Biometrika (2019), **xx**, x, *pp*. 1–3
Printed in Great Britain

Supplementary material for '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

APPENDIX A

Under factorization condition (2), we can derive an explicit expression for $a(t)$ as a function of $G(t)$, $F(x)$ and $c(t,x)$. For this, we represent $\mathrm{pr}(T \in \mathrm{d}t)$ as

$$\begin{aligned}\mathrm{pr}(T \in \mathrm{d}t) = \mathrm{d}G(t) &= \mathrm{pr}(T \in \mathrm{d}t \mid X > t)\,\mathrm{pr}(X > t) + \mathrm{pr}(T \in \mathrm{d}t \mid X < t)\,\mathrm{pr}(X < t)\\ &= \mathrm{d}A(t)\,\mathrm{pr}(X > t) + \mathrm{pr}(T \in \mathrm{d}t \mid X < t)\,\mathrm{pr}(X < t),\end{aligned}$$

which implies that

$$\mathrm{d}A(t) = \left\{\mathrm{d}G(t) - \int_{x=0}^{t} \mathrm{pr}(T \in \mathrm{d}t \mid X = x)\,\mathrm{d}F(x)\right\} \Big/ \mathrm{pr}(X > t),$$

i.e., $a(t) = g(t)b(t)$, where

$$b(t) = \left\{1 - g(t)^{-1} \int_{0}^{t} c(t,x)\,\mathrm{d}F(x)\right\} \Big/ \mathrm{pr}(X > t)$$

for $g(t) > 0$ and $b(t) = 0$ for $g(t) = 0$. When quasi-independence (1) does not hold, $b(t) \neq 1$.

Overall independence and quasi-independence between T and X arise under certain restrictions on these functions. For example, under factorization condition (2), the functions $a(t)$, $b(t)$ and $c(t,x)$ satisfy the following:

(i) Under overall independence of T and X, i.e., independence in both the observable and the unobservable regions, $a(t) = g(t)$ and $c(t,x) = g(t)$, and thus $b(t) = 1$.

(ii) Under quasi-independence (1), $a(t) = g(t)$ and hence $b(t) = 1$, but $c(t, x)$ is not necessarily equal to $g(t)$. In fact, $c(t, x)$ could be a function of both t and x. One example of this is given by selecting $f(x)$, then determining $g(x) \geqslant 0$ through

$$\frac{g(x)}{1 - G(x)} = \frac{xf'(x)}{xf(x) - F(x)} \tag{A1}$$

where $f'(x)$ is the derivative of $f(x)$, and finally setting $c(t, x) = g(t)F(t)/\{tf(x)\}$ for $t > x$. It is straightforward to verify that this choice of $g(x)$ given $f(x)$ satisfies both constraints listed in § 1 of the main paper after (2), as well as $a(t) = g(t)$. Note that this choice of $g(x)$ requires that $xf(x) > F(x)$ for all x, which in turn requires that the support of f be bounded, or else $f(x) \geqslant 1/(2x)$ for all sufficiently large x, which violates integrability. In addition, (A1) implies that

$$\frac{\mathrm{d}}{\mathrm{d}x}\big[-\log\{1 - G(x)\}\big] = \frac{\mathrm{d}}{\mathrm{d}x}\big[\log\{xf(x) - F(x)\}\big],$$

so that for some positive constant C, $1 - G(x) = C/\{xf(x) - F(x)\}$. This implies that f must be bounded from below; otherwise, as x goes to 0, the left-hand side would approach 1 and the right-hand side ∞. Thus, in this example X has to have a bounded support, $[b, B]$, with $b > 0$ and $B < \infty$. This is plausible in many contexts. In addition, it must be that $C/\{xf(x) - F(x)\} \leqslant 1$, and it should be a nonincreasing function. This implies that $xf'(x) \geqslant 0$. If X is a nonnegative random variable, $f(x)$ should be a nondecreasing density, which is possible if the support of X is bounded.

APPENDIX B

Under complete independence between T and X, the Kaplan–Meier estimator (4), the nonparametric maximum likelihood estimator for $S(x)$, was shown to be uniformly consistent for $S(x)$ by Woodroofe (1985) for left-truncated data, and by Andersen et al. (1993, Theorem IV.3.1) for left-truncated and right-censored data. The likelihoods that contribute to estimation of $S(x)$ are identical and equal to L_2 under any of three conditions: complete independence between T and X, quasi-independence (1), or factorization (2). Hence, the estimator (4) is the nonparametric maximum likelihood estimator of $S(x)$ under (1) or (2) as well, and its uniform consistency can be proved in the same way as under complete independence between X and T.

Moreover, the Kaplan–Meier estimator (4) is consistent under any of three factorization conditions, but it is a consistent estimator for different parameters. Under conditions (1) or (2), it is a consistent estimator of $S(x)$; and under condition (5) without quasi-independence, it is a consistent estimator of $1 - A^*(x)$. Nonidentifiability arises because we cannot know what we estimate, $S(x)$ or $1 - A^*(x)$.

Here we show that under factorization condition (5) without quasi-independence, for both censoring models, although $S(x)$ is nonidentifiable, $G(t)$ is identifiable. Under both models for censoring, the overall likelihood for left-truncated and right-censored data can be expressed as

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

where

$$\tilde{L}_1^* = 1, \quad \tilde{L}_2^* = \prod_{i=1}^{n} \frac{\mathrm{pr}(X \in \mathrm{d}y_i \mid T = t_i)^{\delta_i}\, \mathrm{pr}(X > y_i \mid T = t_i)^{1-\delta_i} I(t_i < y_i)}{\mathrm{pr}(X > t_i \mid T = t_i)},$$

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

Under (5), $\tilde{L}_2^* \tilde{L}_3^*$ depends on $\mathrm{pr}(X = x \mid T = t)$ and $G(t)$, whereas under (1) $\tilde{L}_2^* \tilde{L}_3^*$ depends on $S(x)$ and $G(t)$. Since in nonparametric estimation the likelihood $\tilde{L}_2^* \tilde{L}_3^*$ is the same under (1) and under (5), we cannot distinguish the parameter $S(x)$ from $A^*(x) = \mathrm{pr}(X = x \mid T = t)$, but $G(t)$ is identifiable.

Under (5), the estimation of two parameters, $\mathrm{pr}(X = x \mid T = t)$ and $G(t)$, can be done in two steps. First, the nonparametric maximum likelihood estimator of

$$\mathrm{pr}(X > t \mid T = t) = \int_t^\infty a^*(x)\, \mathrm{d}x = 1 - A^*(t)$$

is the estimator (6) in the main paper, the standard Kaplan–Meier estimator that accounts for delayed entry and right censoring. It can be shown that the estimator (6) maximizes $\tilde{L}_2^*$. Second, the nonparametric maximum likelihood estimator of $G(t)$ is the estimator (7) in the main paper, with the nonparametric maximum likelihood estimator of $\mathrm{pr}(X > t \mid T = t)$ found in the first step plugged in for $\hat{S}(t)$. This nonparametric maximum likelihood estimator of $G(t)$ is derived from maximization of $\tilde{L}_3^*$, which has a multinomial structure.

We note that under (5) and right censoring, we do not need the estimator of $S(t)$ in order to estimate $G(t)$. But we need the estimator of $\mathrm{pr}(X > t \mid T = t)$ whatever it estimates. Under (5) without quasi-independence, $\mathrm{pr}(X > t \mid T = t) = \int_t^\infty a^*(x)\, \mathrm{d}x$. Under quasi-independence, $\mathrm{pr}(X > t \mid T = t) = S(t)$.

We also remark that under left truncation, there exists another standard identifiability problem, where instead of $S(x)$ we can only identify the conditional survival function $S(x)/S(t_1)$ for $x > t_1$, where $t_1 = \min\{t_1, \ldots, t_n\}$ (Andersen et al., 1993, p. 264).

References

ANDERSEN, P., BORGAN, O., GILL, R. D. & KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer.

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