Statistical inference for right-censored data with nonignorable missing censoring indicators

Zhihua Sun † ①, Tianfa Xie ②, Hua Liang ③

①School of Mathematical sciences, Graduate University, Chinese Academy of Sciences, Beijing 100049
E-mail: sunzh@amss.ac.cn
③College of Applied Sciences, Beijing University of Technology, Beijing 100124
E-mail: xietf@bjut.edu.cn
②Dept of Biostatistics and Computational Biology, University of Rochester Rochester, New York 14642
E-mail: hliang@bst.rochester.edu

Abstract We consider the statistical inference for right-censored data when censoring indicators are missing but nonignorable, and propose an adjusted imputation product-limit estimator. The proposed estimator is shown to be consistent and converges to a Gaussian process. Furthermore, we develop an empirical process-based testing method to check the MAR (missing at random) mechanism, and establish asymptotic properties for the proposed test statistic. To determine the critical value of the test, a consistent model-based bootstrap method is suggested. We conduct simulation studies to evaluate the numerical performance of the proposed method and compare it with existing methods. We also analyze a real data set from a breast cancer study for an illustration.

Keywords: MAR mechanism testing Nonignorable missing censoring indicators Survival function Quasi-likelihood

1 Introduction

For right-censored data, the censoring indicator may be missing. In this paper, we study the estimation of the survival function for right-censored data with nonignorable missing censoring indicators. The nonignorable missing mechanism means that the missing process of the censoring indicators depends on the underlying values of the censoring indicators.

Let T and C be the survival time and the censoring time variables with the distribution functions F and G, respectively. Assume that T and C are independent. One observes (X, δ, ξ) with $X = \min(T, C)$ and $\delta = I_{(T \leqslant C)}$, where I_A is the indicator function of the set A, and ξ denotes the missing indicator with $\xi = 1$ when δ is observed and 0 otherwise. We assume that $P(\xi = 1|X, \delta)$ follows a parametric model, denoted by $\pi(X, \delta, \alpha)$ with $\pi(\cdot, \cdot, \alpha)$ being a known function and α an unknown parameter. The probability models such as the logistic model, can be applied. Like [2], [3]and [4] , we assume a parametric model $m(x, \theta)$ for $E(\delta|X = x)$. The logistic model and the model, $m(x, \theta) = x^{\theta_2}/(\theta_1 + x^{\theta_2})$ with θ_1 and θ_2 being unknown parameters, are frequently used.

2 Estimation

2.1 Estimation of the unknown parameters

Let $(X_i, \delta_i, \xi_i)_{i=1}^n$ be a sample from (X, δ, ξ). We first estimate the unknown parameters θ and α in $m(X, \theta)$ and $\pi(X, \delta, \alpha)$. Note that $E(\xi\delta|X) = m(X, \theta)\pi(X, 1, \alpha)$ and $Var(\xi\delta|X) = E(\xi\delta|X)\{1 - E(\xi\delta|X)\}$. Then the quasi-likelihood estimating equation method can be used to estimate the unknown parameters θ and α. We construct a quasi-likelihood estimating equation as follows:

$$\sum_{i=1}^{n} \left\{ \begin{array}{c} \frac{\dot{m}(X_i,\theta)}{m(X_i,\theta)} \\ \\ \\ \frac{\dot{\pi}(X_i,1,\alpha)}{\pi(X_i,1,\alpha)} \end{array} \right\} \frac{\xi_i\delta_i - m(X_i, \theta)\pi(X_i, 1, \alpha)}{1 - m(X_i, \theta)\pi(X_i, 1, \alpha)} = 0, \qquad (2.1)$$

where $\dot{m}(x, \theta)$ and $\dot{\pi}(x, 1, \alpha)$ respectively denote the derivatives of $m(x, \theta)$ and $\pi(x, 1, \alpha)$ with respect to θ and α. We denote the solution of (2.1) by $\hat{\theta}_n$ and $\hat{\alpha}_n$. Let $H(\cdot)$ be the distribution of X, $H^1(x) = \int_0^x m(u, \theta)H(du)$, and τ be a constant satisfying $H(\tau) < 1$.

Theorem 1 *Under Conditions (C1)-(C3) in the Appendix, we have*

$$\sqrt{n}\begin{pmatrix} \hat{\theta}_n - \theta \\ \\ \hat{\alpha}_n - \alpha \end{pmatrix} \xrightarrow{L} N(0, V^{-1}).$$

The details of V^{-1} *ca be referred to the full paper.*

2.2 Estimation of the survival function

The imputation method is inappropriate in the presence of nonignorable missingness.We consider the imputation estimator suggested by [6, 7] :

$$S_n(t) = \prod_{i:X_i \leqslant t} \left(\frac{n - R_i}{n - R_i + 1} \right)^{\xi_i\delta_i + (1-\xi_i)\hat{m}_n(X_i)}, \qquad (2.2)$$

where $\hat{m}_n(X)$ is a nonparametric estimator of $m(X) = E(\delta|X)$ or other appropriate value, and R_i is the rank of X_i for $i = 1, 2, \cdots, n$. It can be verified that the estimator (2.2) is inconsistent under the nonignorable missing mechanism. Note that $0 < m(x, \theta) < 1$. By $E(\xi|X) = E(\xi|X, \delta = 1)E(\delta|X) + E(\xi|X, \delta = 0)\{1 - E(\delta|X)\}$, we have

$$\begin{aligned} E\{\xi\delta + (1 - \xi)m(X)|X\} &= E[m(X)\{1 - m(X)\}\{E(\xi|X, \delta = 1) - E(\xi|X, \delta = 0)\}] + m(X) \\ &\neq m(X) \qquad (2.3) \end{aligned}$$

under the nonignorable mechanism. The last equation holds if and only if $E(\xi|X, \delta = 1) = E(\xi|X, \delta = 0)$, which is just the MAR assumption. This fact can be considered as a hint of the inconsistency of $S_n(t)$.

However, we can observe from (2.3) that

$$E\left[\xi\delta + (1 - \xi)m(X) - m(X)\{E(\xi|X, \delta = 1) - E(\xi|X)\}|X\right] = m(X). \qquad (2.4)$$

This equality inspires us to propose an adjusted imputation estimator:

$$\hat{S}_n(t) = \prod_{i:X_i \leqslant t} \left(\frac{n - R_i}{n - R_i + 1} \right)^{[\xi_i\delta_i + (1-\xi_i)m(X_i,\hat{\theta}_n) - m(X_i,\hat{\theta}_n)\{\pi(X_i,1,\hat{\alpha}_n) - \hat{U}_n(X_i)\}]} \qquad (2.5)$$

where $\hat{U}_n(x)$ is a local kernel estimator of $E(\xi|X = x)$; that is, $\hat{U}_n(x) = \frac{\sum_{i=1}^n \xi_i K_b(x - X_i)}{\sum_{i=1}^n K_b(x - X_i)}$ with $K(\cdot)$ being a kernel function and b being a bandwidth sequence. Then it is easy to show that under the mild conditions, $\hat{U}_n(x)$ is a consistent estimator of $E(\xi|X = x)$, which is denoted by $U(x)$.

Remark 1 By $E(\xi\delta|X) = m(X,\theta)\pi(X,1,\alpha) \leqslant m(X,\theta)$, it is easy to see that the complete-case estimator $\hat{S}_C(t) = \prod\limits_{i:X_i \leqslant t} \left(\frac{n-R_i}{n-R_i+1}\right)^{\delta_i \xi_i}$ overestimates the true survival function.

Theorem 2 *Under Conditions (C1)-(C5), we have*

(A) Strong consistence: $\sup_{0\leqslant t \leqslant \tau} |\hat{S}_n(t) - S(t)| \xrightarrow{a.s.} 0$; *and*

(B) Asymptotic normality: $\sqrt{n}\{\hat{S}_n(t) - S(t)\} \xrightarrow{D} S(t)w(t)$, *where* $S(t)$ *is the true survival function and* $w(t)$ *denotes a centered Gaussian process with the covariance function given by*

$$\begin{aligned}\operatorname{cov}\{w(s), w(t)\} = E&\left[\frac{\{m(X,\theta) + \pi(X,1,\alpha) - m(X,\theta)\pi^2(X,1,\alpha)\}\, m(X,\theta)}{\{1-H(X)\}^2} I_{(X\leqslant t)}\right] \\ &+\Gamma_2(s)V_{11}^{[-1]}\Gamma_1(t)^\top - \Gamma_2(s)V_{11}^{[-1]}\Gamma_2(t)^\top + \Gamma_2(t)V_{12}^{[-1]}\Gamma_3(s)^\top \\ &+\Gamma_2(t)V_{21}^{[-1]}\Gamma_3(s)^\top + \Gamma_2(t)V_{11}^{[-1]}\Gamma_1(s)^\top - \Gamma_3(s)V_{21}^{[-1]}\Gamma_1(t)^\top \\ &-\Gamma_3(s)V_{22}^{[-1]}\Gamma_3(t)^\top - \Gamma_3(t)V_{21}^{[-1]}\Gamma_1(s)^\top \qquad (2.6)\end{aligned}$$

for $0 \leqslant t \leqslant s < \tau$ *where the notations can be referred to the original paper.*

3 Testing MAR assumption

The simulation results in Section 4 show that the proposed estimator $\hat{S}_n(t)$ is in the shade of the estimators of [6 , 8] under the MAR mechanism. Therefore, it is expected to test whether MAR condition holds. Since MAR condition is equal to $E(\xi\delta|X) = E(\xi|X)m(X)$, our residual-marked empirical process testing statistic is defined as:

$$\mathcal{T}_n = \int R_n^2(X) dH_n(X), \qquad (3.7)$$

where $R_n(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \{\xi_i\delta_i - \hat{U}_n(X_i)m(X_i,\hat{\theta}_n)\} I_{(X_i\leqslant x)}$ and $H_n(x)$ is the empirical distribution function of X based on sample $(X_i, i = 1, 2, \cdots, n)$. Denote $IF_x(X,\xi,\delta) = \xi\{\delta - m(X,\theta)\}\, I_{(X\leqslant x)} - E\{U(X)\dot{m}(X,\theta)I_{(X\leqslant x)}\}IF_x^\theta(X,\xi,\delta)$ with

$$IF_x^\theta(X,\xi,\delta) = |V|\frac{\xi\delta - m(X,\theta)\pi(X,1,\alpha)}{1-m(X,\theta)\pi(X,1,\alpha)}\left\{V_{11}^{[-1]}\frac{\dot{m}(X,\theta)}{m(X,\theta)} + V_{12}^{[-1]}\frac{\dot{\pi}(X,\theta)}{\pi(X,\theta)}\right\},$$

where $|V|$ denotes the determinant of V.

Theorem 3 Suppose that Conditions (C1)-(C5) in the Appendix hold. Under the MAR mechanism, $R_n(x)$ converges to a centered Gaussian process $R(x)$ with the covariance function $Cov(R(x_1), R(x_2))$ $= E\{IF_{x_1}(X,\xi,\delta)\, IF_{x_2}(X,\xi,\delta)\}$. As a consequence, $\mathcal{T}_n$ converges in distribution to $\int R^2(X)\, dH(X)$. Since the asymptotic distribution of the testing statistic is complex, we apply a model-based bootstrap technique to define the critical value of the test $\mathcal{T}_n$. Here we omit the details.

4 Simulation studies and a real data example

Under some settings, we conducted simulation studies and a real data analysis to validate the estimating and testing procedures. We compare these proposed estimator and the existing estiamtors in terms of the MISE (mean integral square error) on the interval [0, 0.6]. The results, based on 1000 replications with sample sizes 100 and 200, are reported in Table 1.

Table 1. Simulation results for Example 1: MISE×100 of the four estimators—$\hat{S}_n(t)$: the proposed estimator; $\hat{S}_{nW}(t)$: Wang and Ng's nonparametric estimator; $\hat{S}_{nS}(t)$: Subramanian's estimator; $\hat{S}_{nV}(t)$: van der Laan and McKeague's estimator and $\hat{S}_C(t)$: complete-case estimator under the different configurations.

Scenario	Censor	Missing	n	$\hat{S}_n(t)$	$\hat{S}_{nW}(t)$	$\hat{S}_{nS}(t)$	$\hat{S}_{nV}(t)$	$\hat{S}_C(t)$
(i)	A	NMAR1	100	0.189	0.405	0.484	0.801	1.971
			200	0.111	0.384	0.416	0.713	1.880
		NMAR2	100	0.192	0.161	0.205	0.819	1.982
			200	0.112	0.103	0.121	0.724	1.894
	B	NMAR1	100	0.166	0.199	0.226	0.933	1.155
			200	0.095	0.155	0.161	0.799	1.086
		NMAR2	100	0.149	0.172	0.194	0.847	0.823
			200	0.083	0.127	0.131	0.710	0.740
(ii)	A	MAR1	100	0.402	0.138	0.187	2.644	1.476
			200	0.346	0.073	0.089	2.543	1.387
		MAR2	100	0.613	0.134	0.168	2.651	0.946
			200	0.596	0.063	0.075	2.592	0.830
	B	MAR1	100	0.208	0.119	0.157	1.741	2.079
			200	0.157	0.058	0.068	1.318	1.367
		MAR2	100	0.283	0.116	0.146	1.818	1.439
			200	0.245	0.058	0.067	1.739	1.330
(iii)	A	DMP1	100	0.195	0.417	0.493	2.065	0.786
			200	0.116	0.365	0.395	1.880	0.749
		DMP2	100	0.199	0.163	0.205	2.082	0.802
			200	0.117	0.102	0.119	1.892	0.760
	B	DMP1	100	0.162	0.202	0.228	0.933	1.117
			200	0.101	0.150	0.156	0.787	1.096
		DMP2	100	0.150	0.174	0.194	0.835	0.783
			200	0.089	0.124	0.128	0.701	0.760

Table 2. Simulation results for Example 2: Empirical sizes and powers of the test under different sample sizes, missing and censoring rates.

		100		200	
	missing/ censoring	A	B	A	B
sizes	MAR1*	0.058	0.069	0.045	0.057
	MAR2*	0.071	0.075	0.036	0.031
power	NMAR1*	0.585	0.428	1.000	0.992
	NMAR2*	0.346	0.222	0.991	0.876

Example 1. To illustrate the proposed method, we investigate a data set of 169 elderly women with Stage II breast cancer. We calculate the proposed estimator $\hat{S}_n(t)$, the complete-case estimator, and the estimators of [6, 10], and depict these estimators in Figure 1. We have further applied the testing procedure in Section 3 to check whether the missing process is MAR or not. The P-value is calculated to be 0.012. Hence the MAR assumption can not be accepted and it is advisable to estimate the survival function by applying the proposed method. Other parametric probability models for $E(\delta|X)$ and $E(\xi|X, \delta)$, such as logistic models, are also applied to analyse the data and similar results are obtained.

References

[1] Kaplan E. L. and Meier P.(1958) " Nonparametric estimation from incomplete observations." *Journal of the American Statistical Association*, **53**: 457–481

[2] Lu W. and Tsiatis A. A. (2006) " Semiparametric transformation models for the case-cohort study." *Biometrika*, **93**:207–214

[3] Gao G. and Tsiatis A. A. (2005) " Semiparametric estimators for the regression coefficients in the linear transformation competing risks model with missing cause of failure. " *Biometrika*, **92**: 875–891

[4] Dikta G. (1998) " On semiparametric random censorship models." *Journal of Statistical Planning and Inference*, **66**:253–279

[5] Cheng P. E. (1994) " Nonparametric estimation of mean functionals with data missing at random." *Journal of the American Statistical Association*, **89**: 81–87

[6] Wang Q. and Ng K. W. (2008) " Asymptotically efficient product-limit estimators with censoring indicators missing at random." *Statistica Sinica*, **18**: 749–768

[7] Lo S. H. (1991) " Estimating a Survival Function with Incomplete Cause-of-death Data." *Journal of Multivariate Analysis*, **39**: 217–235

[8] Subramanian S. (2006) " Survival analysis for the missing censoring indicator model using kernel density estimation techniques." *Statistical Methodology*, **3**: 125–136

[9] Zhu L. and Ng K. W. (2003) " Checking the adequacy of a partial linear model." *Statistica Sinica*, **13**: 763–781

[10] McKeague I. W. and Subramanian, S. (1998)" Product-Limit Estimators and Cox Regression with Missing Cause-of-Failure Information." *Scandinavian Journal of Statistics*, **25**: 589–601

[11] Goetghebeur E. and Ryan L. (1995) " Analysis of competing risks survival data when some failure types are missing." *Biometrika*, **82**: 821–833

[12] Fan J. and Gijbels I. (1996) Local polynomial modelling and its applications. Chapman & Hall, London

[13] Foutz R. V. (1977) " On the unique consistent solution to the likelihood equations." *Journal of the American Statistical Association*, **72**: 147–148

[14] van der Vaart A. W. and Wellner J. A. (1996) Weak Convergence and Empirical Processes. Springer, New York

[15] Lo S. H and Singh K. (1986) " The product-limit estimator and the bootstrap: some asymptotic representations. " *Probability Theory and Related Fields*, **71**: 455–465

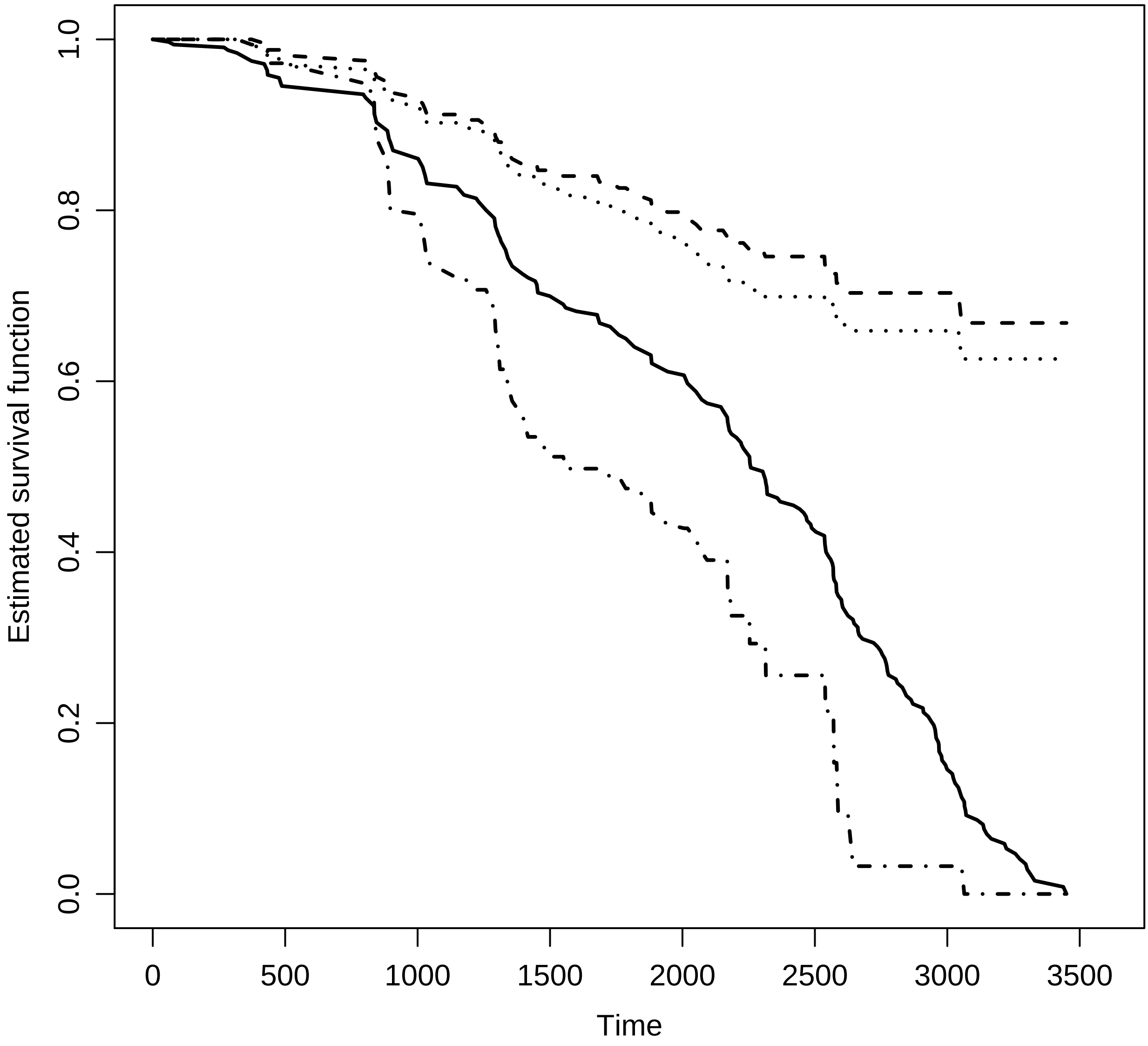

Figure 1. The estimated survival curves for Example 3. Solid line: the proposed estimator $\hat{S}_n(t)$; dotted line: Wang and Ng's nonparametric estimator; dashed line: the complete-case estimator; and dashed-dotted line: van der Laan and McKeague's estimator.