Biometrika (2019), **106**, 3, *pp.* 724–731
Printed in Great Britain

doi:10.1093/biomet/asz023
Advance Access publication 13 May 2019

Nonidentifiability in the presence of factorization for truncated data

By B. VAKULENKO-LAGUN

Department of Biostatistics, Harvard T. H. Chan School of Public Health, 655 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.
blagun@hsph.harvard.edu

J. QIAN

Department of Biostatistics and Epidemiology, University of Massachusetts, 715 N. Pleasant Street, Amherst, Massachusetts 01003, U.S.A.
qian@schoolph.umass.edu

S. H. CHIOU and R. A. BETENSKY

Department of Biostatistics, Harvard T. H. Chan School of Public Health, 655 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.
schiou@hsph.harvard.edu betensky@hsph.harvard.edu

Summary

A time to event, X, is left-truncated by T if X can be observed only if $T < X$. This often results in oversampling of large values of X, and necessitates adjustment of estimation procedures to avoid bias. Simple risk-set adjustments can be made to standard risk-set-based estimators to accommodate left truncation when T and X are quasi-independent. We derive a weaker factorization condition for the conditional distribution of T given X in the observable region that permits risk-set adjustment for estimation of the distribution of X, but not of the distribution of T. Quasi-independence results when the analogous factorization condition for X given T holds also, in which case the distributions of X and T are easily estimated. While we can test for factorization, if the test does not reject, we cannot identify which factorization condition holds, or whether quasi-independence holds. Hence we require an unverifiable assumption in order to estimate the distribution of X or T based on truncated data. This contrasts with the common understanding that truncation is different from censoring in requiring no unverifiable assumptions for estimation. We illustrate these concepts through a simulation of left-truncated and right-censored data.

Some key words: Constant-sum condition; Kendall's tau; Left truncation; Right censoring; Survival data.

1. Introduction

Truncated survival data arise when observation of the time to event, X, occurs only when it falls within a subject-specific interval. Left truncation occurs when X is observed only if $T < X$, where T is the time to sampling, i.e., the truncation variable. It often arises in longitudinal cohort studies in which a subcohort is sampled on the basis of having had a post-baseline assessment prior to the event of interest. Another example is when the time origin of interest, such as onset of cognitive impairment, may occur prior to entry into the cohort, and the endpoint of interest is time from onset of cognitive impairment to death. Estimation must account for the truncation to avoid bias due to the selection based on the magnitude of X. A critical condition that enables simple risk-set adjustment to standard risk-set-based estimators

(Lynden-Bell, 1971; Woodroofe, 1985; Wang et al., 1986) was described by Tsai (1990) as quasi-independence, or independence in the observed region, i.e., $T < X$. In particular, this means that $H(x,t) = \int_0^t \int_0^x \mathrm{d}F(u)\,\mathrm{d}G(v)I(v < u)/\alpha$, where $H(x,t) = \mathrm{pr}(X \leqslant x, T \leqslant t \mid T < X)$, $F(x) = \mathrm{pr}(X \leqslant x)$, $G(t) = \mathrm{pr}(T \leqslant t)$, $\alpha = \mathrm{pr}(T < X)$, and $I(A) = 1$ if the event A holds and 0 otherwise. For simplicity of presentation, we assume that X and T are continuous random variables and that H, F and G have densities h,f and g with respect to Lebesgue measure. Nonetheless, all of our results apply also to discrete random variables, as they are based on nonparametric maximum likelihoods (Vardi, 1989). The quasi-independence assumption expressed in terms of densities is

$$h(x,t) = f(x)g(t)I(t < x)/\alpha. \tag{1}$$

This does not imply that the sampled random variables are conditionally independent given $T < X$; it is not equivalent to $H(\mathrm{d}x, \mathrm{d}t) = \mathrm{pr}(X \in \mathrm{d}x \mid T < X) \times \mathrm{pr}(T \in \mathrm{d}t \mid T < X)$ for $t < x$, where we use the notation $\mathrm{pr}(X \in \mathrm{d}x)$ as a shorthand for $\mathrm{pr}(x \leqslant X < x + \mathrm{d}x)$.

Examination of the likelihood based on left-truncated data elucidates the simplification in estimation that arises from quasi-independence, and reveals that weaker conditions also admit this simplification for estimation of the distribution of X or T, but not of both. The factorization condition that enables estimation of the distribution of X is

$$\mathrm{pr}(T \in \mathrm{d}t \mid X = x) = \mathrm{d}A(t) = a(t)\,\mathrm{d}t \quad (t < x), \tag{2}$$

where $a(t) \geqslant 0$ need not equal $g(t)$ and is defined on the support of T in the observable region $T < X$. In the unobservable region, we define $c(t,x)\,\mathrm{d}t = \mathrm{pr}(T \in \mathrm{d}t \mid X = x)$ for $x < t$. For (x,t) in the support of (X,T), $a(t)$ and $c(t,x)$ are constrained by

$$\int_0^\infty \mathrm{pr}(T \in \mathrm{d}t \mid X = x) = \int_0^x a(t)\,\mathrm{d}t + \int_x^\infty c(t,x)\,\mathrm{d}t = 1,$$

$$g(t) = \int_{x=t}^\infty a(t)\,\mathrm{d}F(x) + \int_{x=0}^t c(t,x)\,\mathrm{d}F(x).$$

In the Supplementary Material we derive an explicit expression for $a(t)$ as a function of $G(t)$, $F(x)$ and $c(t,x)$, under the factorization condition (2). We also discuss the special cases of overall independence and quasi-independence. The factorization condition is similar to the condition of Keiding (1992) that $h(x,t) = f(x)g^*(t)$, upon identifying $g^*(t)$ as $a(t)/\alpha$. The factorization condition is reminiscent of the constant-sum condition for right-censored data (Williams & Lagakos, 1977; Betensky, 2000), under which dependent censoring can be ignored and the Kaplan–Meier estimator is valid.

Proposition 4 shows that the distribution of T is not identifiable under (2) alone; it also requires a complementary factorization condition. The two factorization conditions together constitute quasi-independence (1), under which both distributions can be estimated. We explain in § 3 that the observed data can be used to test whether neither factorization condition holds, but cannot be used to identify which condition holds if either does. Therefore, we require unverifiable assumptions in order to estimate the distribution of X or T based on truncated data. This contrasts with the common understanding that truncation is distinct from censoring and requires no unverifiable assumptions for estimation.

2. Nonparametric likelihood estimation

2.1. *Estimation in the absence of censoring*

First we consider estimation in the absence of right censoring. The likelihood of the observed data $\{(t_i,x_i) : i = 1,\ldots,n,\ t_i < x_i\}$ under left truncation and no censoring is

$$\prod_{i=1}^n \mathrm{pr}(X \in \mathrm{d}x_i,\ T \in \mathrm{d}t_i \mid T < X) = L_1L_2L_3 \tag{3}$$

where

$$L_1 = \prod_{i=1}^{n} \frac{\mathrm{pr}(T \in \mathrm{d}t_i \mid X = x_i) I(t_i < x_i)}{\mathrm{pr}(T \in \mathrm{d}t_i \mid X > t_i)}, \quad L_2 = \prod_{i=1}^{n} \frac{\mathrm{pr}(X \in \mathrm{d}x_i) I(t_i < x_i)}{\mathrm{pr}(X > t_i)},$$

$$L_3 = \prod_{i=1}^{n} \frac{\mathrm{pr}(X > t_i)\, \mathrm{pr}(T \in \mathrm{d}t_i \mid X > t_i)}{\int_0^\infty \mathrm{pr}(X > u)\, \mathrm{pr}(T \in \mathrm{d}u \mid X > u)}.$$

PROPOSITION 1. *Under the factorization condition (2), the nonparametric maximum likelihood estimator of $S(x) = 1 - F(x)$ is the risk-set-adjusted Kaplan–Meier estimator*

$$\hat{S}(x) = \prod_{x_i \leqslant x} \left\{ 1 - \frac{\sum_{j=1}^{n} I(x_j = x_i)}{\sum_{j=1}^{n} I(t_j \leqslant x_i \leqslant x_j)} \right\}. \tag{4}$$

Proof. Under (2), L_1 is equal to 1 since its denominator is equal to its numerator, $\mathrm{d}A(t)$:

$$\begin{aligned} \mathrm{pr}(T \in \mathrm{d}t \mid X > t) &= \int_{x=t}^{\infty} \mathrm{pr}(T \in \mathrm{d}t \mid X = x, X > t)\, \mathrm{pr}(X \in \mathrm{d}x \mid X > t) \\ &= \mathrm{d}A(t) \int_{x=t}^{\infty} \mathrm{pr}(X \in \mathrm{d}x \mid X > t) = \mathrm{d}A(t). \end{aligned}$$

Thus, $L_2 L_3$ effectively constitutes the full likelihood, with unknown parameters $\mathrm{pr}(X \in \mathrm{d}x)$ and $\mathrm{pr}(T \in \mathrm{d}t \mid X > t)$. The standard risk-set-adjusted Kaplan–Meier estimator given by (4) is the maximum likelihood estimator of $S(x)$ based on L_2, and also equals that based on $L_2 L_3$ (Wang, 1991). This is because, in the absence of parametric assumptions on $\mathrm{pr}(T \in \mathrm{d}t \mid X > t)$, L_3 is a multinomial likelihood with maximum value n^{-n} if there are no ties in $t_1, \ldots, t_n$, which is attained when each factor in its product is set to the corresponding sample proportion. A similar argument holds in the presence of ties. □

Since (4) is the maximizer of L_2, if it also maximizes the full likelihood (3), then $L_1 L_3$ must be constant with respect to $S(x)$. If this latter condition implies factorization (2), then it would follow that (2) is a necessary condition for (4) to be the nonparametric maximum likelihood estimator of $S(x)$. We conjecture that this is false.

Under complete independence between T and X, (4) was shown to be uniformly consistent by Woodroofe (1985). Since the likelihoods that contribute to estimation of $S(x)$ are identical and equal to L_2 under any of the three conditions of complete independence between T and X, quasi-independence (1), or factorization (2), the uniform consistency of (4) under (1) or (2) can be proved in the same way as under complete independence between X and T.

The likelihood (3) can also be expressed as $L_1^* L_2^* L_3^*$ where

$$L_1^* = \prod_{i=1}^{n} \frac{\mathrm{pr}(X \in \mathrm{d}x_i \mid T = t_i) I(t_i < x_i)}{\mathrm{pr}(X \in \mathrm{d}x_i \mid T < x_i)}, \quad L_2^* = \prod_{i=1}^{n} \frac{\mathrm{pr}(T \in \mathrm{d}t_i) I(t_i < x_i)}{\mathrm{pr}(T < x_i)},$$

$$L_3^* = \prod_{i=1}^{n} \frac{\mathrm{pr}(T < x_i)\, \mathrm{pr}(X \in \mathrm{d}x_i \mid T < x_i)}{\int_0^\infty \mathrm{pr}(T < u)\, \mathrm{pr}(X \in \mathrm{d}u \mid T < u)}.$$

A complementary factorization condition to (2) for X given T is

$$\mathrm{pr}(X \in \mathrm{d}x \mid T = t) = \mathrm{d}A^*(x) = a^*(x)\, \mathrm{d}x \quad (t < x), \tag{5}$$

where $a^*(x) \geqslant 0$ need not equal $f(x)$. Conditions (2) and (5) together are equivalent to quasi-independence, as stated in the following proposition.

PROPOSITION 2. *Under conditions (2) and (5), $a(t) = g(t)$ and $a^*(x) = f(x)$, which implies quasi-independence (1). Conversely, quasi-independence (1) implies both (2) and (5).*

Proof. Under (2) and (5), $h(x,t) = f(x)a(t)I(t < x)/\alpha = a^*(x)g(t)I(t < x)/\alpha$, implying $a(t) = g(t)$ and $a^*(x) = f(x)$, i.e., quasi-independence. Under (1), $h(x,t) = f(x)g(t)I(t < x)/\alpha$, implying (2) and (5) with $a(t) = g(t)$ and $a^*(x) = f(x)$. □

Under (5), the likelihood for estimation of $G(t)$ effectively reduces to L_2^*, and its estimation is dual to that of $S(x)$ (Wang, 1991). This is summarized in the following proposition.

Proposition 3. *Under* (5), *the nonparametric maximum likelihood estimator of* $G(t)$ *is*

$$\hat{G}(t) = \prod_{t_i \geqslant t} \left\{ 1 - \frac{\sum_{j=1}^{n} I(t_j = t_i)}{\sum_{j=1}^{n} I(t_j \leqslant t_i \leqslant x_j)} \right\}.$$

Proof. This follows from Proposition 1 via reversal of time, by treating $-T$ as left-truncated by $-X$.□

Propositions 1–3 lead to the following corollary.

Corollary 1. *Quasi-independence yields the standard risk-set-adjusted estimators of the distributions of both* X *and* T *as the nonparametric maximum likelihood estimators.*

Assumptions (1), (2) and (5) are indistinguishable given the observed data. We formalize this conclusion in Propositions 4 and 5 and Corollary 2. Proposition 4 shows that the likelihood under (2) is equivalent to that under (1), implying that these conditions cannot be distinguished. Proposition 5 shows that (5) and (1) cannot be distinguished.

Proposition 4. *Assuming factorization* (2), *quasi-independence* (1) *cannot be determined from the observed data. As a consequence, while* $f(x)$ *is identifiable under* (2), $g(t)$ *is not.*

Proof. This follows from the equivalence of the likelihood functions under quasi-independence (1) and factorization (2). Under the latter, the likelihood (3) is

$$L_2 L_3 = \prod_{i=1}^{n} \frac{\mathrm{d}F(x_i) I(t_i < x_i)}{S(t_i)} \prod_{i=1}^{n} \frac{S(t_i)\,\mathrm{d}A(t_i)}{\int_0^\infty S(u)\,\mathrm{d}A(u)}.$$

Under quasi-independence (1), the likelihood (3) is

$$L_2 L_3 = \prod_{i=1}^{n} \frac{\mathrm{d}F(x_i) I(t_i < x_i)}{S(t_i)} \prod_{i=1}^{n} \frac{S(t_i)\,\mathrm{d}G(t_i)}{\int_0^\infty S(u)\,\mathrm{d}G(u)}.$$

Since $a(t)$ is defined only on the observable region, it is unique only up to a constant factor. Assuming that $A(t)$ and $G(t)$ have positive mass at the observed times $t_1, \ldots, t_n$ only, but not assuming their functional forms, the contributions to the likelihood from L_3 are identical under (1) and (2). Hence $G(t)$ is nonidentifiable from the data. □

Proposition 5. *Assuming factorization* (5), *quasi-independence* (1) *cannot be determined from the observed data. Thus, while* $g(t)$ *is identifiable under* (5), $f(x)$ *is not.*

Corollary 2. *Quasi-independence* (1) *cannot be distinguished from the factorization condition* (2) *only, or from the factorization condition* (5) *only, based on the observed data.*

2.2. *Estimation under right censoring*

The nonidentifiability problem persists in the presence of right censoring. There are two practical models for right censoring in the presence of left truncation (Qian & Betensky, 2014): one is on the residual time scale, i.e., censoring of $X - T$, and the other is on the original time scale, i.e., censoring of X. We extend the likelihood decomposition (3) to accommodate these models.

We first consider the independent residual censoring assumption. Suppose that D is a residual censoring time such that $D \perp\!\!\!\perp (T, X) \mid T < X$, where $\perp\!\!\!\perp$ denotes independence, and that censoring of X occurs at $C = T + D$, the total censoring time starting from the time origin. The observed data then comprise $Y = \min(X, C)$, T and δ, where $\delta = 1$ if $T < X \leqslant C$ and $\delta = 0$ if $T < C < X$. This model is appropriate when censoring occurs only after entry into the study. The likelihood contribution for an uncensored observation is the same as that in (3):

$$\begin{aligned}
&\mathrm{pr}(Y \in \mathrm{d}y,\ \delta = 1,\ T \in \mathrm{d}t \mid T < X)\\
&\quad = \mathrm{pr}(X \in \mathrm{d}y,\ T + D > y,\ T \in \mathrm{d}t \mid T < X)\\
&\quad = \mathrm{pr}(X \in \mathrm{d}y,\ T \in \mathrm{d}t \mid D > y - t,\ T < X)\,\mathrm{pr}(D > y - t \mid T < X)\\
&\quad \propto \mathrm{pr}(X \in \mathrm{d}y,\ T \in \mathrm{d}t \mid T < X),
\end{aligned}$$

where the final relation follows from $D \perp\!\!\!\perp (T, X) \mid T < X$ and the noninformativeness of the distribution of D for that of X. The contribution for a censored observation is

$$\begin{aligned}
&\mathrm{pr}(Y \in \mathrm{d}y,\ \delta = 0,\ T \in \mathrm{d}t \mid T < X)\\
&\quad = \mathrm{pr}(X > y,\ T + D \in \mathrm{d}y,\ T \in \mathrm{d}t \mid T < X)\\
&\quad = \mathrm{pr}(X > y,\ T \in \mathrm{d}t \mid D = y - t,\ T < X)\,\mathrm{pr}(D \in \mathrm{d}(y - t) \mid T < X)\\
&\quad \propto \mathrm{pr}(X > y,\ T \in \mathrm{d}t \mid T < X).
\end{aligned}$$

The probability $\mathrm{pr}(X > y, T \in \mathrm{d}t \mid T < X)$ can be expressed as

$$\frac{\mathrm{pr}(T \in \mathrm{d}t \mid X > y)}{\mathrm{pr}(T \in \mathrm{d}t \mid X > t)} \times \frac{\mathrm{pr}(X > y)}{\mathrm{pr}(X > t)} \times \frac{\mathrm{pr}(X > t)\,\mathrm{pr}(T \in \mathrm{d}t \mid X > t)}{\int_0^\infty \mathrm{pr}(X > u)\,\mathrm{pr}(T \in \mathrm{d}u \mid X > u)}\, I(t < y),$$

where the first term is unity under the factorization condition (2).

We next derive the likelihood decomposition under the censoring scheme on the original time scale, where C is measured from the time origin, with $C \perp\!\!\!\perp X \mid T$ assumed, and $\mathrm{pr}(T < C) = 1$ as in Tsai (1990). The condition $\mathrm{pr}(T < C) = 1$ ensures that censoring can occur only for the sampled individuals. The overall likelihood under this censoring scheme equals that under the residual censoring model, given the assumed noninformativeness of C given T for X and that of D given $T < X$. Thus, under both models for censoring, the overall likelihood for left-truncated and right-censored data is

$$\prod_{i=1}^{n} \mathrm{pr}(Y \in \mathrm{d}y_i,\ \delta = \delta_i,\ T \in \mathrm{d}t_i \mid T < X) \propto \tilde{L}_1 \tilde{L}_2 \tilde{L}_3$$

where

$$\tilde{L}_1 = \prod_{i=1}^{n} \frac{\mathrm{pr}(T \in \mathrm{d}t_i \mid X = y_i)^{\delta_i}\,\mathrm{pr}(T \in \mathrm{d}t_i \mid X > y_i)^{1-\delta_i} I(t_i < y_i)}{\mathrm{pr}(T \in \mathrm{d}t_i \mid X > t_i)},$$

$$\tilde{L}_2 = \prod_{i=1}^{n} \frac{\mathrm{pr}(X \in \mathrm{d}y_i)^{\delta_i}\,\mathrm{pr}(X > y_i)^{1-\delta_i} I(t_i < y_i)}{\mathrm{pr}(X > t_i)},$$

$$\tilde{L}_3 = \prod_{i=1}^{n} \frac{\mathrm{pr}(X > t_i)\,\mathrm{pr}(T \in \mathrm{d}t_i \mid X > t_i)}{\int_0^\infty \mathrm{pr}(X > u)\,\mathrm{pr}(T \in \mathrm{d}u \mid X > u)}.$$

Under (2), $\tilde{L}_1 = 1$. As in the uncensored case, $\tilde{L}_2$ is the only component of the likelihood that contributes to estimation of $S(x)$ by the risk-set-adjusted Kaplan–Meier estimator (Wang, 1991)

$$\hat{S}(x) = \prod_{y_i \leqslant x} \left\{ 1 - \frac{\sum_{j=1}^{n} I(y_j = y_i)\delta_i}{\sum_{j=1}^{n} I(t_j \leqslant y_i \leqslant y_j)} \right\}. \tag{6}$$

As in Proposition 4, under (2) the data cannot inform whether $\text{pr}(T \in \text{d}t \mid X > t)$ equals $\text{d}G(t)$ or $\text{d}A(t)$. Nonetheless, the nonparametric maximum likelihood estimator of $\text{pr}(T \leqslant t \mid X > t)$, assuming that it is a distribution function, is

$$\left\{ \sum_{j=1}^{n} \frac{1}{\hat{S}(t_j)} \right\}^{-1} \sum_{i=1}^{n} \frac{1}{\hat{S}(t_i)} I(t_i \leqslant t). \tag{7}$$

In the setting of independent X and (T, C) with $\text{pr}(T < C) = 1$, (7) estimates $G(t)$ (Wang, 1991). Under factorization (2) without quasi-independence, (7) maximizes the likelihood $\tilde{L}_3$ given $\hat{S}(x)$ and estimates a normalized version of $A(t)$ and not $G(t)$. Under factorization (5) without quasi-independence, an alternative decomposition of the likelihood with similar arguments yields the analogous result for estimation of $G(t)$ and $A^*(x)$, as shown in the Supplementary Material.

3. Testing for the factorization condition

A statistic commonly used to test for quasi-independence is the conditional Kendall's tau (Tsai, 1990; Martin & Betensky, 2005). In the presence of censoring, this is defined as $\tau_c = E[\text{sgn}\{(Y_i - Y_j)(T_i - T_j)\} \mid \Lambda_{ij}]$, where $\text{sgn}(a) = I(a > 0) - I(a < 0)$ and $\Lambda_{ij} = \{\max(T_i, T_j) \leqslant \min(Y_i, Y_j)\} \cap \{\delta_i \,\text{sgn}(Y_j - Y_i) = 1 \cup \delta_j \,\text{sgn}(Y_i - Y_j) = 1\}$ denotes the event that the pair (i,j) is comparable and orderable. A consistent estimator of τ_c is the basis of a test for the null hypothesis of (2) or (5), versus the alternative of neither (2) nor (5). This is justified by $\tau_c = 0$ under (2) or (5). We derive this for (2); the calculations are similar under (5):

$$\begin{aligned} \text{pr}(\Lambda_{ij})\tau_c &= E[\text{sgn}\{(Y_1 - Y_2)(T_1 - T_2)\}I(\Lambda_{12})] \\ &= \text{pr}(\delta_1 = 1, T_1 < T_2 < X_1 < Y_2) - \text{pr}(\delta_2 = 1, T_1 < T_2 < X_2 < Y_1) \\ &\quad + \text{pr}(\delta_2 = 1, T_2 < T_1 < X_2 < Y_1) - \text{pr}(\delta_1 = 1, T_2 < T_1 < X_1 < Y_2). \end{aligned} \tag{8}$$

Under the residual censoring model and factorization condition (2), and upon defining $S_D(u) = \text{pr}(D > u)$ as the survival function of D, we can express $\text{pr}(\delta_1 = 1, T_1 < T_2 < X_1 < Y_2)$ as

$$\begin{aligned} &\alpha^2 \int_{t=0}^{\infty} \int_{u=t}^{\infty} \text{pr}(T_2 \in \text{d}t, T_1 < t, X_1 \in \text{d}u, C_1 > u, X_2 > u, C_2 > u \mid T_1 < X_1, T_2 < X_2) \\ &= \int_{t=0}^{\infty} \int_{u=t}^{\infty} \left\{ \int_{s=0}^{t} S_D(u - s)\, \text{d}A(s) \right\} S(u) S_D(u - t)\, \text{d}F(u)\, \text{d}A(t). \end{aligned}$$

The second term of (8) can be expressed similarly, and the remaining two terms are trivially equivalent to the first two terms upon relabelling the indices. A similar result applies under the original-scale censoring model. This demonstrates that $\tau_c = 0$ under either factorization condition, and the conditional Kendall's tau provides a valid test for the null of either (2) or (5). If the test does not reject the null hypothesis, then under (2), Proposition 4 states that $a(t)$ cannot be distinguished from $g(t)$ in the observable region, and so quasi-independence cannot be distinguished from factorization. This holds for any test of factorization in the absence of external information.

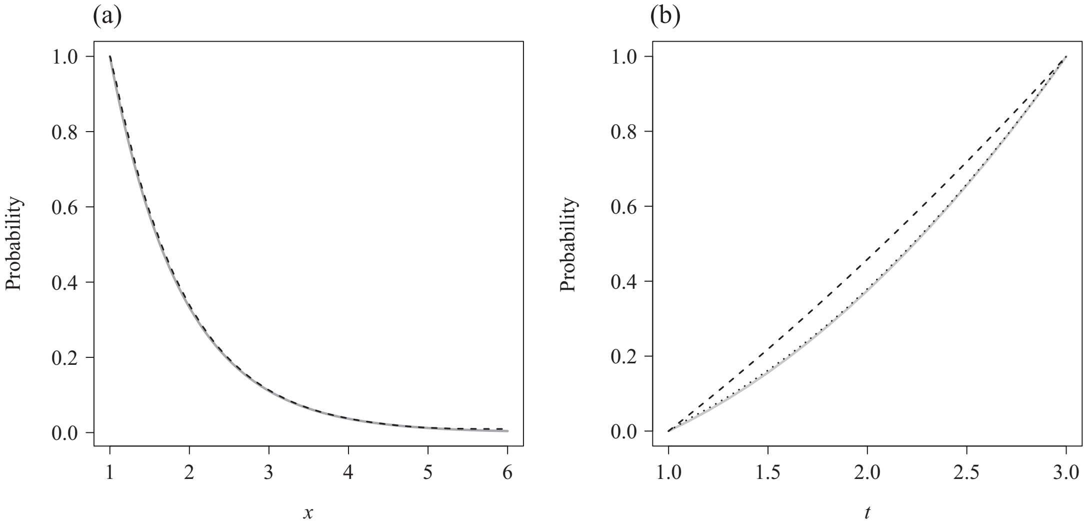

Fig. 1. Simulation results for the estimation of: (a) $\text{pr}(X > x \mid X > 1)$, with a grey solid line depicting the true curve and a black dashed line depicting the average of Kaplan–Meier estimates; (b) $\text{pr}(T \leqslant t)$, with a grey solid line depicting the true $A(t)$, a black dotted line depicting the average of the estimates of Wang (1991), and a black dashed line depicting $G(t) = \text{pr}(T \leqslant t)$.

4. Simulation

We conducted a simulation study to illustrate empirically that the factorization condition (2) alone, without quasi-independence (1), is sufficient for the validity of the risk-set-adjusted Kaplan–Meier estimator for the distribution of X, as stated in Proposition 1. We also demonstrate that Kendall's tau yields a valid test of the factorization condition even in the absence of quasi-independence. Finally, we illustrate Proposition 4, that under (2) without the complementary condition (5), we may not be able to estimate the truncation distribution $G(t)$. Let

$$\text{pr}(T \in \mathrm{d}t \mid X = x) = \begin{cases} 0.5\, I(1 \leqslant t \leqslant 3)\, \mathrm{d}t, & x < 1, \\ 0.25t\, I(1 \leqslant t \leqslant 3)\, \mathrm{d}t, & x \geqslant 1, \end{cases}$$

where $X \sim \text{Ex}(\theta)$ and we set $\theta = 1.1$. It follows that $g(t) = \{0.5 - 0.5\exp(-\theta) + 0.25t\exp(-\theta)\}I(1 \leqslant t \leqslant 3)$, $a(t) = 0.25t\, I(1 \leqslant t \leqslant 3)$, and $c(t,x) = 0.5$ if $x < 1$ and $c(t,x) = 0.25t$ if $1 \leqslant x < t \leqslant 3$. Since $a(t) \neq g(t)$, quasi-independence does not hold. We generated right censoring through an independent residual censoring time $D \sim \text{Un}[0,3]$. Each sample consisted of $n = 200$ triples $\{\min(X, T + D),\ T,\ \delta = I(X \leqslant T + D)\} \mid T < X$. This yielded 88% truncation and 30% censoring based on 1000 replications.

Our first aim is to check the validity of the risk-set-adjusted Kaplan–Meier estimator of the conditional distribution, $\text{pr}(X > x \mid X > 1)$. The full marginal $\text{pr}(X > x)$ is not estimable because there is no information for $X < 1$. Figure 1(a) displays the averaged adjusted Kaplan–Meier estimate, which is indistinguishable from its target, confirming that the adjusted Kaplan–Meier estimator is valid under condition (2) and does not require the stronger condition (1). We also applied the conditional Kendall's tau test of Martin & Betensky (2005) and obtained an estimated Type I error of 0.041, which supports the validity of the test for either factorization condition even in the absence of quasi-independence. Figure 1(b) shows that the estimator (7) estimates $A(t)$ and not $G(t)$, as expected from Proposition 4.

5. Discussion

We have shown that the commonly accepted requirement of quasi-independence of T and X is stronger than the factorization condition (2) that is actually needed for nonparametric estimation of the distribution of X. While we can test for factorization, the observed data do not allow us to distinguish between quasi-independence (1) and the two factorization conditions (2) and (5). This highlights an identification problem that has not been recognized in the literature; an unverifiable assumption is therefore required in order to estimate the distribution of X based on truncated data. In some observational studies, the origin may be observed for all subjects and the delayed study entry time may be externally determined, such as by calendar date. In this case, T is known for the whole population and so $G(t)$ is known. If factorization is not rejected via Kendall's tau test, knowledge of $G(t)$ enables the factorization condition (2) to be distinguished from quasi-independence (1) and the factorization condition (5). In particular, if factorization holds and $\hat{A}(t)$ does not estimate $G(t)$, it follows from Proposition 3 that condition (5) does not hold, which means that (2) must hold and, importantly, we can estimate $F(x)$.

Acknowledgement

The first two authors contributed equally. We thank Micha Mandel and Richard Cook for helpful comments. We acknowledge funding from the U.S. National Institutes of Health.

Supplementary material

Supplementary material available at *Biometrika* online includes the derivation of an explicit expression for $a(t)$ given $G(t)$, $F(x)$ and $c(t,x)$ under factorization condition (2); it also contains a proof that under (5) and for both censoring models, although $S(x)$ is nonidentifiable, $G(t)$ is identifiable and its nonparametric maximum likelihood estimator is similar to the estimator (7).

References

Betensky, R. A. (2000). On nonidentifiability and noninformative censoring for current status data. *Biometrika* **87**, 218–21.

Keiding, N. (1992). Independent delayed entry. In *Survival Analysis: State of the Art*, J. P. Klein & P. K. Goel, eds. Dordrecht: Springer, pp. 309–26.

Lynden-Bell, D. (1971). A method of allowing for known observational selection in small samples applied to 3CR quasars. *Mon. Not. R. Astron. Soc.* **155**, 95–118.

Martin, E. C. & Betensky, R. A. (2005). Testing quasi-independence of failure and truncation times via conditional Kendall's tau. *J. Am. Statist. Assoc.* **100**, 484–92.

Qian, J. & Betensky, R. A. (2014). Assumptions regarding right censoring in the presence of left truncation. *Statist. Prob. Lett.* **87**, 12–7.

Tsai, W.-Y. (1990). Testing the assumption of independence of truncation time and failure time. *Biometrika* **77**, 169–77.

Vardi, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation. *Biometrika* **76**, 751–61.

Wang, M.-C. (1991). Nonparametric estimation from cross-sectional data. *J. Am. Statist. Assoc.* **86**, 130–43.

Wang, M.-C., Jewell, N. P. & Tsai, W.-Y. (1986). Asymptotic properties of the product limit estimate under random truncation. *Ann. Statist.* **14**, 1597–605.

Williams, J. S. & Lagakos, S. W. (1977). Models for censored survival analysis: Constant-sum and variable-sum models. *Biometrika* **64**, 215–24.

Woodroofe, M. (1985). Estimating a distribution function with truncated data. *Ann. Statist.* **13**, 163–77.

[*Received on* 30 *August* 2017. *Editorial decision on* 17 *December* 2018]