APPLIED METHODS OF STATISTICAL ANALYSIS.

STATISTICAL COMPUTATION AND SIMULATION

PROCEEDINGS

OF THE INTERNATIONAL WORKSHOP

18-20 September 2019

Novosibirsk

2019

UDC 519.22(063)
A 67

E d i t o r s:

Prof. Boris Lemeshko, Prof. Mikhail Nikulin,
Prof. Narayanaswamy Balakrishnan

A 67 **Applied Methods of Statistical Analysis. Statistical Computation and Simulation** - AMSA'2019, Novosibirsk, Russia, 18-20 September, 2019: Proceedings of the International Workshop. - Novosibirsk: NSTU publisher, 2019. - 574 pp.
ISSN 2313-870X

ISSN 2313-870X **UDC 519.22(063)**

Applied Methods of Statistical Analysis. Statistical Computation and Simulation

On estimation algorithms in nonparametric analysis of the current status right-censored data

IRINA YU. MALOVA[1,3] AND SERGEY V. MALOV[2,3]
[1] *Higher School of Technology and Energy, St. Petersburg State University of Industrial Technologies and Design, St. Petersburg, Russia*
[2] *Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg, Russia*
[3] *Department of Mathematics, St. Petersburg Electrotechnical University (LETI), St. Petersburg, Russia*
e-mail: malovs@sm14820.spb.edu

Abstract

We consider nonparametric estimation algorithms for current status right-censored data model. In the model right-censored event times are not observed exactly, but at some inspection times. The model covers right-censored data, current status data and life table survival data with a single inspection time. We consider the nonparametric estimation algorithms to obtain three nonparametric estimators for the survival function of failure time: maximum likelihood, pseudo maximum likelihood and the naïve estimator. We discuss large sample properties of the estimators. Using the standard R packages we perform simulations, which compare the estimators under small and moderate sample sizes.
Keywords: survival data, right censoring, interval censoring, current status data, nonparametric estimation.

Introduction

Right-censored survival data model is widely applicable in practice in spite of in many cases the event times (failure or censoring) are not observed exactly, and the investigator observes time interval containing a failure time for each of not missed at follow up individuals having symptoms of disease at the endpoint. In the current status right-censored data model the event is observed in a random inspection time if it occurs before the inspection time or not observed otherwise.

Let T and U be the independent failure and censoring times respectively. Right-censored observation consists of the event time $X = T \wedge U$ and the indicator $\delta = \mathbb{1}_{\{T \le U\}}$. The current status right-censored observation is given as $(W, \kappa, \kappa\delta)$, where $\kappa = \mathbb{1}_{\{X \le W\}}$ and W is a random inspection time, which is independent of (T, U). The observed data is a sample from the distribution $(W, \kappa, \kappa\delta)$ and the main target of statistical analysis is the distribution function F of failure time T.

The right-censored survival data model is well developed. The Kaplan–Meier [19] estimator is widely applicable to estimate the survival function of failure time from right-censored data. Consistency and asymptotic normality of the Kaplan–Meier estimator are obtained first in [5]. The point process technique allows to get functional convergence results for the Kaplan–Meier estimator ([1, 9, 10]; see also

[7, 2]). Note that, the Kaplan–Meier estimator requires the exact event time to be observed, which may fail in practice. In the interval censored data model [27] the event times are not observed exactly. The nonparametric maximum likelihood estimator (NPMLE) for the current status data model can be obtained as a solution of the isotonic regression model [3] using Convex Minorant Algorithm or by using the EM-algorithm [26, 27]. Asymptotic behavior of the NPMLE at any fixed point studied in [11, 16]. Groeneboom & Wellner [16] discussed wide range of asymptotic results on the NPMLE.

The current status right-censored data model discussed in this paper is highly related to the particular case of the current status data with competing risks. The NPMLE and the nonparametric pseudo maximum likelihood estimator (NPPMLE) of parameters from the current status data with competing risks, and the EM-algorithms to get the estimators are given in [17]. Another naïve (ad-hoc) estimator is considered in [18], along with the NPMLE. Consistency and rate of convergence results for the NPMLE are obtained in [14], and weak convergence results are given in [15]. Consistency of the estimators in the current status right-censored data model and the rate of convergence results are obtained in [22].

The current status data and the life table data with a single observation time are particular cases of the model we discuss in this paper. The life table survival data model was widely used at the beginning of survival analysis [4, 6, 8]. The standard life table (actuarial) estimator is generally used to estimate the parameter $F(w_0)$. Breslow & Crowley [5] show that there is no consistent nonparametric estimator of completely unknown distribution function F at the observation time w_0 in the life table survival data model. Nevertheless, in many real cases the asymptotic bias of the standard life table estimator is relatively small [20]. The extended life table estimator that is inconsistent too was investigated in [24].

This work focuses on estimation in current status right-censored data model and investigates properties of nonparametric estimators under small and moderate sample sizes. We consider the NPMLE, the NPPMLE and the naïve estimator, which are obtaining from the corresponding estimators of the baseline current status data model with two competing risks. The maximum likelihood approach and some asymptotic properties of the estimators are discussed in Section 1. The estimation algorithms are displayed in Section 2. Some properties of the estimators obtained by simulations are reported in Section 3, and supplementary tables are postponed to Section 4.

1 The maximum likelihood approach

In this section we display the likelihood function for the interval right-censored data and discuss the nonparametric estimators.

Assume that the failure time T, the censoring time U and the observation time W are independent with the distribution functions F, G and J respectively; $\gamma_T = \sup\{x : F(x) < 1\}$ and $\gamma_G = \sup\{x : G(x) < 1\}$. Let (T_i, U_i, W_i) be a sample from the distribution (T, U, W), and $(W_i, \kappa_i, \kappa_i\delta_i)$ be the observed current status right-censored data, where $X_i = T_i \wedge U_i$, $\delta_i = \mathbb{1}_{\{T_i \le U_i\}}$ and $\kappa_i = \mathbb{1}_{\{X_i \le W_i\}}$, $i = 1, \ldots, n$.

We slightly abuse the notations denoting F, G, J and H for both the distribution functions and the corresponding measures.

The maximum likelihood estimate. Let $\mathcal{Q}$ be the set of nondecreasing nonnegative cadlag functions $Q : \mathbb{R} \to [0,1]$, such that $\lim_{x\to-\infty} Q(x) = 0$;

$$\mathbb{Q} = \{(Q, Q^*) : Q, Q^* \in \mathcal{Q} \text{ and } Q(x) + Q^*(x) \leq 1, x \in \mathbb{R}\}$$

be the set of parameters of the model. The log-likelihood function for the interval right-censored data is defined for $(Q, Q^*) \in \mathbb{Q}$ as follows:

$$\begin{aligned} LL(Q, Q^*) = \sum\nolimits_{i=1}^{n} & \kappa_i \delta_i \log Q(W_i) \\ & + \kappa_i (1-\delta_i) \log Q^*(W_i) + (1-\kappa_i) \log(1 - Q(W_i) - Q^*(W_i)), \end{aligned} \tag{1}$$

where $Q(x) = \int_0^x (1-G_-)dF = \int_0^x (1-H_-)d\Lambda$, $Q^* \equiv H - Q$ and $H \equiv 1-(1-F)(1-G)$ is the distribution function of the event time X, Λ is the cumulative hazard function corresponding to F restricted to $D_H = \{x : H(x) < 1\}$. A parameter $(\widehat{Q}_n, \widehat{Q}_n^*)$, which maximizes (1) over $(Q, Q^*) \in \mathbb{Q}$ is the NPMLE.

The pseudo maximum likelihood estimate. Let

$$\mathbb{Q}_H = \{(Q, Q^*) \in \mathbb{Q} : Q + Q^* \equiv H\}.$$

The likelihood function (1) can be rewritten as the sum of two terms $LL(W, \kappa, \kappa\delta; F, G) = LL^m(W, \kappa; H) + LL^r(W, \kappa, \kappa\delta; R)$ with

$$LL^m(W, \kappa; H) = \sum\nolimits_{i=1}^{n} \big(\kappa_i \log(H(W_i)) + (1-\kappa_i) \log(1 - H(W_i))\big)$$

and

$$LL^r(W, \kappa, \kappa\delta; R) = \sum\nolimits_{i=1}^{n} \big(\kappa_i \delta_i \log R(W_i) + \kappa_i (1-\delta_i) \log(1 - R(W_i))\big),$$

where $R(w) = Q(w)/H(w) = P(\delta = 1 | X \leq w) = \int_0^w (1 - H_-)d\Lambda / H(w)$. The functions Q and Q^* can be written as follows:

$$Q(x) = \int_0^x p dH \quad \text{and} \quad Q^*(x) = \int_0^x (1-p) dH, \tag{2}$$

where $p = \frac{d\Lambda}{d\Lambda^H}$ is the Radon–Nikodym derivative of the measure Λ with respect to Λ^H. Moreover, any measurable function $p : \mathbb{R} \to [0,1]$ defines the distributions of T and U (possibly improper) under any fixed distribution function H [22]. Let $\widetilde{H}_n$ be the sub distribution function, which maximizes LL^m and $\widetilde{R}_n \equiv \widetilde{Q}_n / \widetilde{H}_n$ maximizes LL^r under $H \equiv \widetilde{H}_n$, and $\widetilde{Q}(x) = \int_0^x p d\widetilde{H}_n$. Then $(\widetilde{Q}, \widetilde{Q}^*)$ such that $\widetilde{Q}^* \equiv \widetilde{H} - \widetilde{Q}$ is the NPPMLE for the parameter (Q, Q^*).

The naïve (ad hoc) approach is based on the separate estimation of the parameters Q and Q^* from the observations with $T < U$ and $T \geq U$ respectively. The naïve estimator $\widehat{\widehat{Q}}_n$ for the parameter Q is obtaining by maximizing

$$\Psi(W, \kappa\delta, Q) = \sum\nolimits_{i=1}^{n} (\kappa_i \delta_i \log Q(W_i) + (1 - \kappa_i \delta_i) \log(1 - Q(W_i))$$

on Q having atoms at the observation points W_i with $\kappa_i\delta_i = 1$, and the naïve estimator $\widehat{\widehat{Q}}{}^*_n$ for the parameter Q^* is obtaining by maximizing $\Psi(W, \kappa(1-\delta), Q^*)$ on Q^* having atoms at the observation points W_i with $\kappa_i(1-\delta_i)$, $i = 1, \dots, n$. The naïve estimator can be obtained by the regular convex minorant algorithm from right-censored data analysis. The true disadvantage of the naïve estimator is that the constraint $\widehat{\widehat{Q}}_n + \widehat{\widehat{Q}}_n \leq 1$ may fail in the general case.

Recovering the distributions of failure and censoring times. In order to recover the distribution of T from Q and Q^* we use that $\Lambda(x) = \int_0^x (1 - Q_- - Q^*_-)^{-1} dQ$. Hence,

$$S(t) = \prod_{x \leq t} \Big(1 - \frac{dQ(x)}{1 - Q(x_-) - Q^*(x_-)}\Big), \tag{3}$$

where $S \equiv 1 - F$. The distribution of censoring time U is determined by the cumulative hazard function $\Lambda^G(x) = \int_0^x \frac{1-F_-}{(1-F)(1-H_-)} dQ^*$ and, therefore, $G(t) = 1 - \prod_{x\leq t}\big(1 - d\Lambda^G(x)\big)$. Alternatively, $G(t) = \int_0^t (1-F)^{-1} dQ^*$, $t \in D_H$.

Large sample properties of the estimators. The large sample properties of the nonparametric estimator S_n $(S_n = \widehat{S}_n, \widetilde{S}_n, \widehat{\widehat{S}}_n)$ for the distribution of failure time are determined by the large sample properties of the corresponding estimator (Q_n, Q^*_n) for the parameter (Q, Q^*) that is the particular case of the estimator for the current status data with two competing risks model. The uniform consistency and the rate of convergence results for all the estimators S_n were obtained in [22]. In the absolutely continuous case it was proved that under the condition $H \ll J$, for any $\tau < \gamma_F \wedge \gamma_G$

$$\sup_{x \leq \tau} |F_n(x) - F(x)| \to 0,$$

as $n \to \infty$ almost sure. The uniform consistency result under the assumption $H \ll J$ remains correct in general case. The condition $H \ll J$ is important, otherwise there is no consistent estimator for the parameter S (see [23, 24]). The rate of convergence in the absolutely continuous case is obtained, under $H \ll J$ and the bounded property $M^{-1} \leq \frac{dH}{dJ} \leq M$ for some $M > 1$, in the $L_1(J)$ norm restricted to the interval $[0, \tau]$,

$$\|F_n - F\|_{1,J([0,\tau])} = O_P(n^{-1/3} \log^{1/3} n). \tag{4}$$

Remark 1. *(i). The rate of convergence in (4) is obtained from the refined rate of uniform convergence results for the corresponding estimators H_n of the event time distribution function H.*
(ii). We may expect the rate of convergence $O_P(n^{-1/3})$ in (4) taking into account the rate of convergence $O_P(n^{-1/3})$ of the estimators Q_n and Q^*_n to the parameters Q and Q^* in $L_1(J)$ (and even in $L_2(J)$ norm), but the L_p rate of convergence of the estimators Q_n and Q^*_n is not implies the same rate of convergence for the corresponding estimator S_n.
(iii). Local weak convergence theorems for the estimators $(\widehat{Q}_n, \widehat{Q}^*_n)$ and $(\widehat{\widehat{Q}}_n, \widehat{\widehat{Q}}{}^*_n)$ are given in [15], but there is no way to use these results in order to obtain weak convergence theorem for the corresponding estimators S_n.

2 Estimation algorithms

In this section we discuss algorithms for the NPMLE and the NPPMLE introduced in Section 1. Let $W_{(1)}, \ldots, W_{(r)}$ be the set of observation times in ascending order without replications. The likelihood function (1) can be rewritten in terms of parameters $(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$: $\theta_i = Q(W_{(i)})$ and $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_r^*)$: $\theta_i^* = Q^*(W_{(i)})$, $i = 1, \ldots, r$, as follows:

$$\psi(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \sum_{i=1}^{r} \kappa\delta_{(i)} \log \theta_i + \kappa\bar{\delta}_{(i)} \log \theta_i^* + \bar{\kappa}_{(i)} \log(1 - \theta_i - \theta_i^*),$$

where $\bar{\kappa}_{(i)}$ ($\kappa\delta_{(i)}$, $\kappa\bar{\delta}_{(i)}$) is the total number of observations $W_j = W_{(i)}$ such that $\kappa_j = 0$ ($\kappa_j = 1$ and $\delta_j = 1$, $\kappa_j = 1$ and $\delta_j = 0$, respectively). The optimization problem is to maximize $\psi(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ on $(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \in \overline{\mathbb{S}}$, where

$$\overline{\mathbb{S}} = \{(\boldsymbol{\theta}, \boldsymbol{\theta}^*) : 0 \le \theta_1 \le \ldots \le \theta_r, 0 \le \theta_1^* \le \ldots \le \theta_r^*, \theta_r + \theta_r^* \le 1\},$$

Let

$$\mathbb{S} = \{(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \in \overline{\mathbb{S}} : \theta_i = \theta_{i-1} \textit{ if } \kappa\delta_{(i)} + \bar{\kappa}_{(i)} = 0 \quad \textit{and} \quad \theta_i^* = \theta_{i-1}^* \textit{ if } \kappa\bar{\delta}_{(i)} + \bar{\kappa}_{(i)} = 0,\ i = 1, \ldots, r\}$$

with the notations $\theta_0 = \theta_0^* = 0$. The NPMLE $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}^*)$, which maximizes ψ over $(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \in \mathbb{S}$, is maximizes ψ over $(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \in \overline{\mathbb{S}}$. Moreover, $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}^*)$ is uniquely defined, and $\hat{\theta}_r + \hat{\theta}_r^* = 1$ iff $\bar{\kappa}_{(r)} = 0$ [14].

The maximum likelihood estimation requires first to get the NPMLE $(\widehat{Q}, \widehat{Q}^*)$ of the parameter (Q, Q^*) and then recovering the survival function $\widehat{S}_n$ of failure time by formula (3). The first step reduced to the maximum likelihood estimation in the current status data with two competing risks model. The EM-algorithm due to [17] to get the NPMLE for the parameter (Q, Q^*) is working too slow, and one can use the iterated convex minorant (ICM) algorithm (see [12]) based on the characterization of the NPMLE from current status data with competing risk in [14]. Alternatively, the NPMLE for the parameter (Q, Q^*) can be obtained by using the support reduction algorithm [13], which is realized in the R-package *MLEcens* [21].

The pseudo likelihood estimation consists of three steps. At the first step we get the NPMLE $\widetilde{H}_n$ of the parameter H from the interval censored data (X_i, W_i), $i = 1, \ldots, n$. The convex minorant algorithm is a common way to get the maximum likelihood estimator $\widetilde{H}_n$ [16]. At the second step we get the estimator $(\widetilde{Q}_n, \widetilde{Q}_n^*)$, which maximizes LL^r under $\widetilde{Q}_n^* = \widetilde{H}_n - \widetilde{Q}_n$. We study an algorithm to obtain $\widetilde{R}_n \equiv \widetilde{Q}_n / \widetilde{H}_n$ under known $H \equiv \widetilde{H}_n$ from the observed data. Let $W_{(1)}^{**}, \ldots, W_{(m)}^{**}$ be the set of admissible step points of the estimator $\widetilde{R}_n$ in ascending order, including the observation times W_i with $\kappa_i = 1$; $h_1 = H(W_{(1)}^{**}) > 0$ and $h_i = H(W_{(i)}^{**}) - H(W_{(i-1)}^{**}) > 0$ for all $i = 2, \ldots, m$; $\delta_{(i)}^{**} = \sum_{j: W_j = W_{(i)}^{**}} \delta_j$ be the number of observed failures at $W_{(i)}^{**}$, $i = 1, \ldots, m$. It follows from (2) that $\widehat{R}(W_{(s)}^{**}) = \sum_{i=1}^{s} h_i \zeta_i / \sum_{i=1}^{s} h_i$. Then the pseudo-likelihood function L^r can be rewritten in the following way:

$$L^r(\boldsymbol{\zeta}) = \exp(LL^r(F,G)) = \prod_{s=1}^{m}\Big(\sum_{i=1}^{s} h_i\zeta_i/\sum_{i=1}^{s} h_i\Big)^{\delta^{**}_{(s)}}\Big(\sum_{i=1}^{s} h_i(1-\zeta_i)/\sum_{i=1}^{s} h_i\Big)^{\nu_s-\delta^{**}_{(s)}}$$
$$\cong \prod_{s=1}^{m}\Big(\sum_{i=1}^{s} h_i\zeta_i\Big)^{\delta^{**}_{(s)}}\Big(\sum_{i=1}^{s} h_i(1-\zeta_i)\Big)^{\nu_s-\delta^{**}_{(s)}},$$

where $\boldsymbol{\zeta} = (\zeta_1,\ldots,\zeta_m)$: $\zeta_i = p(W^{**}_{(i)}) \in [0,1]$, ν_i is the total number of observed events at $W^{**}_{(i)}$, $i = 1,\ldots,m$. The estimation problem reduces to maximizing the expression

$$\phi(\boldsymbol{\zeta}) = \sum\nolimits_{s=1}^{m}\Big(\delta^{**}_{(s)}\log\Big(\sum\nolimits_{i=1}^{s} h_i\zeta_i\Big) + (\nu_s-\delta^{**}_{(s)})\log\Big(\sum\nolimits_{i=1}^{s} h_i(1-\zeta_i)\Big)\Big) \quad (5)$$

over the set of $\boldsymbol{\zeta} \in [0,1]^m$. Finally, at the third step one use the reconstruction formula (3) to obtain the NPPMLE $\widetilde{S}_n$ for S.

3 Simulations

In this section we consider specific designs (DS) to evaluate finite-sample performance of the NPMLE, NPPMLE and the naïve estimator from simulated data. We perform simulations of the current status right-censored data with different rates of observations with known status (failure or censoring) $p_\kappa = P(X \le W)$, which are applicable for estimation of the parameter Q, and different rates of observed failures $p_\delta = P(\delta = 1|\kappa = 1)$, under the three sample sizes of 200, 500 and 1000. We denote $\Gamma(a,b)$ is the gamma distribution and $\mathcal{W}(a,b)$ is the Weibull distribution with the shape parameter $a > 0$ and scale parameter $b > 0$; $\mathrm{E}(1/b) = \Gamma(1,b)$ is the exponential distribution; $\mathrm{LN}(m,b)$ and $\mathrm{FN}(m,b)$ is the lognormal and the folded-normal distribution with parameters $m \in \mathbb{R}$ and $b > 0$ respectively. The following table 1 collects main features of the experimental designs used for the simulations.

Table 1. Main features of the experimental designs

DS	T	U	W	p_κ	p_δ	DS	T	U	W	p_κ	p_δ
A	$\Gamma(1/2,1)$	$\Gamma(2,1)$	LN(0,1)	0.83	0.91	D	$\Gamma(2,1)$	E(1)	FN(0,1)	0.54	0.19
B	$\Gamma(1/2,1)$	E(1)	FN(0,1)	0.80	0.73	E	$\frac{1}{3}\Gamma(2,\frac{1}{5})+\frac{2}{3}\Gamma(10,\frac{1}{5})$	E(1/2)	E(1)	0.50	0.49
C	$\Gamma(3,1)$	E(1)	E(1)	0.52	0.07	F	$\frac{1}{2}\mathcal{W}(\frac{1}{2},1)+\frac{1}{2}\mathcal{W}(5,1)$	E(1)	E(3/2)	0.48	0.47

The same experimental designs were used in [22] to perform large sample properties of the NPMLE, NPPMLE and its bootstrapped version by simulations.

In order to perform simulations we use R statistical software [25]. The function *computeMLE()* of *MLEcens* package is used to create the MLE $(\widehat{Q}_n, \widehat{Q}^*_n)$ for the parameter (Q,Q). We use the the convex minorant algorithm realization *gcmlcm()* of package *pdrtool* to get the estimator $\widetilde{H}$ for the distribution of the event time H, and the function *lbfgsb3()* of the same name package to solve the optimization problem in (5) under $H \equiv \widetilde{H}$ and obtain the estimator $\widetilde{Q}_n$. Finally, we obtain the estimators $\widehat{S}_n$, $\widetilde{S}_n$ and $\widetilde{\widetilde{S}}_n$ for the survival function S of failure time from $(\widehat{Q}_n, \widehat{Q}^*_n)$, $(\widetilde{Q}_n, \widetilde{Q}^*_n)$ by the reconstruction formula (3).

For the NPMLE, NPPMLE and naïve estimators we display the estimation bias (Section 4, Table I) and the mean absolute estimation error (Section 4, Table II) at the quartiles (Q25,MED,Q75) and 95%-quantile Q95 of the failure time distribution, as well as the supremum $\sup_{x\in[0,Q]}|S_n(x) - S(x)|$ (Section 4, Table III) and the L_1 norm adjusted to the interval length $\|S_n - S\|_{1,[0,Q]}/J((0,Q])$ (Section 4, Table IV) restricted to the interval $[0, Q]$ for Q =Q25,MED,Q75,Q95. The results are obtained separately by using 10^4 replications.

First, we note that the finite sample performance of the estimators is highly related to the experimental design features. The designs A and B display very good approximation quality for MED$-Q95$ quantiles, but there is an obvious problems in the estimation of the survival distribution of failure time at first quartile $Q25$, especially under the experimental design A because of $\frac{dQ}{dJ}(w) \to \infty$ as $w \to 0_+$. On the other hand, the number of observations is insufficient to get good enough nonparametric estimates under the designs C and D having a very small rate of observed failures. All the estimators display good enough finite sample performance under the designs E and F with the bimodal distributions of failure time. The $L_1(J)$ divergence display quite small estimation error for all the designs except the design A, and the uniform norm divergence is too high under these sample sizes. Moreover, both the $L_1(J)$ and the uniform estimation errors are not highly dependent of the population sizes from 200 to 1000.

Roughly, the nonparametric estimators show very similar finite sample performance for each of the designs. More careful look at the results allows us to give some preference to the NPMLE, which displays a little bit smaller divergence in almost all the cases. In most of cases the NPPMLE performs a little bit better results then the naïve estimator, but it displays a huge bias (overestimation of the survival function) at Q75 and Q95 quantile points that should be explained by accumulation of the bias and the estimation error appears under estimation of the event time distribution H and the competing risks components (Q, Q^*) under fixed $H \equiv \widetilde{H}$ in the adverse experimental conditions.

4 Supplementary tables

Table I. The estimation bias

DS	N	NPMLE				NPPMLE				Naïve			
		Q25	MED	Q75	Q95	Q25	MED	Q75	Q95	Q25	MED	Q75	Q95
A	200	0.2314	0.0309	6.9E-4	-0.009	0.2314	0.0282	0.0013	-0.0138	0.2337	0.0885	0.0947	0.137
	500	0.2088	0.0141	9.4E-5	-0.0128	0.2086	0.012	0.0012	-0.007	0.2141	0.0768	0.1002	0.1505
	1000	0.1804	0.0085	1.5E-4	-0.0101	0.1803	0.0066	7.9E-4	-1.1E-4	0.1899	0.0735	0.1029	0.1565
B	200	0.0716	0.0177	-0.0045	-0.0029	0.0602	0.0171	0.0150	0.0132	0.1048	0.1112	0.1711	0.1664
	500	0.0314	0.0087	-0.0012	-0.0139	0.0242	0.0099	0.0112	0.0022	0.073	0.105	0.1721	0.2128
	1000	0.0192	0.0052	-0.0010	-0.0184	0.0151	0.0074	0.0085	-0.0025	0.0624	0.1019	0.171	0.2353
C	200	-0.0439	-0.0914	0.0561	0.2349	0.0112	0.1732	0.4094	0.6086	0.0475	0.0558	0.0447	0.1064
	500	-0.0277	-0.1052	0.0146	0.1959	0.0018	0.0932	0.3181	0.5163	0.0735	0.1515	0.1711	0.0989
	1000	-0.0173	-0.0926	-0.0097	0.173	0.0052	0.0429	0.2500	0.4471	0.0781	0.1853	0.2528	0.1257
D	200	-0.0081	-0.0932	-0.0318	0.1589	0.0106	0.0238	0.1893	0.3865	0.0554	0.073	-0.0657	0.0987
	500	-0.0051	-0.0605	-0.0669	0.1193	0.0084	0.0027	0.1278	0.3222	0.0616	0.1278	0.0042	0.0882
	1000	-0.0045	-0.0337	-0.0858	0.0945	0.0062	0.0015	0.0852	0.2776	0.0627	0.1492	0.0865	0.0802
E	200	0.0146	-0.0392	-0.0700	-0.0097	0.0176	-0.0045	0.0594	0.1742	0.0608	0.1036	0.2079	0.2467
	500	0.0084	-0.0183	-0.0510	-0.022	0.0106	-0.0010	0.0383	0.1169	0.059	0.1200	0.2311	0.3046
	1000	0.0057	-0.0102	-0.0290	-0.0248	0.0086	-3.1E-4	0.0237	0.0796	0.0584	0.1265	0.2359	0.3278
F	200	-0.0075	-0.0382	-0.0378	0.053	-0.0067	0.0057	0.0851	0.1612	0.0449	0.1289	0.2572	0.0964
	500	-0.0047	-0.021	-0.0269	0.043	-0.0039	7.5E-4	0.0491	0.0852	0.0486	0.135	0.2653	0.0951
	1000	-0.0032	-0.0121	-0.0166	0.0347	-0.0020	0.0018	0.0306	0.0526	0.0508	0.1377	0.265	0.0946

Table II. The absolute error

DS	N	NPMLE				NPPMLE				Naïve			
		Q25	MED	Q75	Q95	Q25	MED	Q75	Q95	Q25	MED	Q75	Q95
A	200	0.2413	0.0946	0.0560	0.0355	0.2412	0.0945	0.0583	0.0537	0.2404	0.1161	0.1096	0.1446
	500	0.2257	0.0637	0.0402	0.0288	0.2256	0.0643	0.0425	0.0424	0.2244	0.0934	0.1059	0.1519
	1000	0.2024	0.0485	0.032	0.0236	0.2019	0.0492	0.034	0.0328	0.2016	0.083	0.1055	0.1569
B	200	0.1251	0.0774	0.0649	0.0487	0.1215	0.0796	0.074	0.0804	0.1278	0.1214	0.1749	0.1887
	500	0.0770	0.0561	0.0451	0.0372	0.0761	0.0591	0.0521	0.0645	0.0882	0.1099	0.173	0.2198
	1000	0.0571	0.0442	0.0350	0.0327	0.0573	0.0475	0.0416	0.0540	0.0726	0.1041	0.1713	0.2375
C	200	0.1534	0.2540	0.2696	0.2803	0.1341	0.2809	0.4635	0.6194	0.1424	0.2826	0.3063	0.1780
	500	0.1009	0.2052	0.2188	0.2390	0.1023	0.2165	0.3763	0.5278	0.1065	0.2352	0.3216	0.1689
	1000	0.0741	0.1687	0.1889	0.2143	0.0804	0.1747	0.3103	0.4584	0.0969	0.2214	0.3367	0.1899
D	200	0.0784	0.1706	0.1811	0.2009	0.0801	0.179	0.3076	0.4101	0.0894	0.1893	0.2463	0.1647
	500	0.0561	0.1205	0.1520	0.1619	0.0601	0.1317	0.2474	0.3459	0.0770	0.1694	0.2437	0.1544
	1000	0.0442	0.0855	0.1406	0.1384	0.0485	0.1045	0.2086	0.3014	0.0719	0.1671	0.2432	0.1467
E	200	0.0574	0.1186	0.1606	0.0643	0.0590	0.1116	0.1979	0.2216	0.0730	0.1327	0.2306	0.2634
	500	0.0404	0.0830	0.1172	0.0522	0.0425	0.0830	0.1334	0.1636	0.0639	0.1297	0.236	0.3083
	1000	0.0321	0.0653	0.0867	0.0464	0.034	0.0678	0.0979	0.1229	0.0609	0.1302	0.2376	0.3287
F	200	0.0558	0.1321	0.1258	0.0919	0.0559	0.1256	0.1851	0.2089	0.0638	0.1509	0.2662	0.1498
	500	0.0395	0.0935	0.0898	0.0719	0.0412	0.0940	0.1186	0.1336	0.0567	0.1434	0.2672	0.1481
	1000	0.0312	0.0715	0.0693	0.0608	0.0328	0.0763	0.0871	0.0997	0.0547	0.1417	0.2654	0.1472

Table III. The supremum norm divergence

		NPMLE				NPPMLE				Naïve			
DS	N	Q25	MED	Q75	Q95	Q25	MED	Q75	Q95	Q25	MED	Q75	Q95
A	200	0.2490	0.3506	0.3512	0.3512	0.2490	0.3505	0.351	0.3511	0.2487	0.3515	0.3537	0.3564
	500	0.2466	0.3059	0.3060	0.3060	0.2466	0.3058	0.3059	0.3059	0.2462	0.3069	0.3084	0.3131
	1000	0.2423	0.2767	0.2767	0.2767	0.2424	0.2767	0.2767	0.2767	0.2420	0.2779	0.2795	0.2869
B	200	0.2036	0.2327	0.2381	0.2393	0.2025	0.2305	0.2373	0.2438	0.1998	0.2429	0.2752	0.3167
	500	0.1661	0.1805	0.1836	0.1841	0.1645	0.1801	0.1846	0.1888	0.1650	0.1988	0.2380	0.2940
	1000	0.1379	0.1480	0.1499	0.1502	0.137	0.1486	0.1520	0.1549	0.1401	0.1734	0.2182	0.2849
C	200	0.2126	0.3879	0.4621	0.5364	0.1836	0.3432	0.5240	0.6872	0.183	0.3607	0.4755	0.5256
	500	0.1524	0.323	0.3913	0.457	0.1475	0.2897	0.4451	0.5963	0.1394	0.2901	0.4287	0.4942
	1000	0.1205	0.2683	0.3449	0.4044	0.1230	0.2504	0.3845	0.5249	0.1252	0.2655	0.4136	0.4951
D	200	0.1323	0.2673	0.3465	0.3960	0.1297	0.2530	0.3916	0.5178	0.1306	0.2573	0.3777	0.4199
	500	0.1014	0.1957	0.2978	0.3351	0.1038	0.1995	0.3258	0.4421	0.1101	0.2207	0.3453	0.3898
	1000	0.0837	0.1513	0.2663	0.2972	0.0877	0.1662	0.2832	0.3909	0.0998	0.2084	0.3335	0.3794
E	200	0.1189	0.1953	0.2862	0.3005	0.1171	0.1856	0.2860	0.3637	0.1227	0.1908	0.3013	0.3913
	500	0.0909	0.1447	0.2193	0.2340	0.0919	0.1447	0.2189	0.2800	0.1021	0.1732	0.2893	0.3871
	1000	0.0748	0.1184	0.1751	0.1905	0.0768	0.1217	0.1803	0.2255	0.0913	0.1663	0.2819	0.383
F	200	0.1146	0.2122	0.2812	0.2905	0.1132	0.2002	0.2895	0.3578	0.1161	0.2022	0.3288	0.4004
	500	0.0878	0.1623	0.2149	0.2208	0.0880	0.1616	0.2223	0.2632	0.0964	0.1859	0.3120	0.3855
	1000	0.072	0.1324	0.1730	0.1776	0.0726	0.1365	0.1852	0.2120	0.0859	0.1787	0.3019	0.3768

Table IV. The L_1 norm divergence

		NPMLE				NPPMLE				Naïve			
DS	N	Q25	MED	Q75	Q95	Q25	MED	Q75	Q95	Q25	MED	Q75	Q95
A	200	0.2159	0.1525	0.0854	0.0653	0.2159	0.1541	0.0876	0.0709	0.2154	0.1643	0.1188	0.1224
	500	0.2096	0.1001	0.0590	0.0459	0.2095	0.1015	0.0610	0.0493	0.2085	0.1195	0.1018	0.1152
	1000	0.1980	0.0735	0.0452	0.0353	0.1979	0.0747	0.0469	0.0379	0.1971	0.0972	0.0944	0.1130
B	200	0.1207	0.0979	0.0803	0.0717	0.1194	0.0979	0.0831	0.0849	0.1184	0.1174	0.1366	0.1675
	500	0.0876	0.0688	0.0568	0.0510	0.0864	0.0699	0.0604	0.0607	0.0894	0.0951	0.1246	0.1653
	1000	0.0663	0.0528	0.0441	0.0395	0.0656	0.0544	0.0478	0.0476	0.0718	0.0845	0.119	0.1642
C	200	0.0337	0.0541	0.0645	0.0683	0.0307	0.0496	0.0651	0.0734	0.031	0.0512	0.0642	0.0677
	500	0.0231	0.0381	0.0469	0.0501	0.0226	0.0375	0.0498	0.0568	0.0231	0.0392	0.0515	0.0556
	1000	0.0177	0.0290	0.0367	0.0396	0.0181	0.0302	0.0401	0.0461	0.0199	0.0348	0.0470	0.0518
D	200	0.0391	0.0596	0.0704	0.0712	0.0377	0.0592	0.0741	0.0759	0.0400	0.0643	0.0787	0.0797
	500	0.0281	0.0414	0.0508	0.0515	0.0283	0.0438	0.0552	0.0567	0.0318	0.0541	0.0673	0.0684
	1000	0.0220	0.0318	0.0396	0.0403	0.0227	0.0349	0.0442	0.0455	0.0277	0.0500	0.0633	0.0644
E	200	0.0437	0.0572	0.0667	0.0703	0.0426	0.0566	0.0665	0.0771	0.0468	0.0659	0.0778	0.0905
	500	0.0313	0.0405	0.0470	0.0504	0.0314	0.0413	0.0480	0.0556	0.0375	0.058	0.0707	0.0854
	1000	0.0246	0.0317	0.0365	0.0394	0.0250	0.0329	0.0381	0.0434	0.0326	0.0547	0.0678	0.0833
F	200	0.0499	0.0610	0.0697	0.0741	0.0491	0.0597	0.0699	0.0946	0.0562	0.0677	0.0826	0.1133
	500	0.0348	0.0432	0.0493	0.0524	0.0350	0.0439	0.0506	0.0652	0.046	0.0585	0.0739	0.1096
	1000	0.0270	0.0338	0.0383	0.0408	0.0275	0.0349	0.0401	0.0500	0.0414	0.0549	0.0704	0.1089

Acknowledgements

This research was supported in part by St.-Petersburg State University (grants No 1.52.1647.2016).

References

[1] Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics* **6**(4), pp. 701–726.

[2] Andersen, P. K., Borgan, Ø., R. D. Gill and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* New York: Springer-Verlag.

[3] Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T. and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics* **26**(4), 641–647.

[4] Berkson and Gage (1950). Calculation of survival rates for cancer. *Proceedings of Staff Meetings of the Mayo Clinic*, Vol. **25**, pp. 270–286.

[5] Breslow, N. and Crowley, J. (1974) A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship. *The Annals of Statistics*, Vol. **2**(3), pp. 437–453.

[6] Cutler and Ederer (1958). Maximum utilization of the life table method in analyzing survival. *J. Chron.Dis.*, . Vol. **8**, pp. 699–712.

[7] Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* New York: John Wiley & Sons.

[8] Gehan (1969). Estimating survival function from the life table. *J.Chron.Dis.*, Vol. **21**, pp. 629–644.

[9] Gill, R. D. (1980a). *Censoring and stochastic integrals.* Mathematical Centre Tracts 124. Amsterdam: Mathematisch Centrum. V, 178 p.

[10] Gill, R. D. (1980b). Nonparametric estimation based on censored observations of a markov renewal process. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete* **53**(1), 97–116.

[11] Groeneboom, P. (1991). *Nonparametric maximum likelihood estimators for interval censoring and deconvolution.* Technical Report 378, Department of Statistics, Stanford University.

[12] Groeneboom, P. and Jongbloed, G. (2014). *Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[13] Groeneboom, P., G. Jongbloed, and J. Wellner (2008). The support reduction algorithm for computing non-parametric function estimates in mixture models. *Scandinavian Journal of Statistics* **35**(3), 385–399.

[14] Groeneboom, P., M. Maathuis, and J. Wellner (2008a). Current status data with competing risks: Consistency and rates of convergence of the mle. *Ann. Statist.* **36**(3), 1031–1063.

[15] Groeneboom, P., Maathuis, M. and Wellner, J. (2008b). Current status data with competing risks: Limiting distribution of the mle. *Ann. Statist.* **36**(3), 1064–1089.

[16] Groeneboom, P. and J. A. Wellner (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation.* DMV Seminar Band 19. Basel: Birkhauser Verlag.

[17] Hudgens, M. G., Satten, G. and Longini, I. (2001). Nonparametric maximum likelihood estimation for competing risks survival data subject to interval censoring and truncation. *Biometrics* **57**(1), 74–80.

[18] Jewell, N., van der Laan, M. and Henneman, T. (2003). Nonparametric estimation from current status data with competing risks. *Biometrika* **90**(1), 183–197.

[19] Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**(282), 457–481.

[20] Lawless, J. F. (2002). *Statistical Models and Methods for Lifetime Data, 2-nd edition.* Wiley Series in Probability and Statistics. John Wiley & Sons.

[21] Maathuis, M. (2013). *MLEcens: Computation of the MLE for bivariate (interval) censored data.* R package version 0.1-4.

[22] Malov S.V. (2019). Nonparametric estimation for current status interval right-censored data model. *Statistica Neerlandica*, 1–21. Published online https://doi.org/10.1111/stan.12180.

[23] Malov S.V. & O'Brien S.J. (2015). On Survival Categorical Methods Based on an Extended Actuarial Estimator. In *Applied Methods of Statistical Analysis. Nonparametric Approach.* Proceedings of the International Workshop AMSA'15 (Novosibirsk & Belokurikha, September 14-19, 2015), pp. 147–156.

[24] Malov S.V. & O'Brien S.J. (2018). Life Table Estimator Revisited. *Communication in Statistics – Theory and Methods* **47**(9), 2126–2133.

[25] R Core Team (2014). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

[26] Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association* **69**, 169–173.

[27] Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society.* Series B **38**, 290–295.

**APPLIED METHODS OF STATISTICAL ANALYSIS.
STATISTICAL COMPUTATION AND SIMULATION**

Proceedings
of the international workshop
Novosibirsk, 18-20 September 2019

Научное издание

**ПРИКЛАДНЫЕ МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА.
СТАТИСТИЧЕСКИЕ РАСЧЕТЫ И МОДЕЛИРОВАНИЕ**

ТРУДЫ
Международного семинара
Новосибирск, 18-20 сентября 2019 г.

На английском языке

Подписано в печать 16.08.2019. Формат 60 × 84 1/8.
Усл. печ. л. 71,75. Уч.-изд. л. 133,46.
Тираж 60 экз. Заказ № 1172.

Отпечатано в типографии
Новосибирского государственного технического университета
630073, г. Новосибирск, пр. К. Маркса, 20