and $T_2 \geq 0$, primary interest is often in estimating the survival distribution $S(t) = P(T > t)$ (univariate setting) and $S(t_1, t_2) = P(T_1 > t_1, T_2 > t_2)$ (bivariate setting). Survival data are (typically) subject to right censoring, for such data we review, in Sect. 2 of the paper, nonparametric estimation of $S(t) = P(T > t)$. In Sect. 3 we review nonparametric estimation for $S(t_1, t_2) = P(T_1 > t_1, T_2 > t_2)$. After a general introduction (Sect. 3.3), we discuss univariate censoring and one-component censoring. We clearly explain how the censoring scheme determines the definition of the nonparametric estimator and for each estimator we study the asymptotic normality and give an explicit analytic expression for the asymptotic variance (contributing to an open question in Hougaard (2000), p. 457, where he writes 'generally, expressions for the variance are not available').

The advantage of nonparametric estimation when compared to (semi)parametric estimation is that no underlying model assumptions are made (i.e. the inference is completely data driven). Also note that even if (semi)parametric models are used to estimate the survival distribution, nonparametric estimators remain instrumental for goodness-of-fit purposes.

2 Nonparametric estimation of the univariate survival function

2.1 The right random censoring model

In survival analysis, the main object of interest is a nonnegative random variable T, called survival time (or lifetime, failure time, event time,...).

A typical feature is that T is not always observed. Instead of T one sometimes observes some other nonnegative random variable C, called censoring time. In the right random censorship model the observable variables are

$$Y = T \wedge C \quad \text{and} \quad \delta = I(T \leq C)$$

where $a \wedge b = \min(a, b)$ and where I is the indicator function, defined for every event A as $I(A) = 1$ if and only if A holds and zero otherwise.

The right random censorship model assumes that

$$T \text{ and } C \text{ are independent,}$$

the independent censoring assumption. Let $T_1, \ldots, T_n \overset{i.i.d.}{\sim} T$ be independent and identically distributed (i.i.d.) random variables with distribution function F and survival function $S = 1 - F$ and $C_1, \ldots, C_n \overset{i.i.d.}{\sim} C$ with distribution function G. The observations in the model are (Y_i, δ_i), with $Y_i = T_i \wedge C_i$ and $\delta_i = I(T_i \leq C_i)$, $i = 1, \ldots, n$. We clearly have $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} Y$ with distribution function H and due to the independence assumption

$$1 - H(t) = (1 - F(t))(1 - G(t)). \tag{1}$$

The estimation should therefore be based on $(Y_i, \delta_i), i = 1, \ldots, n$. For the uncensored observations we define the following subdistribution function

$$\begin{aligned} H^1(t) &= P(Y \leq t, \delta = 1) = P(T \leq t, T \leq C) \\ &= \int_0^t \int_{y-}^{\infty} G(dx)F(dy) = \int_0^t (1 - G(y-))F(dy). \end{aligned} \tag{2}$$

We also introduce the following notation. For any distribution function L, the upper endpoint of support is denoted by τ_L, i.e. $\tau_L = \inf\{t : L(t) = 1\}$. From (1) it follows that $\tau_H = \tau_F \wedge \tau_G$.

2.2 Identifiability

An important preliminary question is of course the identifiability of the survival function, i.e. the possibility of obtaining the survival function of T from the observations on $Y = T \wedge C$ and $\delta = I(T \leq C)$.

Theorem 1 below shows that the independent censoring assumption on T and C is sufficient for identifiability of the survival function of T. We refer to Tsiatis (1975) for the weaker assumption which says that knowledge of the copula function of T and C is also sufficient. See also Ebrahimi et al. (2003).

Theorem 1 *Assume that T and C are independent with continuous distribution functions F and G. Then, for $t < \tau_H$,*

$$S(t) = \exp\left\{-\int_0^t \frac{H^1(dy)}{1 - H(y)}\right\}.$$

Proof From (2) and the continuity of G we obtain

$$H^1(t) = \int_0^t (1 - G(y))F(dy).$$

Together with (1) this gives

$$\frac{H^1(dy)}{1 - H(y)} = \frac{F(dy)}{1 - F(y)}$$

or, for $t < \tau_H$,

$$1 - F(t) = \exp\left(-\int_0^t \frac{H^1(dy)}{1 - H(y)}\right).$$

2.3 The Kaplan–Meier estimator

The classical nonparametric maximum likelihood estimator under right censoring for the survival function is the estimator of Kaplan and Meier (1958), also called the product-limit estimator (see Supplementary Material) or nonparametric maximum likelihood estimator (see Johansen 1978). For values of t in the range of the data, it is defined as

$$\widehat{S}(t) = 1 - \widehat{F}(t) = \prod_{\substack{i=1 \\ Y_i \le t, \delta_i = 1}}^{n} \left(1 - \frac{d_i}{n_i}\right) \tag{3}$$

with d_i the number of events and n_i the number of subjects/objects at risk at time Y_i, $i = 1, \ldots, n$.

For continuous survival distributions we have that, with probability one, only one event can happen at a time. We then have $d_i = 0$ for $\delta_i = 0$ and $d_i = 1$ for $\delta_i = 1$. This is the situation that we consider in the sequel. In fact we assume that T and C have continuous distribution functions.

Hence the proposed estimator is a step function with jumps at the event times, i.e. the Y_i having $\delta_i = 1, i = 1, \ldots, n$.

The Kaplan–Meier estimator in (3) can then be rewritten as

$$\begin{aligned}\widehat{S}(t) &= \prod_{\substack{i=1\\ Y_i \le t, \delta_i = 1}}^{n} \left(1 - \frac{1}{n_i}\right)\\ &= \prod_{\substack{i=1\\ Y_{(i)} \le t}}^{n} \left(1 - \frac{1}{n-i+1}\right)^{\delta_{(i)}} I(t < Y_{(n)})\end{aligned} \tag{4}$$

with $Y_{(1)} \le \ldots \le Y_{(n)}$ the ordered Y_i's and $\delta_{(1)}, \ldots, \delta_{(n)}$ the corresponding indicators.

The Kaplan–Meier estimator can also be represented as a sum:

$$\widehat{S}(t) = 1 - \widehat{F}(t) = \sum_{i=1}^{n} \Delta_i I(Y_{(i)} > t)$$

where Δ_i is the jump at $Y_{(i)}$. We have

$$\begin{aligned}\Delta_i &= \widehat{S}(Y_{(i)}-) - \widehat{S}(Y_{(i)}) = \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}} - \prod_{j=1}^{i} \left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}}\\ &= \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}} = \frac{\delta_{(i)}}{n} \prod_{j=1}^{i-1} \left(\frac{n-j+1}{n-j}\right)^{1-\delta_{(j)}}\\ &= \frac{\delta_{(i)}}{n} \frac{1}{1-\widehat{G}(Y_{(i)}-)}\end{aligned}$$

with $\widehat{G}$ the Kaplan–Meier estimator for G, i.e. the Kaplan–Meier estimator based on the sample $(Y_i, 1-\delta_i), i = 1, \ldots, n$. Hence,

$$\begin{aligned}\widehat{S}(t) &= \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_{(i)}}{1-\widehat{G}(Y_{(i)}-)} I(Y_{(i)} > t)\\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{1-\widehat{G}(Y_i-)} I(Y_i > t)\\ &\equiv \widehat{S}_{RR}(t),\end{aligned}$$

which is the estimator of Robins and Rotnitzky (1992), also called the 'inverse-probability-of censoring weighted average'. See also Satten and Datta (2001).

Remark 1 The Robins-Rotnitzky estimator stems from the identifying equation approach (see Suplementary Material).

2.4 The Lin-Ying estimator

In this section we want to present an estimator which is new in the univariate case. It has been proposed in the bivariate case by Lin and Ying (1993), but their simple idea can also be used in the univariate situation.

If T and C are independent, then for $t < \tau_H$, (1) implies

$$S(t) = \frac{1 - H(t)}{1 - G(t)}$$

and a simple nonparametric estimator for $S(t)$ is, for $t < Y_{(n)}$,

$$\widehat{S}_{LY}(t) = \frac{1 - H_n(t)}{1 - \widehat{G}(t)} = \frac{1}{n} \frac{1}{1 - \widehat{G}(t)} \sum_{i=1}^{n} I(Y_i > t)$$

where $H_n(t) = n^{-1} \sum_{i=1}^{n} I(Y_i \leq t)$ is the empirical distribution function of $Y_1, \ldots, Y_n$ and $\widehat{G}(t)$ is the Kaplan–Meier estimator for $G(t)$. A nice feature of this estimator is that it jumps at every observation Y_i, $i = 1, \ldots, n$. A drawback is that $\widehat{S}_{LY}$, as estimator of a monotone function S, is not guaranteed to be monotone.

Remark 2 To compare $\widehat{S}_{RR}(t)$ and $\widehat{S}_{LY}(t)$ note that the Lin-Ying estimator has $(1 - \widehat{G}(t))^{-1}$ in front of the summation whereas the Robins-Rotnizky estimator has the weights $\frac{\delta_i}{1 - \widehat{G}(Y_i-)}$, $i = 1, \ldots, n$, inside the sum.

2.5 Asymptotic behaviour of the Kaplan–Meier estimator

The asymptotic properties of the Kaplan–Meier estimator have been studied in great detail in several papers. In this survey we restrict attention to uniform strong consistency and asymptotic normality.

The oldest Glivenko-Cantelli type result is proved in Földes and Rejtő (1981).

Theorem 2 (Földes and Rejtő 1981)
Assume that T and C are independent and that F and G are continuous.
Then, for any $t_0 < \tau_H = \inf\{t : 1 - H(t) = 0\}$,

$$\sup_{0 \leq t \leq t_0} | \widehat{S}(t) - S(t) | = O(n^{-1/2} (\log \log n)^{1/2}) \ a.s.$$

Related important papers are Stute and Wang (1993b) and Gill (1994).

Theorem 3 (Lo and Singh 1986; Major and Rejtő 1988)
Assume that T and C are independent and that F and G are continuous.
Then, for any $t < \tau_H$,

$$\widehat{S}(t) = S(t) - \frac{1}{n} \sum_{i=1}^{n} \xi(t; Y_i, \delta_i) + R_n(t) \ \ a.s.$$

with, for any $t_0 < \tau_H$,

$$\sup_{0 \le t \le t_0} | R_n(t) | = O(n^{-1} \log n) \ \ a.s.$$

The i.i.d. random variables $\xi(t; Y_i, \delta_i)$ in this representation are given by

$$\begin{aligned}
\xi(t; Y_i, \delta_i) &= S(t) \left\{ - \int_0^{Y_i \wedge t} \frac{H^1(dy)}{(1 - H(y))^2} + \frac{I(Y_i \le t, \delta_i = 1)}{1 - H(Y_i)} \right\} \\
&= S(t) \left\{ \int_0^t \frac{I(Y_i \le y) - H(y)}{(1 - H(y))^2} H^1(dy) + \frac{I(Y_i \le t, \delta_i = 1) - H^1(t)}{1 - H(t)} \right. \\
&\quad \left. - \int_0^t \frac{I(Y_i \le y, \delta_i = 1) - H^1(y)}{(1 - H(y))^2} H(dy) \right\}.
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
&E\xi(t; Y, \delta) = 0 \\
&Cov(\xi(t; Y, \delta), \xi(t'; Y, \delta)) = S(t)S(t') \int_0^{t \wedge t'} \frac{H^1(dy)}{(1 - H(y))^2}.
\end{aligned} \tag{5}$$

The calculation of this covariance expression is long and tedious and can be found in Breslow and Crowley (1974), Appendix, p. 450–452.

Corollary 1 *Assume the conditions of Theorem 2. Then, for any $t < \tau_H$,*

$$n^{1/2}(\widehat{S}(t) - S(t)) \xrightarrow{d} N\left(0; S^2(t) \int_0^t \frac{H^1(dy)}{(1 - H(y))^2}\right).$$

Remark 3 The Kaplan–Meier estimator has been extended to the regression case, where next to the observations of (Y, δ) also another variable X, called covariate, is observed. The pioneering paper on nonparametric estimation of the conditional survival function $S(t \mid x) = P(T > t \mid X = x)$ is Beran (1981). He studied the conditional Kaplan–Meier estimator of $S(t \mid x)$ defined by

$$\begin{aligned}
\widehat{S}(t \mid x) &= \prod_{\substack{i=1 \\ Y_i \le t, \delta_i = 1}}^{n} \left(1 - \frac{w_{ni}(t, h_n)}{\sum_{j=1}^n w_{nj}(t, h_n) I(Y_j \ge Y_i)}\right) \\
&= \prod_{\substack{i=1 \\ Y_{(i)} \le t}}^{n} \left(1 - \frac{w_{n(i)}(t, h_n)}{1 - \sum_{j=1}^{i-1} w_{n(j)}(t, h_n)}\right)^{\delta_{(i)}}
\end{aligned}$$

where $Y_{(1)} \leq Y_{(2)} \leq \ldots Y_{(n)}$ are the ordered Y_j's. Also $\delta_{(j)}$ and $w_{n(j)}(t, h_n)$ are the censoring indicator and weight corresponding to that ordering.

The $w_{ni}(t, h_n)$ are some smoothing weights depending on a given probability density function (called kernel) and a nonnegative sequence $\{h_n\}$, tending to 0 as $n \to \infty$ (bandwidth sequence).

Note that $w_n(t, h_n) = n^{-1}$ gives the classical Kaplan–Meier estimator as in (4). Properties of the Beran estimator such as the generalization of the representation of Lo and Singh (Theorem 3) are in González-Manteiga and Cadarso Suarez (1994), Van Keilegom and Veraverbeke (1997).

Remark 4 The Kaplan–Meier estimator has also been extended to the case of dependent censoring, that is T and C are allowed to be dependent. Then, instead of assuming that $P(T \leq t, C \leq c) = F(t)G(c)$ we assume that for a given copula $\mathcal{C}$, $P(T \leq t, C \leq c) = \mathcal{C}(F(t), G(c))$ (see Sklar's theorem in Nelsen 2006).

For this more general setting identifiability is discussed in Tsiatis (1975). Further important references include Zheng and Klein (1995) and Rivest and Wells (2001). The regression case for dependent T and C is considered in Braekers and Veraverbeke (2005).

2.6 Asymptotic behaviour of the Lin-Ying estimator

From Sect. 2.4 we have, for $t < \tau_H$,

$$\begin{aligned}
\widehat{S}_{LY}(t) - S(t) &= \frac{1 - H_n(t)}{1 - \widehat{G}(t)} - \frac{1 - H(t)}{1 - G(t)} \\
&= \frac{1}{(1 - \widehat{G}(t))(1 - G(t))}\{-(1-G(t))(H_n(t) - H(t)) \\
&\quad + (1-H(t))(\widehat{G}(t)-G(t))\} . \qquad (6)
\end{aligned}$$

Theorem 4 *Assume that T and C are independent and that F and G are continuous. Then, for any $t_0 < \tau_H$,*

$$\sup_{0 \leq t \leq t_0} | \widehat{S}_{LY}(t) - S(t) |= O(n^{-1/2}(\log\log n)^{1/2}) \;\; a.s.$$

Proof If $t_0 < \tau_H$, it follows from (6) and the uniform consistency of $\widehat{G}$ that there exist positive constants K_1 and K_2 such that

$$\sup_{0 \leq t \leq t_0} | \widehat{S}_{LY}(t) - S(t) | \leq K_1 \sup_{0 \leq t \leq t_0} | H_n(t) - H(t) | + K_2 \sup_{0 \leq t \leq t_0} | \widehat{G}(t) - G(t) | .$$

Apply the law of iterated logarithm result to the first term and the corresponding result for the Kaplan–Meier estimator (see Theorem 2) to the second term.

Theorem 5 *Assume that T and C are independent and that F and G are continuous.*

Then, for any $t < \tau_H$,

$$\widehat{S}_{LY}(t) = S(t) - \frac{1}{n}\sum_{i=1}^{n} \psi_{LY}(t; Y_i, \delta_i) + \widetilde{R}_n(t) \;\; a.s.$$

with, for any $t_0 < \tau_H$,

$$\sup_{0 \le t \le t_0} \mid \widetilde{R}_n(t) \mid = O(n^{-1} \log n) \;\; a.s.$$

The i.i.d. random variables $\psi_{LY}(t; Y_i, \delta_i)$ are given by

$$\psi_{LY}(t; Y_i, \delta_i) = \frac{1}{1 - G(t)}\{I(Y_i \le t) - H(t)\}$$
$$- \frac{1 - F(t)}{1 - G(t)} \left\{ (1 - G(t)) \left[\int_0^t \frac{I(Y_i \le y) - H(y)}{(1 - H(y))^2} H^0(dy) \right.\right.$$
$$\left.\left. + \frac{I(Y_i \le t, \delta_i = 0) - H^0(t)}{1 - H(t)} - \int_0^t \frac{I(Y_i \le y, \delta_i = 0) - H^0(y)}{(1 - H(y))^2} H(dy) \right] \right\}$$

where $H^0(t) = P(Y \le t, \delta = 0)$ is the subdistribution function of the censored obervations.

Proof From (6) and the consistency of $\widehat{G}$ (using a Slutsky argument) it follows by linearization that $\widehat{S}_{LY}(t) - S(t)$ has the same asymptotic distribution as

$$-\frac{1}{1 - G(t)}\{H_n(t) - H(t)\} + \frac{1 - F(t)}{1 - G(t)} \left\{ \widehat{G}(t) - G(t) \right\}.$$

Plugging in the asymptotic representation of Theorem 3 for $\widehat{G} - G$ gives the desired result. This asymptotic representation for $\widehat{G} - G$ is obtained from Theorem 3 by interchanging the role of F and G and now $1 - \delta_i$ in the role of δ_i.

Corollary 2 *Assume the conditions of Theorem 5. Then, for any fixed $t < \tau_H$,*

$$n^{1/2}(\widehat{S}_{LY}(t) - S(t)) \xrightarrow{d} N(0; Var(\psi_{LY}(t; Y, \delta)).$$

A long but rather straightforward calculation gives that

$$Var(\psi_{LY}(t; Y, \delta)) = S^2(t) \int_0^t \frac{H^1(dy)}{(1 - H(y))^2}$$

and

$$Cov(\psi_{LY}(t;Y,\delta),\psi_{LY}(t';Y,\delta)) = S(t)S(t')\int_0^{t\wedge t'} \frac{H^1(dy)}{(1-H(y))^2}$$

which is exactly the same as for the Kaplan–Meier estimator.

Proof From the expression for $\psi_{LY}(t;Y;\delta)$ in Theorem 5:

$$\begin{aligned}
&\mathrm{Var}(\psi_{LY}(t;Y,\delta)) = \frac{1}{(1-G(t))^2}H(t)(1-H(t))\\
&+(1-F(t))^2\int_0^t \frac{H^0(dy)}{(1-H(y))^2}\\
&-2\frac{1-F(t)}{1-G(t)}E\left\{I(Y\le t)\left[\int_0^t \frac{I(Y\le y)-H(y)}{(1-H(y))^2}H^0(dy)\right.\right.\\
&\left.\left.+\ \frac{I(Y\le t,\delta=0)-H^0(t)}{1-H(t)} - \int_0^t \frac{I(Y\le y,\delta=0)-H^0(t)}{(1-H(y))^2}H(dy)\right]\right\}
\end{aligned}$$

where we used the fact that $E[\psi_{LY}(t;Y,\delta)]=0$ in the covariance term.
Write the expectation above as $E\{(1)+(2)-(3)\}$.

$$\begin{aligned}
E(1) &= (1-H(t))\int_0^t \frac{H(y)}{(1-H(y))^2}H^0(dy)\\
E(2) &= H_0(t)\\
E(3) &= (1-H(t))\int_0^t \frac{H^0(y)}{(1-H(y))^2}H(dy).
\end{aligned}$$

Now using $H^0(dy) = (1-F(y))G(dy)$,

- $$\int_0^t \frac{H(y)}{(1-H(y))^2}H^0(dy) = \int_0^t \frac{1-(1-F(y))(1-G(y))}{(1-F(y))(1-G(y))^2}G(dy)$$
$$= \int_0^t \frac{1}{1-H(y)}\frac{G(dy)}{1-G(y)} + \ln(1-G(t)) = \int_0^t \frac{H^0(dy)}{(1-H(y))^2} + \ln(1-G(t)).$$

- $$\int_0^t \frac{H^0(y)}{(1-H(y))^2} H(dy) = \int_0^t H^0(y) d\left(\frac{1}{1-H(y)}\right)$$

$$= \frac{H^0(t)}{1-H(t)} - \int_0^t \frac{1}{1-H(y)} H^0(dy) = \frac{H^0(t)}{1-H(t)} + \ln(1-G(t)).$$

Hence, $E\{(1)+(2)-(3)\} = (1-H(t) \int_0^t \frac{H^0(dy)}{(1-H(y))^2}$.

$$\text{Var}(\psi_{LY}(t; Y, \delta)) = \frac{1}{(1-G(t))^2}$$

$$H(t)(1-H(t)) - (1-F(t))^2 \int_0^t \frac{H^0(dy)}{(1-H(y))^2}.$$

Use $H(y) = H^0(y) + H^1(y)$ to obtain

$$\text{Var}(\psi_{LY}(t; Y, \delta) = (1-F(t))^2 \int_0^t \frac{H^1(dy)}{(1-H(y))^2}$$

$$+ \frac{1}{(1-G(t))^2} H(t)(1-H(t)) - (1-F(t))^2 \int_0^t \frac{H(dy)}{(1-H(y))^2}.$$

Since $\int_0^t \frac{H(dy)}{(1-H(y))^2} = \frac{H(t)}{1-H(t)}$, we have

$$\text{Var}(\psi_{LY}(t; Y, \delta)) = (1-F(t))^2 \int_0^t \frac{H^1(dy)}{(1-H(y))^2}.$$

Remark 5 We are grateful to one of the referees for insisting on a further (very long) calculation, in line with the calculations for the asymptotic variance, of the asymptotic covariance. Given that the asymptotic covariances of Lin-Ying estimator and the Kaplan–Meier estimator coincide, the Lin-Ying process in t is first order asymptotic equivalent with the Kaplan-Meier process in t.

3 Nonparametric estimation of the bivariate survival function

3.1 The bivariate right random censoring model

In the bivariate setting we have a vector (T_1, T_2) of nonnegative random variables, subject to right random censoring by a vector (C_1, C_2) of nonnegative censoring variables. The observable variables are (Y_1, Y_2) and (δ_1, δ_2) with, for $j = 1, 2$,

$$Y_j = T_j \wedge C_j \quad \text{and} \quad \delta_j = I(T_j \leq C_j).$$

The observations in the model are $(Y_{1i}, Y_{2i}, \delta_{1i}, \delta_{2i})$ with, for $i = 1, \ldots, n$ and $j = 1, 2$, $Y_{ji} = T_{ji} \wedge C_{ji}$ and $\delta_{ji} = I(T_{ji} \leq C_{ji})$ and $(Y_{1i}, Y_{2i}, \delta_{1i}, \delta_{2i})$ are i.i.d. as $(Y_1, Y_2, \delta_1, \delta_2)$. Note that (T_{1i}, T_{2i}), $i = 1, \ldots, n$, is an i.i.d. sequence with joint distribution function $F(t_1, t_2)$ and joint survival function $S(t_1, t_2)$, and (C_{1i}, C_{2i}), $i = 1, \ldots, n$, is an i.i.d. sequence with joint distribution function $G(t_1, t_2)$ and joint survival function $S_G(t_1, t_2)$.

3.2 Identifiability

For the bivariate censoring model there is a result analogous to Theorem 1. It is due to Langberg and Shaked (1982) and shows that the survival function of (T_1, T_2) is identifiable under the assumption of independence of the vectors (T_1, T_2) and (C_1, C_2). We refer to Pruitt (1993) for a discussion on possible other sufficient conditions (depending on the type of data).

Theorem 6 (Langberg and Shaked 1982)
Assume that (T_1, T_2) and (C_1, C_2) are independent and that the marginal distributions F_1, F_2, G_1, G_2 of T_1, T_2, C_1, C_2 are continuous. Then $S(t_1, t_2)$ is identifiable on the set $\Omega \cup \widetilde{\Omega}$, where

$$\Omega = \{(t_1, t_2) : t_2 < \tau_{H_2}, t_1 < \tau_{H_1}(t_2)\}$$
$$\widetilde{\Omega} = \{(t_1, t_2) : t_1 < \tau_{H_1}, t_2 < \tau_{H_2}(t_1)\}$$

with τ_{H_1}, τ_{H_2}, $\tau_{H_1}(t_2)$, $\tau_{H_2}(t_1)$ the right endpoints of support of $H_1(t) = P(Y_1 \leq t)$, $H_2(t) = P(Y_2 \leq t)$, $P(Y_1 \leq v \mid Y_2 > t_2)$, $P(Y_2 \leq v \mid Y_1 > t_1)$.
For all $(t_1, t_2) \in \Omega$ we have

$$S(t_1, t_2) = \exp\left(-\int_0^{t_2} \frac{dP(Y_2 \leq u, \delta_2 = 1)}{1 - H_2(u)}\right) \exp\left(-\int_0^{t_1} \frac{dP(Y_1 \leq v, \delta_1 = 1 \mid Y_2 > t_2)}{P(Y_1 > v \mid Y_2 > t_2)}\right) \tag{7}$$

and for all $(t_1, t_2) \in \widetilde{\Omega}$ we have

$$S(t_1, t_2) = \exp\left(-\int_0^{t_1} \frac{dP(Y_1 \leq u, \delta_1 = 1)}{1 - H_1(u)}\right) \exp\left(-\int_0^{t_2} \frac{dP(Y_2 \leq v, \delta_2 = 1 \mid Y_1 > t_1)}{P(Y_2 > v \mid Y_1 > t_1)}\right).$$

Proof We have, using independence of T_2 and C_2,

$$\begin{aligned} S(t_1, t_2) &= P(T_2 > t_2)P(T_1 > t_1 \mid T_2 > t_2) \\ &= P(T_2 > t_2)P(T_1 > t_1 \mid Y_2 > t_2). \end{aligned} \tag{8}$$

From the assumptions and Theorem 1, we have that the first factor in (8) is equal to the first factor in (7).

From the assumptions it also follows that $T_1 \mid Y_2 > t_2$ and $C_1 \mid Y_2 > t_2$ are independent and have continuous distributions. Apply again Theorem 1, to see that the second factor in (8) is equal to the second factor in (7).

The second expression for $S(t_1, t_2)$ for $(t_1, t_2) \in \widetilde{\Omega}$ follows similarly, starting from

$$S(t_1, t_2) = P(T_1 > t_1)P(T_2 > t_2 \mid Y_1 > t_1).$$

3.3 Bivariate extensions of the Kaplan–Meier estimator

Under the assumption of independence of the vector (T_1, T_2) and (C_1, C_2), several nonparametric estimators for the survival function $S(t_1, t_2)$ have been proposed in the literature.

To obtain a nonparametric estimator of the survival function Dabrowksa (1988) has used a two-dimensional product-limit approximation (see also Pruitt 1991). Prentice and Cai (1992) used an approximation based on Peano series and van der Laan (1996) has taken an approach based on nonparametric maximum likelihood ideas. A look at the proposed solutions shows that bivariate censoring complicates nonparametric inference and makes it a hard problem. All proposals have one or more drawbacks such as, lack of monotonicity, non-uniqueness, slow rate of convergence, no analytic variance expression. See also Gill (1992) and see Prentice and Zhao (2018) for an excellent recent review with focus on these approaches. These complicated estimators will not be discussed in this survey.

Our approach to study nonparametric estimation of $S(t_1, t_2)$ given bivariate right censored time to event data follows the inverse probability weighting (IPW) idea of Robins and Rotnitzky (see also Burke 1988; Satten and Datta 2001; Lopez 2012). After a general starting point, we consider specific bivariate censoring schemes and for these we work out the asymptotic distribution theory of the proposed nonparametric estimators in detail.

Also the simpler nonparametric estimators we propose share some of the drawbacks mentioned above. For example IPW estimators have been criticized for not using all the information contained in the data, and the Lin-Ying type estimators do not define a true distribution and are not necessarily monotone. Our estimators, however, show remarkable good behaviour in concrete applied situations (see the simulations in Geerdens et al. 2016; Abrams et al. 2021, 2023). Their performance can also depend on the time region where they are used (Geerdens et al. 2016). It is clear that the finite sample quality always needs to checked by detailed simulations (see e.g. Prentice and Zhao (2018)).

As, for example, in Burke (1988), introduce the following subdistribution function

$$H^{11}(t_1, t_2) = P(Y_1 \leq t_1, Y_2 \leq t_2, \delta_1 = 1, \delta_2 = 1).$$

We have, under independence of (T_1, T_2) and (C_1, C_2),

$$H^{11}(t_1, t_2) = P(T_1 \leq t_1, T_2 \leq t_2, T_1 \leq C_1, T_2 \leq C_2)$$

$$= \int_0^{t_1}\int_0^{t_2} P(C_1 \geq y_1, C_2 \geq y_2) F(dy_1, dy_2)$$

$$= \int_0^{t_1}\int_0^{t_2} S_G(y_1-, y_2-) F(dy_1, dy_2).$$

Hence

$$F(dy_1, dy_2) = \frac{1}{S_G(y_1-, y_2-)} H^{11}(dy_1, dy_2)$$

or

$$S(t_1, t_2) = \int_{t_1}^{\infty}\int_{t_2}^{\infty} \frac{1}{S_G(y_1-, y_2-)} H^{11}(dy_1, dy_2). \tag{9}$$

In the Supplementary Material we show how (9) can be obtained from the identifying equation idea. This will also be demonstrated for (10) (Sect. 3.4) and (15) (Sect. 3.6).

An estimator for $S(t_1, t_2)$ is obtained by plugging in appropriate estimators $\widehat{H}^{11}$ for H^{11} and $\widehat{S}_G$ for S_G. For $\widehat{H}^{11}$ we can take

$$\widehat{H}^{11}(t_1, t_2) = \frac{1}{n}\sum_{i=1}^{n} \delta_{1i}\delta_{2i} I(Y_{1i} \leq t_1, Y_{2i} \leq t_2)$$

which gives

$$\widehat{S}(t_1, t_2) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_{1i}\delta_{2i}}{\widehat{S}_G(Y_{1i}-, Y_{2i}-)} I(Y_{1i} > t_1, Y_{2i} > t_2).$$

Since, in general C_1 and C_2 are not independent, there exists a (survival) copula $\mathcal{C}$ such that $S_G(t_1, t_2) = \mathcal{C}(S_{G_1}(t_1), S_{G_2}(t_2))$ with $S_{G_j}(t_j) = 1 - G_j(t_j)$ and $G_j(t_j)$ the marginal distributions corresponding to $G(t_1, t_2)$, $j = 1, 2$.

The presence of $\mathcal{C}$ complicates the situation. If $\mathcal{C}$ is known, S_G can be estimated as

$$\widehat{S}_G(t_1, t_2) = \mathcal{C}(\widehat{S}_{G_1}(t_1), \widehat{S}_{G_2}(t_2))$$

where $\widehat{S}_{G_j}(t_j) = 1 - \widehat{G}_j(t_j)$ with $\widehat{G}_j(t_j)$ the Kaplan–Meier estimator of $G_j(t_j)$, $j = 1, 2$.

The study of $\widehat{S}_G(t_1, t_2)$ in the general setting, although possible, is challenging and an explicit expression for the asymptotic variance is hard to obtain (see p. 457 in Hougaard 2000). However, for specific censoring schemes, explicit estimators for $\widehat{S}_G(t_1, t_2)$—and hence for $\widehat{S}(t_1, t_2)$—can be given and the asymptotic normality of

$\widehat{S}(t_1, t_2)$ can be obtained with an explicit analytic expression for the asymptotic variance.

In the sequel we study in detail univariate censoring (Sects. 3.4 and 3.5) and one-component censoring (Sects. 3.6–3.8).

3.4 Estimation of the bivariate survival function under univariate censoring

In this situation (T_1, T_2) is subject to right censoring by a single censoring variable C with univariate distribution function $G(c) = P(C \le c)$.

We assume that (T_1, T_2) and C are independent.

Denote $Y_1 = T_1 \wedge C$, $Y_2 = T_2 \wedge C$, $\delta_1 = I(T_1 \le C)$, $\delta_2 = I(T_2 \le C)$. Also, with $a \vee b = \max(a, b)$,

$$\begin{aligned} H^{11}(t_1, t_2) &= P(Y_1 \le t_1, Y_2 \le t_2, \delta_1 = 1, \delta_2 = 1) \\ &= P(T_1 \le t_1, T_2 \le t_2, T_1 \vee T_2 \le C) \\ &= \int_0^{t_1}\int_0^{t_2} [1 - G((y_1 \vee y_2)-)]\, F(dy_1, dy_2). \end{aligned}$$

As in Sect. 3.3 we obtain, for continuous G,

$$\begin{aligned} S(t_1, t_2) &= \int_{t_1}^{\infty}\int_{t_2}^{\infty} \frac{1}{1 - G((y_1 \vee y_2)-)} H^{11}(dy_1, dy_2) \\ &= \int_{t_1}^{\infty}\int_{t_2}^{\infty} \frac{1}{1 - G(y_1 \vee y_2)} H^{11}(dy_1, dy_2). \end{aligned} \tag{10}$$

As empirical version for $H^{11}(t_1, t_2)$ we again use

$$\widehat{H}^{11}(t_1, t_2) = \frac{1}{n}\sum_{i=1}^{n} \delta_{1i}\delta_{2i} I(Y_{1i} \le t_1, Y_{2i} \le t_2).$$

To estimate G, we note that C is observed if $T_1 > C$ or $T_2 > C$, i.e. $T_1 \vee T_2 > C$. Therefore G can be estimated by a Kaplan–Meier estimator $\widehat{G}$, calculated from $\{C_i \wedge (T_{1i} \vee T_{2i}), I(C_i \le T_{1i} \vee T_{2i}\} = \{Y_{1i} \vee Y_{2i}, \delta_i^{\max}\}$ with $\delta_i^{\max} = 1 - \delta_{1i}\delta_{2i}$.

The estimator for $S(t_1, t_2)$ is

$$\begin{aligned} \widehat{S}(t_1, t_2) &= \int_{t_1}^{\infty}\int_{t_2}^{\infty} \frac{1}{1 - \widehat{G}((y_1 \vee y_2)-)} \widehat{H}^{11}(dy_1, dy_2) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_{1i}\delta_{2i}}{1 - \widehat{G}((Y_{1i} \vee Y_{2i})-)} I(Y_{1i} > t_1, Y_{2i} > t_2). \end{aligned}$$

In the next theorem, the supports of the underlying distributions are important. Let $\tau_{F_1}, \tau_{F_2}, \tau_G, \tau_{H_1}, \tau_{H_2}$ be the right endpoints of support of the distribution functions F_1, F_2, G, H_1, H_2 of T_1, T_2, C, Y_1, Y_2. We impose the condition

$$\tau_G > \tau_{F_1} \vee \tau_{F_2}. \tag{11}$$

This will imply that $1 - G(y_1 \vee y_2) > 0$ for $y_1 < \tau_{F_1}$ and $y_2 < \tau_{F_2}$.

Another consequence of (11) is that $\tau_{H_1} = \tau_{F_1}$ and $\tau_{H_2} = \tau_{F_2}$. Indeed, since $Y_1 = T_1 \wedge C$ and $Y_2 = T_2 \wedge C$, we have that $\tau_{H_1} = \tau_{F_1} \wedge \tau_G = \tau_{F_1}$ and $\tau_{H_2} = \tau_{F_2} \wedge \tau_G = \tau_{F_2}$.

Also, note that $P(Y_1 > \tau_{H_1}, Y_2 > \tau_{H_2}) = 0$.

Theorem 7 *Assume that (T_1, T_2) and C are independent and that the distribution functions F_1, F_2 and G are continuous. Also assume condition (11). Then, for $t_1 < \tau_{F_1}$, $t_2 < \tau_{F_2}$ with $S(t_1, t_2) > 0$, we have the following asymptotic representation:*

$$\begin{aligned}
&\widehat{S}(t_1, t_2) - S(t_1, t_2) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_{1i}\delta_{2i}}{1 - G(Y_{1i} \vee Y_{2i})} I(Y_{1i} > t_1, Y_{2i} > t_2) - S(t_1, t_2) \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} \int_{t_1}^{\infty} \int_{t_2}^{\infty} \frac{\xi(y_1 \vee y_2; Y_{1i} \vee Y_{2i}, \delta_i^{\max})}{(1 - G(y_1 \vee y_2))^2} H^{11}(dy_1, dy_2) \; + o_P(n^{-1/2})
\end{aligned}$$

where, with $\delta^{\max} = 1 - \delta_1\delta_2$,

$$\begin{aligned}
&\xi(t; Y_1 \vee Y_2, \delta^{\max}) \\
&= (1 - G(t)) \left\{ \int_0^t \frac{I(Y_1 \vee Y_2 \le y) - \widetilde{H}(y)}{(1 - \widetilde{H}(y))^2} \widetilde{H}^1(dy) \right. \\
&\quad + \frac{I(Y_1 \vee Y_2 \le t, \delta^{\max} = 1) - \widetilde{H}^1(t)}{1 - \widetilde{H}(t)} \\
&\quad \left. - \int_0^t \frac{I(Y_1 \vee Y_2 \le y, \delta^{\max} = 1) - \widetilde{H}^1(y)}{(1 - \widetilde{H}(y))^2} \widetilde{H}(dy) \right\}
\end{aligned}$$

and

$$\begin{aligned}
\widetilde{H}(t) &= P(Y_1 \vee Y_2 \le t) \\
\widetilde{H}^1(t) &= P(Y_1 \vee Y_2 \le t, \delta^{\max} = 1) \\
&= P(Y_1 \vee Y_2 \le t, \delta_1 = 0 \text{ or } \delta_2 = 0).
\end{aligned} \tag{12}$$

Proof

$$\begin{aligned}
&\widehat{S}(t_1, t_2) - S(t_1, t_2) \\
&= \int_{t_1}^{\infty} \int_{t_2}^{\infty} \frac{1}{1 - \widehat{G}((y_1 \vee y_2)-)} \widehat{H}^1(dy_1, dy_2) - \int_{t_1}^{\infty} \int_{t_2}^{\infty} \frac{1}{1 - G(y_1 \vee y_2)} H^{11}(dy_1, dy_2)
\end{aligned}$$

$$
\begin{aligned}
&= \int_{t_1}^{\infty}\int_{t_2}^{\infty} \left\{ \frac{1}{1-\widehat{G}((y_1 \vee y_2)-)} - \frac{1}{1-\widehat{G}(y_1 \vee y_2)} \right\} (\widehat{H}^{11}(dy_1, dy_2) - H^{11}(dy_1, dy_2)) \\
&\quad + \int_{t_1}^{\infty}\int_{t_2}^{\infty} \left\{ \frac{1}{1-\widehat{G}(y_1 \vee y_2)} - \frac{1}{1-G(y_1 \vee y_2)} \right\} \widehat{H}^{11}(dy_1, dy_2) \\
&\quad + \int_{t_1}^{\infty}\int_{t_2}^{\infty} \left\{ \frac{1}{1-\widehat{G}((y_1 \vee y_2)-)} - \frac{1}{1-\widehat{G}(y_1 \vee y_2)} \right\} H^{11}(dy_1, dy_2) \\
&\quad + \int_{t_1}^{\infty}\int_{t_2}^{\infty} \frac{1}{1-G(y_1 \vee y_2)} \widehat{H}^{11}(dy_1, dy_2) - S(t_1, t_2) \\
&\equiv (a) + (b) + (c) + (d).
\end{aligned} \tag{13}
$$

We have

$$
\begin{aligned}
\mid (c) \mid &\leq \int_{t_1}^{\tau_{F_1}}\int_{t_2}^{\tau_{F_2}} \frac{\widehat{G}((y_1 \vee y_2)-) - \widehat{G}((y_1 \vee y_2)}{(1-\widehat{G}((y_1 \vee y_2))^2} H^{11}(dy_1, dy_2) \\
&\leq \frac{1}{(1-\widehat{G}(\tau_{F_1} \vee \tau_{F_2}))^2} \sup_{s \leq \tau_{F_1} \vee \tau_{F_2}} \mid \widehat{G}(s-) - \widehat{G}(s) \mid \int_{t_1}^{\tau_{F_1}}\int_{t_2}^{\tau_{F_2}} H^{11}(dy_1, dy_2) \\
&= O_p(n^{-1})
\end{aligned}
$$

since $\widehat{G}(\tau_{F_1} \vee \tau_{F_2})$ is a consistent estimator for $G(\tau_{F_1} \vee \tau_{F_2})$ and since the jump of the Kaplan–Meier estimator is $O_P(n^{-1})$, uniformly (See Sect. 2.3).

Similarly $\mid (a) \mid = O_P(n^{-1})$.

For (b) we replace the expression $(\widehat{G} - G)/[(1-\widehat{G})(1-G)]$ in the integrand by $(\widehat{G} - G)/(1-G)^2$ and use the consistency result for $\widehat{G}$ in Theorem 2. It then follows that

$$
\begin{aligned}
&\widehat{S}(t_1, t_2) - S(t_1, t_2) \\
&\quad = \int_{t_1}^{\infty}\int_{t_2}^{\infty} \frac{1}{1-G(y_1 \vee y_2)} \widehat{H}^{11}(dy_1, dy_2) - S(t_1, t_2) \\
&\quad + \int_{t_1}^{\infty}\int_{t_2}^{\infty} \frac{\widehat{G}(y_1 \vee y_2) - G(y_1 \vee y_2)}{(1-G(y_1 \vee y_2))^2} \widehat{H}^{11}(dy_1, dy_2) \\
&\quad + O_P(n^{-1} \log\log n).
\end{aligned}
$$

In the second term we plug in the asymptotic representation of Lo and Singh (1986) (see Theorem 3). This gives that the second term becomes

$$\frac{1}{n}\sum_{i=1}^{n}\int_{t_1}^{\infty}\int_{t_2}^{\infty}\frac{\xi(y_1\vee y_2; Y_{1i}\vee Y_{2i},\delta_i^{\max})}{(1-G(y_1\vee y_2))^2}\widehat{H}^{11}(dy_1,dy_2)+O(n^{-1}\log n)\ \text{a.s.}$$
$$=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\xi(Y_{1j}\vee Y_{2j}; Y_{1i}\vee Y_{2i},\delta_i^{\max})}{(1-G(Y_{1j}\vee Y_{2j}))^2}I(Y_{1j}>t_1,Y_{2j}>t_2,\delta_{1j}=1,\delta_{2j}=1)$$
$$+O_P(n^{-1}\log n).$$

The double sum term in the above expression is a V-statistic with kernel

$$h((y_{1i},y_{2i},\delta_{1i},\delta_{2i}),(y_{1j},y_{2j},\delta_{1j},\delta_{2j}))$$
$$=\frac{\xi(y_{1j}\vee y_{2j}; y_{1i}\vee y_{2i},\delta_i^{\max})}{(1-G(y_{1j}\vee y_{2j}))^2}I(y_{1j}>t_1,y_{2j}>t_2,\delta_{1j}=1,\delta_{2j}=1).$$

We have

$$E[h((y_1,y_2,\delta_1,\delta_2),(Y_{1j},Y_{2j},\delta_{1j},\delta_{2j}))]$$
$$=\int_{t_1}^{\infty}\int_{t_2}^{\infty}\frac{\xi(y_1'\vee y_2'; y_1\vee y_2,\delta^{\max})}{(1-G(y_1'\vee y_2'))^2}H^{11}(dy_1',dy_2')$$

and

$$E[h((Y_{1i},Y_{2i},\delta_{1i},\delta_{2i}),(y_1,y_2,\delta_1,\delta_2))]$$
$$=\frac{E[\xi(y_1\vee y_2; Y_{1i}\vee Y_{2i},\delta_i^{\max})]}{(1-G(y_1\vee y_2))^2}I(y_1>t_1,y_2>t_2,\delta_1=1,\delta_2=1)$$
$$=0.$$

Hence the Hajek projection of the V-statistic is

$$\frac{1}{n}\sum_{i=1}^{n}\int_{t_1}^{\infty}\int_{t_2}^{\infty}\frac{\xi(y_1\vee y_2; Y_{1i}\vee Y_{2i},\delta_i^{\max})}{(1-G(y_1\vee y_2))^2}H^{11}(dy_1,dy_2)$$

and the remainder term is $o_p(n^{-1/2})$.

This follows from the asymptotic theory for the V-statistic and the corresponding U-statistic (Serfling 1980). The required moment conditions are satisfied since the kernel h is bounded. Indeed, ξ is bounded and $1/(1-G(y_1\vee y_2))\le 1/(1-G(\tau_{F_1}\vee\tau_{F_2}))$

since $\tau_G > \tau_{F_1} \vee \tau_{F_2}$. Note that symmetry of the kernel is not required for this type of result.

This proves the theorem.

Corollary 3 *Assume the conditions of Theorem 7. Then, for any $t_1 < \tau_{F_1}$, $t_2 < \tau_{F_2}$ with $S(t_1, t_2) > 0$, we have*

$$n^{1/2}(\widehat{S}(t_1, t_2) - S(t_1, t_2)) \to N(0; \sigma^2(t_1, t_2))$$

where

$$\begin{aligned}
\sigma^2(t_1, t_2) = &\int_{t_1}^{\infty}\int_{t_2}^{\infty} \frac{1}{1 - G(y_1 \vee y_2)} F(dy_1, dy_2) - S^2(t_1, t_2) \\
&+ \int_{t_1}^{\infty}\int_{t_2}^{\infty}\int_{t_1}^{\infty}\int_{t_2}^{\infty} \left(\int_0^{(y_1 \vee y_2)\wedge(y_1' \vee y_2')} \frac{\widetilde{H}^1(dy)}{(1 - \widetilde{H}(y))^2} \right) F(dy_1, dy_2) F(dy_1', dy_2') \\
&-2S^2(t_1, t_2) \int_0^{t_1 \vee t_2} \frac{\widetilde{H}^1(dy)}{(1 - \widetilde{H}(y))^2}. \qquad (14)
\end{aligned}$$

Proof This follows from the asymptotic representation in Theorem 7 which is of the form

$$\widehat{S}(t_1, t_2) - S(t_1, t_2) = \frac{1}{n}\sum_{i=1}^{n} A_i + \frac{1}{n}\sum_{i=1}^{n} B_i + o_P(n^{-1/2}).$$

In the Supplementary Material we show that

$$\begin{aligned}
Var(A_i) &= \int_{t_1}^{\infty}\int_{t_2}^{\infty} \frac{1}{1 - G(y_1 \vee y_2)} F(dy_1, dy_2) - S^2(t_1, t_2) \\
Var(B_i) &= \int_{t_1}^{\infty}\int_{t_2}^{\infty}\int_{t_1}^{\infty}\int_{t_2}^{\infty} \left(\int_0^{(y_1 \vee y_2)\wedge(y_1' \vee y_2')} \frac{\widetilde{H}^1(dy)}{(1 - \widetilde{H}(y))^2} \right) F(dy_1, dy_2) F(dy_1', dy_2') \\
\mathrm{Cov}(A_i, B_i) &= -S^2(t_1, t_2) \int_0^{t_1 \vee t_2} \frac{\widetilde{H}^1(dy)}{(1 - \widetilde{H}(y))^2}.
\end{aligned}$$

Remark 6 In case of no censoring

$$\sigma^2(t_1, t_2) = S(t_1, t_2)(1 - S(t_1, t_2)).$$

Indeed in this case $\delta_1 = \delta_2 \equiv 1$, $\widetilde{H}^1 \equiv 0$, $G \equiv 1$.

3.5 Estimator of Lin-Ying for the bivariate survival function under univariate censoring

For bivariate survival data subject to univariate censoring an alternative estimator has been proposed by Lin and Ying (1993). It is based on the following simple idea. Given the assumed independence of (T_1, T_2) and C we have

$$\begin{aligned} P(Y_1 > t_1, Y_2 > t_2) &= P(T_1 > t_1, T_2 > t_2, C > t_1, C > t_2) \\ &= S(t_1, t_2) P(C > t_1 \vee t_2). \end{aligned}$$

This leads, for $t_1 \vee t_2 < (Y_1 \vee Y_2)_{(n)}$, to the following estimator

$$\widehat{S}_{LY}(t_1, t_2) = \frac{\frac{1}{n}\sum_{i=1}^{n} I(Y_{1i} > t_1, Y_{2i} > t_2)}{1 - \widehat{G}(t_1 \vee t_2)},$$

where $\widehat{G}$ is the Kaplan–Meier estimator for G given in Sect. 3.4. Note that in the absence of censoring, $\widehat{S}_{LY}$ reduces to the usual bivariate empirical survival function.

Theorem 8 *Assume that (T_1, T_2) and C are independent and that the distribution functions F_1, F_2 and G are continuous. Assume condition (11), i.e. $\tau_G > \tau_{F_1} \vee \tau_{F_2}$. Then, for $t_1 < \tau_{F_1}$, $t_2 < \tau_{F_2}$ with $S(t_1, t_2) > 0$, we have the following asymptotic representation*

$$\widehat{S}_{LY}(t_1, t_2) - S(t_1, t_2) = \frac{1}{n}\sum_{i=1}^{n} \psi_{LY}(t_1, t_2, Y_{1i}, Y_{2i}, \delta_{1i}, \delta_{2i}) + o_P(n^{-1/2})$$

where

$$\begin{aligned} &\psi_{LY}(t_1, t_2, Y_1, Y_2, \delta_1, \delta_2) \\ &= \frac{1}{1 - G(t_1 \vee t_2)}\{I(Y_1 > t_1, Y_2 > t_2) - P(Y_1 > t_1, Y_2 > t_2)\} \\ &+ \frac{S(t_1, t_2)}{1 - G(t_1 \vee t_2)}\xi(t_1 \vee t_2; Y_1 \vee Y_2, \delta^{\max}) \end{aligned}$$

where $\xi(t, Y_1 \vee Y_2, \delta^{\max})$ is defined in (12).

Proof As in the proof of Theorem 5, we note that by linearization and by consistency of $\widehat{G}$, $\widehat{S}_{LY}(t_1, t_2) - S(t_1, t_2)$ has the same asymptotic distribution as

$$\begin{aligned} &\frac{1}{1 - G(t_1 \vee t_2)}\left\{\frac{1}{n}\sum_{i=1}^{n} I(Y_{1i} > t_1, Y_{2i} > t_2) - P(Y_1 > t_1, Y_2 > t_2)\right\} \\ &+ \frac{S(t_1, t_2)}{1 - G(t_1 \vee t_2)}\left\{\widehat{G}(t_1 \vee t_2) - G(t_1 \vee t_2)\right\}. \end{aligned}$$

Then plug in the asymptotic representation for $\widehat{G}(t_1 \vee t_2) - G(t_1 \vee t_2)$.

Corollary 4 *Assume the conditions of Theorem 8. Then, for any* $t_1 < \tau_{F_1}$, $t_2 < \tau_{F_2}$ *with* $S(t_1, t_2) > 0$, *we have*

$$n^{1/2}(\widehat{S}_{LY}(t_1, t_2) - S(t_1, t_2)) \xrightarrow{d} N(0; \sigma^2_{LY}(t_1, t_2))$$

where

$$\begin{aligned}
\sigma^2_{LY}(t_1, t_2) &= \frac{1}{(1 - G(t_1 \vee t_2))^2} P(Y_1 > t_1, Y_2 > t_2)(1 - P(Y_1 > t_1, Y_2 > t_2)) \\
&\quad -2S^2(t_1, t_2) \int_0^{t_1 \vee t_2} \frac{\widetilde{H}^1(dy)}{(1 - \widetilde{H}(y))^2} \\
&= \frac{1}{(1 - G(t_1 \vee t_2))^2} P(Y_1 > t_1, Y_2 > t_2)(1 - P(Y_1 > t_1, Y_2 > t_2)) \\
&\quad -S^2(t_1, t_2) \int_0^{t_1 \vee t_2} \frac{G(dy)}{(1 - G(y))^2(1 - F(y, y))}
\end{aligned}$$

with $\widetilde{H}^1$ *and* $\widetilde{H}$ *as defined in (12) of Sect. 3.4.*

Proof

$$\begin{aligned}
\sigma^2_{LY}(t_1, t_2) &= \frac{1}{(1 - G(t_1 \vee t_2))^2} P(Y_1 > t_1, Y_2 > t_2)(1 - P(Y_1 > t_1, Y_2 > t_2)) \\
&\quad +S^2(t_1, t_2) \int_0^{t_1 \vee t_2} \frac{\widetilde{H}^1(dy)}{(1 - \widetilde{H}(y))^2} \\
&\quad +2\frac{S(t_1, t_2)}{(1 - G(t_1 \vee t_2))^2} E\{I(Y_1 > t_1, Y_2 > t_2)\xi(t_1 \vee t_2; Y_1 \vee Y_2, \delta^{\max})\}.
\end{aligned}$$

The expectation above is equal to

$$\begin{aligned}
&-P(Y_1 > t_1, Y_2 > t_2)(1 - G(t_1 \vee t_2)) \left\{ \int_0^{t_1 \vee t_2} \frac{\widetilde{H}(y)}{(1 - \widetilde{H}(y))^2} \widetilde{H}^1(dy) \right. \\
&\quad \left. + \frac{\widetilde{H}^1(t_1 \vee t_2)}{1 - \widetilde{H}(t_1 \vee t_2))} - \int_0^{t_1 \vee t_2} \frac{\widetilde{H}^1(y)}{(1 - \widetilde{H}(y))^2} \widetilde{H}(dy) \right\} \\
&= -P(Y_1 > t_1, Y_2 > t_2)(1 - G(t_1 \vee t_2)) \int_0^{t_1 \vee t_2} \frac{\widetilde{H}^1(dy)}{(1 - \widetilde{H}(y))^2}
\end{aligned}$$

using the calculation in the proof of Corollary 3.

Hence,

$$\sigma^2_{LY}(t_1,t_2) = \frac{1}{(1-G(t_1\vee t_2))^2} P(Y_1>t_1,Y_2>t_2)(1-P(Y_1>t_1,Y_2>t_2)) \\ -S^2(t_1,t_2)\int_0^{t_1\vee t_2}\frac{\widetilde{H}^1(dy)}{(1-\widetilde{H}(y))^2}.$$

This can be rewritten by using the expressions: $1-\widetilde{H}(y) = (1-G(y))(1-F(y,y))$, $\widetilde{H}^1(dy) = (1-F(y,y))G(dy)$ and $P(Y_1>t_1,Y_2>t_2) = S(t_1,t_2)(1-G(t_1\vee t_2))$.

Remark 7 In case of no censoring

$$\sigma^2_{LY}(t_1,t_2) = S(t_1,t_2)(1-S(t_1,t_2)).$$

Remark 8 Wang and Wells (1997) use a different estimator for the denominator in the Lin and Ying (1993) estimator. Since $G(t_1\vee t_2) = G(t_1)\vee G(t_2)$, they estimate $1-G(t_1\vee t_2)$ by $1-(\widehat{G}(t_1)\vee\widehat{G}(t_2))$:

$$\widehat{S}_{WW}(t_1,t_2) = \frac{\frac{1}{n}\sum_{i=1}^n I(Y_{1i}>t_1,Y_2>t_2).}{1-(\widehat{G}(t_1)\vee\widehat{G}(t_2))}.$$

Similar calculations as before give for the asymptotic variance:

$$\sigma^2_{WW}(t_1,t_2) = \frac{1}{(1-G(t_1\vee t_2))^2} P(Y_1>t_1,Y_2>t_2)(1-P(Y_1>t_1,Y_2>t_2)) \\ -S^2(t_1,t_2)\times\begin{cases}\displaystyle\int_0^{t_1}\frac{G(dy)}{(1-G(y))^2(1-F_1(y))} & \text{if } t_1>t_2\\ \displaystyle\int_0^{t}\frac{G(dy)}{(1-G(y))^2(1-F(y,y))} & \text{if } t_1=t_2=t\\ \displaystyle\int_0^{t_2}\frac{G(dy)}{(1-G(y))^2(1-F_2(y))} & \text{if } t_1<t_2\end{cases}.$$

Since $F(y,y)\le F_1(y)$ and $F(y,y)\le F_2(y)$ we have that $\sigma^2_{WW}(t_1,t_2)\le\sigma^2_{LY}(t_1,t_2)$ (see also (3.5b) and (3.6b) in Wang and Wells (1997)). Also note that Remark 7 is valid for $\sigma^2_{WW}(t_1,t_2)$ since, in case of no censoring, $G\equiv 1$.

3.6 One-component censoring: survival function estimator of Stute

A simplification of the general bivariate setting of Sect. 3.1 is the situation where the component T_1 is fully observed and the component T_2 is subject to right censoring by C. Compare to a regression-like context where the response T_2 is censored and the covariate T_1 is fully observed. So in this model we observe a random sample $(T_{1i}, Y_{2i}, \delta_{2i})$, $i = 1, \ldots, n$, from (T_1, Y_2, δ_2) where $Y_2 = T_2 \wedge C$ and $\delta_2 = I(Y_2 \leq C)$.

In this section we discuss the estimator $\widehat{S}_S(t_1, t_2)$ for the survival function $S(t_1, t_2)$ introduced by Stute (1993a, 1995, 1996) studied the more general context of Kaplan-Meier integrals, i.e. estimation of $\int \varphi(t_1, t_2) F(dt_1, dt_2)$ by $\int \varphi(t_1, t_2)\widehat{F}(dt_1, dt_2)$ for some functions φ and with $\widehat{F}$ an appropriate estimator for F. The condition of independence between (T_1, T_2) and C is now replaced by the following pair of assumptions:

(i) T_2 and C are independent
(ii) $P(T_2 \leq C \mid T_1, T_2) = P(T_2 \leq C \mid T_2)$.

Note that independence of (T_1, T_2) and C implies (i) and (ii) and that the present weaker assumptions allow for dependence between T_1 and C. For a discussion on (ii) we refer to Stute (1996), p. 462, and to Pruitt (1993).

For simplicity we also assume that the distribution functions of T_1, T_2 and C are continuous.

Conditions (i) and (ii) are sufficient for identifiability of the survival function of (T_1, T_2). Indeed, denote

$$\widetilde{H}^{11}(t_1, t_2) = P(T_1 \leq t_1, Y_2 \leq t_2, \delta_2 = 1).$$

Then

$$\begin{aligned}
\widetilde{H}^{11}(t_1, t_2) &= E[I(T_1 \leq t_1, T_2 \leq t_2, T_2 \leq C)] \\
&= E[E[I(T_1 \leq t_1)I(T_2 \leq t_2)I(T_2 \leq C) \mid T_1, T_2]] \\
&= E[I(T_1 \leq t_1, T_2 \leq t_2)E[I(T_2 \leq C) \mid T_2] \\
&= E[I(T_1 \leq t_1, T_2 \leq t_2)(1 - G(T_2-))] \\
&= \int_0^{t_1}\int_0^{t_2} (1 - G(y_2-)F(dy_1, dy_2).
\end{aligned}$$

Hence,

$$F(dy_1, dy_2) = \frac{1}{1 - G(y_2-)} \widetilde{H}^{11}(dy_1, dy_2)$$

or, since G is continuous,

$$S(t_1, t_2) = \int_{t_1}^{\infty}\int_{t_2}^{\infty} \frac{1}{1 - G(y_2)} \widetilde{H}^{11}(dy_1, dy_2). \tag{15}$$

The corresponding estimator for $S(t_1, t_2)$ is

$$\widehat{S}_S(t_1, t_2) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_{2i}}{1-\widehat{G}(Y_{2i}-)} I(T_{1i} > t_1, Y_{2i} > t_2).$$

Considering the results of Stute (1996) for the particular choice $\varphi(x, w) = I_{]t_1,\infty[\times]t_2,\infty[}(x, w)$ and calculating the quantities $\gamma_0, \gamma_1^{\varphi}, \gamma_2^{\varphi}$ in Stute (1996), p. 464 (see the Supplementary Material for details), we obtain the asymptotic representation in Theorem 9 below.

The following integrability assumptions are also required (see (1.3) and (1.4) in Stute 1996).

(iii) $\int_{t_1}^{\infty}\int_{t_2}^{\infty} \frac{1}{1-G(w)} F(dx, dw) < \infty$

(iv) $\int_{t_1}^{\infty}\int_{t_2}^{\infty} \left(\int_0^w \frac{H_2^0(dy)}{(1-H_2(y))^2}\right)^{1/2} F(dx, dw) < \infty$

where $H_2^0(t) = P(Y_2 \le t, \delta_2 = 0)$, $H_2(t) = P(Y_2 \le t)$.

Theorem 9 *Assume conditions (i)–(iv).*

Assume that T_1, T_2, C have continuous distributions.

Then,

$$\widehat{S}_S(t_1, t_2) - S(t_1, t_2) = \frac{1}{n}\sum_{i=1}^{n} \psi_S(t_1, t_2, T_{1i}, Y_{2i}, \delta_{2i}) + o_p(n^{-1/2})$$

where

$$\begin{aligned}\psi_S(t_1, t_2, T_1, Y_2, \delta_2) = &\frac{1}{1-G(Y_2)} I(T_1 > t_1, Y_2 > t_2, \delta_2 = 1) - S(t_1, t_2)\\ &+ \frac{1}{1-H_2(Y_2)} \int_{t_1}^{\infty}\int_{t_2}^{\infty} \frac{I(Y_2 \le w, \delta_2 = 0)}{1-G(w)} \widetilde{H}^{11}(dx, dw)\\ &- \int_{t_1}^{\infty}\int_{t_2}^{\infty}\int_0^{Y_2\wedge w} \frac{H_2^0(dv)}{(1-H_2(v))^2} \frac{1}{1-G(w)} \widetilde{H}^{11}(dx, dw).\end{aligned}$$

Corollary 5 *Assume the conditions of Theorem 9. Then,*

$$n^{1/2}(\widehat{S}_S(t_1, t_2) - S(t_1, t_2)) \xrightarrow{d} N(0; \sigma_S^2(t_1, t_2))$$

where

$$\sigma_S^2(t_1, t_2) = \int_{t_1}^{\infty}\int_{t_2}^{\infty} \frac{1}{1-G(w)} F(dx, dw) - S^2(t_1, t_2)$$

$$
-\int_{t_1}^{\infty}\int_{t_2}^{\infty}\int_{t_1}^{\infty}\int_{t_2}^{\infty}\left(\int_0^{w\wedge w'}\frac{H_2^0(dy)}{(1-H_2(y))^2}\right)F(dx,dw)F(dx',dw').
$$

Proof For the calculation of the asymptotic variance, it is useful to note that ψ_S can also be written as

$$
\begin{aligned}
\psi_S(t_1,t_2,T_1,Y_2,\delta_2) &= \frac{1}{1-G(Y_2)}I(T_1>t_1,Y_2>t_2,\delta_2=1)-S(t_1,t_2)\\
&\quad+\int_{t_1}^{\infty}\int_{t_2}^{\infty}\frac{1}{(1-G(w))^2}\xi_G(w;Y_2,\delta_2)\widetilde{H}^{11}(dx,dw) \qquad (16)
\end{aligned}
$$

where ξ_G is the expression in the asymptotic representation for $\widehat{G}(w)-G(w)$, see Theorem 3.

The variance of the first two terms in (16) is equal to

$$
\begin{aligned}
&\int_{t_1}^{\infty}\int_{t_2}^{\infty}\frac{1}{(1-G(w))^2}\widetilde{H}^{11}(dx,dw)-S^2(t_1,t_2)\\
&=\int_{t_1}^{\infty}\int_{t_2}^{\infty}\frac{1}{1-G(w)}F(dx,dw)-S^2(t_1,t_2).
\end{aligned}
$$

The variance of the third term in (16) equals

$$
\begin{aligned}
&\int_{t_1}^{\infty}\int_{t_2}^{\infty}\int_{t_1}^{\infty}\int_{t_2}^{\infty}E[\xi_G(w;Y_2,\delta)\xi_G(w';Y_2,\delta)]\frac{1}{(1-G(w))^2(1-G(w'))^2}\\
&\quad\widetilde{H}^{11}(dx,dw)\widetilde{H}^{11}(dx',dw')\\
&=\int_{t_1}^{\infty}\int_{t_2}^{\infty}\int_{t_1}^{\infty}\int_{t_2}^{\infty}\left(\int_0^{w\wedge w'}\frac{H_2^0(dy)}{(1-H_2(y))^2}\right)F(dx,dw)F(dx',dw')
\end{aligned}
$$

using the covariance formula (5).

Finally the covariance is equal to

$$
\begin{aligned}
&E\left\{\frac{I(T_1>t_1,Y_2>t_2,\delta_2=1)}{1-G(Y_2)}\int_{t_1}^{\infty}\int_{t_2}^{\infty}\frac{1}{(1-G(w))^2}\xi_G(w;Y_2,\delta_2)\widetilde{H}^{11}(dx,dw)\right\}\\
&=-\int_{t_1}^{\infty}\int_{t_2}^{\infty}E\left\{\frac{I(T_1>t_1,Y_2>t_2,\delta_2=1)}{1-G(Y_2)}\int_0^{Y_2\wedge w}\frac{H_2^0(dv)}{(1-H_2(v))^2}\right\}\frac{1}{(1-G(w))}\widetilde{H}^{11}(dx,dw)\\
&=-\int_{t_1}^{\infty}\int_{t_2}^{\infty}\int_{t_1}^{\infty}\int_{t_2}^{\infty}\left(\int_0^{w\wedge w'}\frac{H_2^0(dv)}{(1-H_2(v))^2}\right)\frac{1}{1-G(w)}\frac{1}{1-G(w')}\widetilde{H}^{11}(dx,dw)\widetilde{H}^{11}(dx',dw')
\end{aligned}
$$

$$= -\int_{t_1}^{\infty}\int_{t_2}^{\infty}\int_{t_1}^{\infty}\int_{t_2}^{\infty}\left(\int_0^{w\wedge w'} \frac{H_2^0(dv)}{(1-H_2(v))^2}\right) F(dx, dw)F(dx', dw').$$

Collecting the terms gives the desired results.

Remark 9 In case of no censoring

$$\sigma_S^2(t_1, t_2) = S(t_1, t_2)(1 - S(t_1, t_2)).$$

3.7 One-component censoring: survival function estimator of Lin-Ying

The idea which led to the estimator of Lin and Ying (1993) discussed in Sect. 3.5 can also be used in the case of one-component censoring. It leads to a new estimator for the bivariate survival function.

If (T_1, T_2) and C are independent and if T_1, T_2 and C have continuous distributions, then

$$\begin{aligned} P(T_1 > t_1, Y_2 > t_2) &= P(T_1 > t_1, T_2 > t_2, C > t_2) \\ &= S(t_1, t_2)(1 - G(t_2)) \end{aligned}$$

or

$$S(t_1, t_2) = \frac{P(T_1 > t_1, Y_2 > t_2)}{1 - G(t_2)}.$$

A simple estimator is given by

$$\widetilde{S}_{LY}(t_1, t_2) = \frac{\frac{1}{n}\sum_{i=1}^n I(T_{1i} > t_1, Y_{2i} > t_2)}{1 - \widehat{G}(t_2)}$$

with $\widehat{G}$ the Kaplan–Meier estimator of G.

Remark 10 Given the way we write $S(t_1, t_2)$ it is natural to assume that (T_1, T_2) and C are independent. Note that this condition implies conditions (i) and (ii) in Sect. 3.6.

Theorem 10 *Assume that (T_1, T_2) and C are independent and that the distribution functions of T_1, T_2 and C are continuous. Then, for $t_2 < \tau_G$,*

$$\widetilde{S}_{LY}(t_1, t_2) - S(t_1, t_2) = \frac{1}{n}\sum_{i=1}^n \widetilde{\psi}_{LY}(t_1, t_2, T_{1i}, Y_{2i}, \delta_{2i}) + o_P(n^{-1/2})$$

where

$$\widetilde{\psi}_{LY}(t_1, t_2, T_1, Y_2, \delta_2) = \frac{1}{1 - G(t_2)}\{I(T_1 > t_1, Y_2 > t_2) - P(T_1 > t_1, Y_2 > t_2)\}$$

$$+S(t_1,t_2)\left\{\int_0^{t_2}\frac{I(Y_2\le y)-H_2(y)}{(1-H_2(y))^2}H_2^0(dy)\right.$$
$$+\frac{I(Y_2\le t_2,\delta_2=0)-H_2^0(t_2)}{1-H_2(t_2)}$$
$$\left.-\int_0^{t_2}\frac{I(Y_2\le y,\delta_2=0)-H_2^0(y)}{(1-H_2(y))^2}H_2(dy)\right\}$$

with $H_2(t)=P(Y_2\le t)$ *and* $H_2^0(t)=P(Y_2\le t,\delta_2=0)$.

Proof Similarly as in the proof of Theorem 8 it follows by linearization of $\widetilde{S}_{LY}-S$ that $\widetilde{S}_{LY}(t_1,t_2)-S(t_1,t_2)$ has the same asymptotic distribution as

$$\frac{1}{1-G(t_2)}\frac{1}{n}\sum_{i=1}^n\{I(T_{1i}>t_1,Y_{2i}>t_2)-P(T_1>t_1,Y_2>t_2\}$$
$$+S(t_1,t_2)\frac{1}{1-G(t_2)}[\widehat{G}(t_2)-G(t_2)].$$

Now replace $\widehat{G}(t_2)-G(t_2)$ by its asymptotic representation.

Corollary 6 *Assume the conditions of Theorem 10. Then, for* $t_2<\tau_G$,

$$n^{1/2}(\widetilde{S}_{LY}(t_1,t_2)-S(t_1,t_2))\xrightarrow{d}N(0;\widetilde{\sigma}^2_{LY}(t_1,t_2))$$

where

$$\widetilde{\sigma}^2_{LY}(t_1,t_2)=\frac{1}{(1-G(t_2))^2}P(T_1>t_1,Y_2>t_2)(1-P(T_1>t_1,Y_2>t_2))$$
$$-S^2(t_1,t_2)\int_0^{t_2}\frac{H_2^0(dy)}{(1-H_2(y))^2}.$$

Proof

$$\widetilde{\sigma}^2_{LY}(t_1,t_2)=\frac{1}{(1-G(t_2))^2}P(T_1>t_1,Y_2>t_2)(1-P(T_1>t_1,Y_2>t_2))$$
$$+S^2(t_1,t_2)\int_0^{t_2}\frac{H_2^0(dy)}{(1-H_2(y))^2}$$
$$+2\frac{S(t_1,t_2)}{1-G(t_2)}E\left\{I(T_1>t_1,Y_2>t_2)\left[\int_0^{t_2}\frac{I(Y_2\le y)-H_2(y)}{(1-H_2(y))^2}H_2^0(dy)\right.\right.$$

$$+\ \frac{I(Y_2 \le t_2, \delta_2 = 0) - H_2^0(t_2)}{1 - H_2(t_2)} - \int_0^{t_2} \frac{I(Y_2 \le y, \delta_2 = 0) - H_2^0(y)}{(1 - H_2(y))^2} H_2(dy) \Bigg]\Bigg\}.$$

The last term equals

$$-2\frac{S(t_1, t_2)}{1 - G(t_2)} P(T_1 > t_1, Y_2 > t_2) \Bigg\{ \int_0^{t_2} \frac{H_2(y)}{(1 - H_2(y))^2} H_2^0(dy)$$

$$+ \frac{H_2^0(t)}{1 - H_2(t_2)} - \int_0^{t_2} \frac{H_2^0(y)}{(1 - H_2(y))^2} H_2(dy) \Bigg\}$$

$$= -2\frac{S(t_1, t_2)}{1 - G(t_2)} P(T_1 > t_1, Y_2 > t_2) \int_0^{t_2} \frac{H_2^0(dy)}{(1 - H_2(y))^2}$$

$$= -2S^2(t_1, t_2) \int_0^{t_2} \frac{H_2^0(dy)}{(1 - H_2(y))^2}$$

where we used a similar calculation as in Theorem 7.

Remark 11 In case of no censoring

$$\widetilde{\sigma}_{LY}^2(t_1, t_2) = S(t_1, t_2)(1 - S(t_1, t_2)).$$

Remark 12 For $t_1 = 0$, we have $S(0, t_2) = P(T_2 > t_2)$ and $\widetilde{\sigma}_{LY}^2(0, t_2) = (P(T_2 > t_2))^2 \int_0^{t_2} \frac{H_2^1(dy)}{(1-H_2(y))^2}$ where $H_2^1(t) = P(Y_2 \le t, \delta_2 = 1)$ which is the asymptotic variance of the Kaplan–Meier estimator for the survival function of T_2.

3.8 One-component censoring: survival function estimator of Akritas

We again consider the bivariate model where T_1 is fully observed and T_2 is subject to censoring by C with distribution function G. Also, $Y_2 = T_2 \wedge C$, $\delta_2 = I(T_2 \le C)$ and the observations are $(T_{1i}, Y_{2i}, \delta_{2i})$, $i = 1, \ldots, n$.

The following estimator for $S(t_1, t_2)$ has been proposed by Akritas (1994) and further studied in Akritas and Van Keilegom (2003).

It is assumed that, given T_1, the variables T_2 and C are independent. The starting point is the following relation

$$S(t_1, t_2) = \int_{t_1}^{\infty} S(t_2 \mid t) F_1(dt) \tag{17}$$

where $S(t_2 \mid t) = P(T_2 > t_2 \mid T_1 = t)$.

The estimator is obtained by plugging in estimators $\widehat{S}_n(t_2 \mid t)$ for $S(t_2 \mid t)$ and $F_{1n}(t)$ for $F_1(t)$, where $F_{1n}(t) = n^{-1} \sum_{i=1}^n I(T_{1i} \leq t)$. This gives

$$\widehat{S}_A(t_1, t_2) = \frac{1}{n} \sum\nolimits_{i=1}^n \widehat{S}_n(t_2 \mid T_{1i}) I(T_{1i} > t_1).$$

For $\widehat{S}_n(t_2 \mid t)$ we use the Beran (1981) estimator (see Remark 3:)

$$\widehat{S}_n(t_2 \mid t) = \prod_{\substack{i=1 \\ Y_{2i} \leq t_2, \delta_{2i}=1}}^n \left(1 - \frac{w_{ni}(t, h_n)}{\sum_{j=1}^n w_{nj}(t, h_n) I(Y_{2j} \geq Y_{2i})}\right).$$

The weights $w_{ni}(t, h_n)$ are Nadaraya-Watson weights with

$$w_{ni}(t, h_n) = K\left(\frac{t - T_{1i}}{h_n}\right) \Big/ \sum\nolimits_{j=1}^n K\left(\frac{t - T_{1j}}{h_n}\right),$$

where K is a known probability density function and $\{h_n\}$ is a sequence of nonnegative constants, tending to 0 as $n \to \infty$. It has been shown (see Van Keilegom and Veraverbeke 1997) that $\widehat{S}_n(t_2 \mid T_{1i}) - S(t_2 \mid T_{1i})$ has the same asymptotic distribution as $-\sum_{j=1}^n w_{nj}(T_{1i}, h_n) \xi_A(t_2; Y_{2j}, \delta_{2j}, T_{1i})$ where

$$\xi_A(t_2; Y_2, \delta_2, t) = S(t_2 \mid t) \left(- \int_0^{Y_2 \wedge t_2} \frac{H_2^1(ds \mid t)}{(1 - H_2(s \mid t))^2} + \frac{I(Y_2 \leq t_2, \delta_2 = 1)}{1 - H_2(Y_2 \mid t)}\right).$$

Here $H_2(y \mid t) = P(Y_2 \leq y \mid T_1 = t)$ and $H_2^1(y \mid t) = P(Y_2 \leq y, \delta_2 = 1 \mid T_1 = t)$.

Due to the censoring of T_2, it will only be possible to estimate $S(t_1, t_2)$ in a certain domain for (t_1, t_2). Indeed, the estimator for $S(t_1, t_2)$ is obtained from relation (17) by plugging in the empirical distribution function for $F_1(t)$ and the conditional Kaplan–Meier estimator for $S(t_2 \mid t)$. To achieve uniformity of the remainder term in the asymptotic representation, we have to stay strictly away from the right endpoint of support of F_1 as well as from the right endpoint of support of $P(Y_2 \leq y \mid T_1 = t)$, for all $t \geq t_1$, the range of the integral in (17).

Hence, in order to define the domain of our estimator, we introduce the following notation (as in Akritas 1994; Akritas and Van Keilegom 2003):

τ_1 = any number strictly less than $\inf\{t : F_1(t) = 1\}$
$\tau_2(t)$ = any number strictly less than $\inf\{y : H_2(y \mid t) = 1\}$

Therefore we use the following domain

$$\Omega_A = \{(t_1, t_2) : t_1 \leq \tau_1, t_2 \leq \inf_{t \geq t_1} \tau_2(t)\}.$$

We will also need the following assumptions (see Akritas and Van Keilegom 2003):

(A1) $\frac{\log n}{nh_n} \to 0, nh^4 \to 0$;
K is a probability density with support $[-1, 1]$, K is twice continuously differentiable, $\int uK(u)du = 0$;
(A2) $F_1(t_1)$ is three times continuously differentiable w.r.t. t_1;
$H_2(t_2 \mid t_1)$ and $H_2^1(t_2 \mid t_1)$ are twice continuously differentiable w.r.t. t_1 and t_2 and for $(t_1, t_2) \in \Omega_A$, all derivatives are uniformly bounded.

Theorem 11 (Akritas 1994; Akritas and Van Keilegom 2003)
Assume that T_2 and C are independent, given T_1. Assume (A1) and (A2). Then, for $(t_1, t_2) \in \Omega_A$, we have the following representation

$$\widehat{S}_A(t_1, t_2) - S(t_1, t_2) = \frac{1}{n}\sum_{i=1}^{n} \psi_A(t_2; Y_{2i}, \delta_{2i}, T_{1i}) + r_n(t_1, t_2)$$

where

$$\psi_A(t_2; Y_2, \delta_2, T_1) = \{S(t_2 \mid T_1)I(T_1 > t_1) - S(t_1, t_2)\} - \xi_A(t_2; Y_2, \delta_2, T_1)I(T_1 > t_1)$$

and

$$\sup_{(t_1,t_2)\in\Omega_A} \mid r_n(t_1, t_2) \mid = o_p(n^{-1/2}).$$

Remark 13 The crucial part of the Akritas estimator is the Beran estimator $\widehat{S}_n(t_2 \mid t)$. It is therefore natural to assume that T_2 and C are conditional independent given T_1. Note that this assumption is not implied or does not imply the independence conditions in Sects. 3.6 and 3.7.

Remark 14 Theorem 11 is a special case of Van Keilegom (2004) in which T_1 is allowed to be censored.

Corollary 7 *Assume the conditions of Theorem 11. Then,*

$$n^{1/2}(\widehat{S}_A(t_1, t_2) - S(t_1, t_2)) \xrightarrow{d} N(0; \sigma_A^2(t_1, t_2))$$

where

$$\begin{aligned}\sigma_A^2(t_1, t_2) &= E[S^2(t_2 \mid T_1)I(T_1 > t_1)] - S^2(t_1, t_2) \\ &\quad + E\left[S^2(t_2 \mid T_1)\left(\int_0^{t_2} \frac{H_2^1(ds \mid T_1)}{(1 - H_2(s \mid T_1))^2}\right) I(T_1 > t_1)\right] \\ &= \int_{t_1}^{\infty} S^2(t_2 \mid t)F_1(dt) - S^2(t_1, t_2)\end{aligned}$$

$$+\int_{t_1}^{\infty} S^2(t_2 \mid t)\left(\int_0^{t_2} \frac{H_2^1(ds \mid t)}{(1-H_2(s \mid t))^2}\right) F_1(dt).$$

Remark 15 In case of no censoring $\int_0^{t_2} \frac{H_2^1(ds|t)}{(1-H_2(s|t))^2} = \frac{F(t_2|t)}{1-F(t_2|t)}$ and hence $\sigma_A^2(t_1, t_2) = S(t_1, t_2)(1 - S(t_1, t_2))$.

4 Applications

Representations are a particularly useful tool to study asymptotic properties of complicated statistical estimators. In this section we demonstrate, for right censored data, how the i.i.d. representations for nonparametric univariate and bivariate survival function estimators have been used as building blocks in the derivation of asymptotic properties of more complicated estimators. Given the large amount of possible applications, we limit ourselves to four concrete examples that have recently been discussed in the statistical literature: nonparametric conditional residual quantile estimation, nonparametric copula estimation, cure models (in survival analysis and banking) and goodness-of-fit in regression models.

4.1 Conditional residual quantiles

The Lo and Singh (1986) representation (Theorem 3 in this paper) has been used to obtain i.i.d. representations for quantiles of the Kaplan–Meier estimator $\widehat{S}(t)$ (Gijbels and Veraverbeke 1988) and also for quantiles of the conditional Kaplan-Meier estimator $\widehat{S}(t \mid x)$ in Remark 3 (Van Keilegom and Veraverbeke 1998).

More recent work is the study of conditional residual quantiles. For a lifetime T_1 and some other variable T_2, containing extra information on T_1, conditional residual lifetime distributions are defined as $P(T_1 - t_1 \leq y \mid T_1 > t_1, T_2 \leq t_2)$ or $P(T_1 - t_1 \leq y \mid T_1 > t_1, T_2 > t_2)$ or $P(T_1 - t_1 \leq y \mid T_1 > t_1, t_{21} < T_2 \leq t_{22})$.

Abrams et al. (2021, 2023) studied asymptotic representations for nonparametric estimators of the quantiles of these distributions. The proposed estimators use the one-component Akritas-Van Keilegom estimator of Sect. 3.8 and the univariate censoring estimators of Sects. 3.4 and 3.5.

The i.i.d. representations in Sect. 3 are key ingredients to study the asymptotic properties of the conditional residual quantile estimators.

4.2 Copulas

Survival copulas can be written as $\mathcal{C}(u_1, u_2) = S(S_1^{-1}(u_1), S_2^{-1}(u_2))$ with S the joint survival function and S_1 and S_2 the marginal survival functions. Using nonparametric estimators S_n, S_{1n} and S_{2n} for S, S_1 and S_2, a nonparametric estimator $\mathcal{C}_n(u_1, u_2)$ for

$\mathcal{C}(u_1, u_2)$ is given by

$$\mathcal{C}_n(u_1, u_2) = S_n(S_{1n}^{-1}(u_1), S_{2n}^{-1}(u_2)).$$

Using the nonparametric estimators for S_n discussed in Sect. 3 and nonparametric Kaplan–Meier based estimators for the marginal quantiles, we obtain estimators for copula functions, which can be studied based on the asymptotic representations given in Sect. 3. See Geerdens et al. (2016) for details. In that paper there is also a comparison with an alternative estimator of Gribkova and Lopez (2015).

4.3 Cure models

There are many contexts (e.g. cancer research) in which subjects in the study never experience the event of interest (e.g. death caused by the cancer). They are called 'cured'.

Several models have been introduced and studied to modify the classical survival analysis in presence of a cured fraction. An up-to-date review paper is Amico and Van Keilegom (2018). In Geerdens et al. (2020) a goodness-of-fit test for a parametric survival function with cure fraction is discussed for the mixture cure model $S(t) = 1-\phi+\phi S_1(t)$ with $1-\phi$ the cure fraction and $S_1(t)$ the survival function of the uncured subjects (the susceptibles). With $\widehat{S}_1(t)$ the Maller and Zhou (1996) estimator for $S_1(t)$ and $\widehat{\theta}$ the maximum likelihood estimator for θ, the Cramér-von Mises distance

$$\wedge_n = \sum_{i=1}^{n} (\widehat{S}_1(Y_i) - S_{1,\widehat{\theta}}(Y_i))^2 \tag{18}$$

with $Y_i = T_i \wedge C_i$, is used to test

$$H_0 : S_1 \in \{S_{1,\theta} : \theta \in \Theta\} \text{ versus } H_a : S_1 \notin \{S_{1,\theta} : \theta \in \Theta\}$$

where Θ is the parameter space of the parameter θ in the assumed parametric form $S_{1,\theta}(t)$ of the survival function $S_1(t)$.

An example of an application of censoring and cure models outside the clinical research but in the domain of finances and banking appeared in the recent PhD thesis of Peláez-Suárez (2022). She uses the conditional cure model

$$S(t \mid x) = 1 - \phi(x) + \phi(x) S_1(t \mid x)$$

with T the time to default (unable to pay the debts incurred by granting a credit) and X a credit score variable. To estimate the default probability

$$P(T \leq t + b \mid T > t, X = x) = \frac{S(t + b \mid x)}{S(t \mid x)} \tag{19}$$

she uses a nonparametric cure model estimator of the conditional survival function $S(\cdot \mid x)$. The latter estimator, in terms of Beran-type estimators (Beran 1981) for the incidence $\phi(x)$ and the latency $S_1(t \mid x)$, is studied in López-Cheda et al. (2017a, 2017b).

To study the asymptotic properties of the goodness-of-fit statistics $\wedge_n$ in (18) or the estimated default probability (19) again i.i.d. representations are essential.

4.4 Goodness-of-fit in regression models

There is also a huge literature on regression models with censored data in which the response T is subject to random right censoring. We mention the two recent papers: González-Manteiga et al. (2020) and Conde-Amboage et al. (2021) and the references therein. Examples are the mean regression model $T = m(X) + \varepsilon$ where $m(X)$ is the conditional mean of T, given X, or the quantile regression model $T = g_\tau(X) + \varepsilon$ where $g_\tau(X)$ is the conditional τ-quantile function of T, given X $(0 < \tau < 1)$. There exist many goodness-of-fit procedures to test the hypothesis that $m(\cdot)$ or $g_\tau(\cdot)$ belong to some class of parametric functions. As discussed for cure models, goodness-of-fit statistics are based on a comparison of a model based parametric estimator and a nonparametric estimator for $m(\cdot)$, resp. $g_\tau(\cdot)$ and, again, the role of i.i.d. representations is crucial to study the asymptotic properties of the goodness-of-fit statistics.

The above examples clearly show the need of i.i.d. representations to study asymptotics in more complicated censoring models. Indeed, also the study of asymptotic properties of nonparametric estimators of the univariate or bivariate survival functions for data subject to left truncation and right censoring or interval censored data will rely on i.i.d. representations. Moreover such representations are and will be highly needed to study more complex data schemes, e.g. censored data in competing risks models and models dealing with dependent censoring.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00184-023-00911-7.

Acknowledgements The authors thank the editor and three referees for their valuable comments that led to an improved version of the manuscript.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Abrams S, Janssen P, Veraverbeke N (2021) Quantiles of the conditional residual lifetime. Statistics 55:1271–1290

Abrams S, Janssen P, Veraverbeke N (2023) Nonparametric estimation of the quantiles of the conditional lifetime distribution. Technical Report (submitted)

Akritas MG (1994) Nearest neighbor estimation of a bivariate distribution under random censoring. Ann Statist 22:1299–1327

Akritas MG, Van Keilegom I (2003) Estimation of bivariate and marginal distributions with censored data. J Royal Statist Soc B 65:457–471

Amico M, Van Keilegom I (2018) Cure models in survival analysis. Annu Rev Statist Appl 5:311–342

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

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

Breslow N, Crowley J (1974) A large sample study of the life table and product limit estimates under random censorship. Ann Statist 2:437–453

Burke M (1988) Estimation of a bivariate distribution under random censorship. Biometrika 75:379–382

Conde-Amboage M, Van Keilegom I, González-Manteiga W (2021) A new lack-of-fit test for quantile regression with censored data. Scand J Statistics 48:665–688

Dabrowksa DM (1988) Kaplan-Meier estimate on the plane. Ann Statist 16:1475–1489

Ebrahimi N, Molefe D, Ying Z (2003) Identifiability and censored data. Biometrika 90:724–727

Földes A, Rejtő L (1981) A LIL type result for the product limit estimator. Z Wahrsch Verw Gebiete 56:75–86

Geerdens C, Janssen P, Van Keilegom I (2020) Goodness-of-fit test for a parametric survival function with cure fraction. TEST 29:768–792

Geerdens C, Janssen P, Veraverbeke N (2016) Large sample properties of nonparametric copula estimators under bivariate censoring. Statistics 50:1036–1055

Gijbels I, Veraverbeke N (1988) Weak asymptotic representations for quantiles of the product limit estimator. J Statist Plan Inference 18:151–160

Gill RD (1992) Multivariate survival analysis. Theory Prob Appl 37:18–31

Gill RD (1994) Glivenko-Cantelli for Kaplan-Meier. Math Methods Statist 3:76–78

González-Manteiga W, Cadarso Suarez C (1994) Asymptotic properties of a generalized Kaplan-Meier estimator with some applications. J Nonparametric Stat 4:65–78

González-Manteiga W, Heuchenne C, Sanchez-Sellero C, Beretta A (2020) Goodness-of-fit tests for censored regression based on artificial data points. TEST 29:599–615

Gribkova S, Lopez O (2015) Nonparametric copula estimation under bivariate censoring. Scand J Stat 42:925–946

Hougaard P (2000) Analysis of Multivariate Survival Data. Springer-Verlag, New York

Johansen S (1978) The product limit estimator as maximum likelihood estimator. Scand J Stat 5:195–199

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457–481

Langberg NA, Shaked M (1982) On the identifiability of multivariate life distributions. Ann Prob 10:773–779

Lin DY, Ying ZA (1993) A simple nonparametric estimator of the bivariate survival function under univariate censoring. Biometrika 80:573–581

Lo S-H, Singh K (1986) The product-limit estimator and the bootstrap: some asymptotic representations. Prob Theory Rel Fields 71:455–465

Lopez O (2012) A generalization of the Kaplan-Meier estimator for analyzing bivariate mortality under right-censoring with applications in model checking for survival copula models. Insur Math Econ 51:505–516

Lopéz-Cheda A, Cao R, Jácome MA (2017a) Nonparametric latency estimation for mixture cure models. TEST 26:353–376

Lopéz-Cheda A, Cao R, Jácome MA, Van Keilegom I (2017b) Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. Comput Stat Data Anal 105:144–165

Major P, Rejtő L (1988) Strong embedding of the estimator of the distribution function under random censorship. Ann Stat 16:1113–1132

Maller R, Zhou S (1996) Survival analysis with long term survivors. Wiley, London

Nelsen RB (2006) An introduction to copulas, 2nd edn. Springer, New York

Peláez-Suárez, R. (2022). Nonparametric estimation of the probability of default in credit risk. Doctoral Thesis, Universidade da Coruña, Spain

Prentice RL, Cai J (1992) Covariance and survivor function estimation using censored multivariate failure time data. Biometrika 79:495–512

Prentice RL, Zhao S (2018) Nonparametric estimation of the multivariate survivor function: the multivariate Kaplan-Meier estimator. Lifetime Data An 24:3–27

Pruitt R (1991) On negative mass assigned by the bivariate Kaplan-Meier estimator. Ann Stat 19:443–453
Pruitt R (1993) Identifiability of bivariate survival curves from censored data. J Am Stat Assoc 88:573–579
Rivest L, Wells MT (2001) A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. J Multiv Anal 79:138–155
Robins JM, Rotnitzky A (1992) Recovery of information and adjustment for dependent censoring using surrogate markers. In: Jewell N, Dietz K Jr, Farewell VT (eds) AIDS epidemiology methodological issues. Birkhauser, Boston, pp 297–331
Satten GA, Datta S (2001) The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average. Am Stat 55:207–210
Serfling R (1980) Approximations theorems of mathematical statistics. Wiley, New York
Stute W (1993a) Consistent estimation under random censorship when covariables are present. J Multiv Anal 45:89–103
Stute W, Wang J-L (1993b) The strong law under random censorship. Ann Stat 21:1591–1607
Stute W (1995) The central limit theorem under random censorship. Ann Stat 23:422–439
Stute W (1996) Distributional convergence under random censorship when covariables are present. Scand J Stat 23:461–471
Tsiatis AA (1975) A nonidentifiability aspect of the problem of competing risks. Proc Nat Acad Sci United States of America 72:20–22
Van Keilegom I, Veraverbeke N (1997) Estimation and bootstrap with censored data in fixed design non-paramteric regression. Ann Inst Statist Math 49:467–491
Van Keilegom I, Veraverbeke N (1998) Bootstrapping quantiles in a fixed design regression model with censored data. J Stat Plan Inference 69:115–131
Van Keilegom I (2004) A note on the nonparametric estimation of the bivariate distribution under dependent censoring. J Nonparametric Stat 16:659–670
van der Laan MJ (1996) Efficient estimation in the bivariate censoring model and repairing NPMLE. Ann Stat 24:596–627
Wang W, Wells MT (1997) Nonparametric estimators of the bivariate survival function under simplified censoring conditions. Biometrika 84:863–880
Zheng M, Klein JP (1995) Estimates of the marginal survival for dependent competing risks based on an assumed copula. Biometrika 82:127–138