Estimation for partially observed semi-Markov processes via self-consistency equations

Odile Pons

INRA, Mathématiques et informatique appliquée, 78352 Jouy-en-Josas cedex, France
(e-mail: odile.pons@jouy.inra.fr)

Abstract. Nonparametric estimators of the survival function $S(t) = P(T \geq t)$ for a censored time variable T has been defined by several methods, in particular by integral self-consistency equations since Efron (1967) [Chang and Yang, 1987]. We establish explicit expressions of the estimators in an additive form and extend this approach to several cases: a left-truncated and right-censored variable, the left-censored or left-truncated sojourn times of a right-censored semi-Markov process.
Keywords: left-truncation, right-censoring, self-consistency, semi-Markov process.

1 Introduction

Semi-Markov processes are non-homogeneous models for the evolution of individuals or systems between several states or submitted to several kinds of damage. They may be applied to data in biomedicine, biology, demography and quality control. For instance, the comparison of two treatments in patients may involve not only the final event, death or recovery, but also their evolution between several health states or their quality of life during a disease. The transition times between the states are not always observed and their values may be missing due to several possible observation scheme. In some cases the estimation of the survival function has only solved by recursive algorithms. This paper presents the usual product-limit estimator of right-censored survival function as a sum and provide closed form expressions of a survival function under left and right censoring or truncation. The estimators are extended to estimate the distribution of sojour times of a semi-Markov process under similar censorship and truncation.

2 Estimation of right-censored and left-truncated variables

2.1 Right-censored variables

Let $(X_i, \delta_i)_{i \leq n}$ be a sample of real time variables and censoring indicators, $X_i = T_i \wedge C_i$ and $\delta_i = 1\{T_i \leq C_i\}$, where T and C have the distribution functions F and G, and survival functions S and $\bar{G}$. Let $N_n(t) = \sum_i \delta_i 1_{\{X_i \leq t\}}$

and $\widehat{S}_n$ satisfying the self-consistency equation

$$\widehat{S}_n(t) = n^{-1} \sum_{i=1}^{n} \{1_{\{X_i > t\}} + (1-\delta_i) 1_{\{X_i \le t\}} \frac{\widehat{S}_n(t)}{\widehat{S}_n(X_i)}\}, \tag{1}$$

then equation (1) uniquely defines an estimator of S if the censoring distribution is continuous,

$$\widehat{S}_n(t) = 1 - \int_0^t \frac{dN_n(s)}{n - \sum_{j=1}^n (1-\delta_j) 1_{\{X_j < s\}} \widehat{S}_n^{-1}(X_j)}$$

and $\widehat{S}_n(t) \equiv \widehat{\bar{F}}_n(t)$, the Kaplan-Meier estimator.

2.2 Numerical example

Let $(X_{(1)} < X_{(2)} < \ldots < X_{(n)})$ be the ordered sample $(X_i)_{i \le n}$ and $\delta_{(i)}$ be the indicator related to $X_{(i)}$. The estimator $\widehat{S}_n$ is as a right-continuous decreasing step function with jumps at the uncensored observations, starting from $\widehat{S}_n(0) = 1$ and with

$$\widehat{S}_n(X_{(i)}) = \widehat{S}_n(X_{(i-1)}) - \frac{\delta_{(i)}}{n - \sum_{j=1}^n (1-\delta_j) 1_{\{X_j \le X_{(i-1)}\}} \widehat{S}_n^{-1}(X_j)}.$$

Consider a sample such that $(\delta_{(i)})_{i \le n} = (1,0,1,1,0,0,0,1,1,1)$, then the sequence $(\widehat{S}_n(X_{(i-1)}, \widehat{S}_n(X_{(i)}) - \widehat{S}_n(X_{(i-1)})_{i \le n}$ takes the values

$$((1, \frac{1}{10}), (\frac{9}{10}, 0), (\frac{9}{10}, \frac{9}{8} \times \frac{1}{10}), (\frac{9}{8} \times \frac{7}{10}, \frac{9}{8} \times \frac{1}{10}), (\frac{9}{8} \times \frac{6}{10}, 0), (\frac{9}{8} \times \frac{6}{10}, 0),$$
$$(\frac{9}{8} \times \frac{6}{10}, 0), (\frac{9}{8} \times \frac{6}{10}, \frac{9}{4 \times 10}), (\frac{9}{2 \times 10}, \frac{9}{4 \times 10}), (\frac{9}{4 \times 10}, \frac{9}{4 \times 10})).$$

The product-limit estimator of Kaplan-Meier is defined as

$$\widehat{\bar{F}}_n(t) = \prod_{X_i \le t} \Big\{1 - \frac{\delta_i}{Y_n(X_i)}\Big\}$$

with $Y_n(t) = \sum_{i=1}^n 1_{\{X_i \ge t\}}$. For the above sample

$$(1 - N_n(X_{(i)})\, Y_n^{-1}(X_{(i)}))_{i \le n} = (\frac{9}{10}, 1, \frac{7}{8}, \frac{6}{7}, 1, 1, 1, \frac{2}{3}, \frac{1}{2}, 0),$$

$\widehat{\bar{F}}_n$ is a step function with jumps at the $X_{(i)}$'s and the values $(\widehat{\bar{F}}_n(X_{(i)}))_{i \le n}$ are

$$(\frac{9}{10}, \frac{9}{10}, \frac{9}{10} \times \frac{7}{8}, \frac{9}{10} \times \frac{6}{8}, \frac{9}{10} \times \frac{6}{8}, \frac{9}{10} \times \frac{6}{8}, \frac{9}{10} \times \frac{6}{8}, \frac{9}{10 \times 2}, \frac{9}{10 \times 4}, 0).$$

2.3 Left-truncated and right-censored variables

Under left-truncation and right-censoring, the variables X_i and δ_i for individual i are observed only if $X_i > U_i$, where T_i, C_i, U_i are independent with dfs $1-S$, G and H respect. Let $Y_n(t) = \sum_{i=1}^n 1_{\{U_i<t\leq X_i\}}$. As in (1), self-consistency property for the estimator of $H(t)$ and $H(t)S(t) = P(U \leq t \leq T)$ may be written

$$\widehat{H}_n(t) = n^{-1}\{Y_n(t^+) + \sum_{i=1}^n 1_{\{U_i<X_i\leq t\}} \frac{\widehat{H}_n(t)}{\widehat{H}_n(X_i)}\}$$

$$\widehat{H}_n(t)\widehat{S}_n(t) = n^{-1}\{Y_n(t^+) + \sum_{i=1}^n (1-\delta_i) 1_{\{U_i<X_i\leq t\}} \frac{\widehat{H}_n(t)\widehat{S}_n(t)}{\widehat{H}_n(X_i)\widehat{S}_n(X_i)}\}, \quad (2)$$

Let $R_U(i)$ and $R_X(i)$ be the ranks of $U(i)$ and $X(i)$. A direct estimator of $H(t)$ as a right-continuous increasing step function with jumps at the observations U_i, $i = 1, \ldots, n$, and starting from $\widehat{H}_n(0) = 0$ is defined by

$$\widehat{H}_n(U_{(i+1)}) = \widehat{H}_n(U_{(i)}) + \frac{1_{\{U_{(i+1)}<X_{R_U(i+1)}\}}}{n - \sum_{j=1}^n 1_{\{X_j<U_j\leq U_{(i)}\}} \widehat{H}_n^{-1}(U_j)}. \quad (3)$$

Moreover, $\widehat{H}_n$ given by (3) is equal to the product-limit estimator of H ([Woodroof, 1985]),

$$\widehat{H}_n^{pl}(t) = \prod_{1\leq i\leq n} \{1 - \frac{1_{\{Y_n(X_i)>0\}} 1_{\{U_i<t\wedge X_i\}}}{Y_n(X_i)}\}.$$

By (2) an estimator $\widehat{S}_n$ is defined as a step function with jumps at the observed X_i, with $\widehat{S}_n(0) = 1$ and such that

$$(\widehat{H}_n\widehat{S}_n)(X_{(i)}) = (\widehat{H}_n\widehat{S}_n)(X_{(i-1)}) - \frac{\delta_{(i)} 1_{\{U_{R_X(i)}<X_{(i)}\}}}{n - \sum_{j=1}^n (1-\delta_j) 1_{\{U_j<X_j\leq X_{(i-1)}\}} (\widehat{H}_n\widehat{S}_n)_n^{-1}(X_j)}.$$

3 Estimation of a semi-Markov process under right-censoring

We consider a n independent observations of a semi-Markov jump process in a finite state space $\{1, \ldots, m\}$. The i^{th} sample path of the process is defined by the sequence of the different sojourn states $J_i = (J_{i,k})_{k\geq 0}$ and by the sequence of the transition times $T_i = (T_{i,k})_{k\geq 0}$, with $T_{i,0} = 0$ and $T_{i,k}$ is the arrival time in state $J_{i,k}$, up to a random time t_i. For $k \geq 1$, $i = 1, \ldots, n$, the sojourn time $X_{i,k} = T_{i,k} - T_{i,k-1}$ in a transient state $J_{i,k-1}$ may therefore be right-censored by a random variable $C_{i,k}$ and the observations are $X_{i,k} \wedge C_{i,k}$ and the indicator $\delta_{i,k} = 1_{\{X_{i,k} \le C_{i,k}\}}$. The variable $C_{i,k}$ is supposed independent of $(T_{i,1}, \cdots, T_{i,k-1})$ and $(J_{i,0}, \cdots, J_{i,k-2})$ but to depend only on $J_{i,k-1}$, $k \ge 1$, $i = 1, \dots, n$. Let K_i be the random number of uncensored transitions of the process (J_i, T_i), $X_i^* = t_i - \sum_{k=1}^{K_i} X_{i,k}$ the last (censored) duration time; if $J_i^* = J_{i,K_i}$ is censored, $\delta_i^* = \delta_{i,K_i+1} = 0$, otherwise $X_i^* = 0$.

The model is defined by the transition functions from j to j', $F_{j'|j}(x) = P(X_{i,k} \le x, J_{i,k} = j' | J_{i,k-1} = j)$ or, equivalently, by the transition probabilities $p_{j'|j} = P(J_{i,k} = j' | J_{i,k-1} = j)$ and the distributions of the sojourn times between two states j and j', $F_{|jj'}(x) = P(X_{i,k} \le x | J_{i,k} = j', J_{i,k-1} = j)$. The distribution of a sojourn time in j is $F_j = \sum_{j'} F_{j'|j}$; the related survival functions are denoted $S_{|jj'}$ and S_j, and $S_{j'|j}(x) = p_{j'|j} - F_{j'|j}(x)$. The censoring variable of the sojourn times in state j has a distribution function G_j. These functions are all assumed to be continuous.

Let $N(j,n)$ be the total number of arrivals in state j, $Y^{nc}(x,j,j',n)$ the total number of sojourn times larger than x before a transition from j to j', $Y^{nc}(x,j,n)$ (resp. $Y^c(x,j,n)$) the total number of uncensored (resp. censored) sojourn times larger than x in j and $Y(x,j,n) = Y^{nc}(x,j,n) + Y^c(x,j,n)$. As in (1), the nonparametric maximum likelihood estimator $\widehat{S}_{n,j}$ of the survival function S_j in state j may be defined as a solution of the self-consistency equation

$$\widehat{S}_{n,j}(x) = \frac{1}{N(j,n)} \left\{ Y(x^+, j, n) + \sum_{i=1}^n (1 - \delta_i^*) 1_{\{J_i^* = j\}} 1_{\{X_i^* \le x\}} \frac{\widehat{S}_{n,j}(x)}{\widehat{S}_{n,j}(X_i^*)} \right\}, \tag{4}$$

with $\widehat{S}_{n,j}(0) = 1$ and (4) determines the Kaplan-Meier estimator of S_j.

For the estimation of $S_{j'|j}$ and $S_{|jj'}$, we assume that the mean number of visits in j, $\pi_j^0 = n^{-1} EN(j,n)$, is finite and $\pi_j^0 p_{jj'} > 0$. Estimators $\widehat{S}_{n,j'|j}$ and $\widehat{S}_{n,|jj'}$ are unique solutions of

$$\begin{aligned} \widehat{S}_{n,j'|j}(x) = \frac{1}{N(j,n)} \Bigg\{ & Y^{nc}(x^+, j, j', n) + \sum_{i=1}^n (1 - \delta_i^*) 1_{\{J_i^* = j\}} 1_{\{X_i^* > x\}} \frac{\widehat{S}_{n,j'|j}(X_i^*)}{\widehat{S}_{n,j}(X_i^*)} \\ & + \sum_{i=1}^n (1 - \delta_i^*) 1_{\{J_i^* = j\}} 1_{\{X_i^* \le x\}} \frac{\widehat{S}_{n,j'|j}(x)}{\widehat{S}_{n,j}(X_i^*)} \Bigg\}, \end{aligned} \tag{5}$$

$$\begin{aligned} \widehat{S}_{n,|jj'}(x) = \frac{1}{N(j,n)} \Bigg\{ & \frac{Y^{nc}(x^+, j, j', n)}{\widehat{p}_{n,jj'}} + \sum_{i=1}^n (1 - \delta_i^*) 1_{\{J_i^* = j\}} 1_{\{X_i^* > x\}} \frac{\widehat{S}_{n,|jj'}(X_i^*)}{\widehat{S}_{n,j}(X_i^*)} \\ & + \sum_{i=1}^n (1 - \delta_i^*) 1_{\{J_i^* = j\}} 1_{\{X_i^* \le x\}} \frac{\widehat{S}_{n,|jj'}(x)}{\widehat{S}_{n,j}(X_i^*)} \Bigg\} \end{aligned} \tag{6}$$

where $\widehat{p}_{n,jj'} = \widehat{S}_{n,j'|j}(0)$. The estimators $\widehat{S}_{n,j'|j}$ and $\widehat{S}_{n,|jj'}$ solutions of equations (5) and (6) are defined from $\widehat{S}_{n,j'|j}(0) = \widehat{p}_{n,jj'}$ and $\widehat{S}_{n,|jj'}(0) = 1$. They are decreasing step functions with jumps at the observed durations before a transition from j to j' and their variations depend on the number of such transitions and on the number of censored durations in state j. The censored durations in j, before and after x, are dispatched onto all the observed duration times in j before a transition to another state according to weights depending on the previously calculated values of $\widehat{S}_{n,j'|j}$ (resp. $\widehat{S}_{n,|jj'}$) and $\widehat{S}_{n,j}$.

We denote $(X_{(1)} < X_{(2)} < \ldots)$ the ordered sample $((X_{i,1}, \ldots, X_{i,K_i}), X_i^*)_{i \leq n}$ and $\delta_{(l)}$ the indicator related to $X_{(l)}$,

$$\widehat{S}_{n,j'|j}(X_{(l-1)}) - \widehat{S}_{n,j'|j}(X_{(l)}) = \frac{Y^{nc}(X_{(l-1)}, j, j', n) - Y^{nc}(X_{(l)}, j, j', n)}{N(j,n) + \int_0^{X_{(l-1)}} \widehat{S}_{n,j}^{-1}(y)\, dY^c(y^+, j, n)}.$$

(5) defines the Kaplan-Meier estimator for $S_{j'|j}$ studied by [Gill, 1980] and $\widehat{S}_{n,|jj'}(x) = \widehat{p}_{n,jj'}^{-1} \widehat{S}_{n,j'|j}(x)$.

A self-consistency equation and a direct estimator of $p(j'|x,j) = P(J_{k,i} = j'|X_{k,i} \geq x, J_{k-1,i} = j)$

$$\widehat{p}_n(j'|x,j) = Y^{-1}(x^+, j, n)\{Y^{nc}(x^+, j, j', n) - \int_{x^+}^{\infty} \widehat{p}_n(j'|y,j)\, dY^c(y,j,n)\}. \tag{7}$$

Equation (7) defines an estimator of $p(j'|x,j)$ as a decreasing step function with jumps at the censored durations in j and at the uncensored durations related to transitions from j to j'. Starting from $\widehat{p}_n(j'|\infty, j) = 0$,

$$\widehat{p}_n(j'|X_{(l-1)}, j) = \widehat{p}_n(j'|X_{(l)}, j) + \frac{1 - \delta_{(l)}\widehat{p}_n(j'|X_{(l)}, j)}{Y^c(X_{(l)}, j, n)}.$$

4 Self-consistent estimation for observations by intervals

4.1 Doubly censored observations

For individual i, the k-th sojourn time $X_{i,k}$ of the process is observed on an interval $[U_{i,k}, C_{i,k}]$ with $U_{i,k} \leq C_{i,k}$ and $\cup_{0 \leq k \leq K_i}[T_{i,k} + U_{i,k}, T_{i,k} + C_{i,k}] \subset [0, t_i]$. The observations are $J_{i,k-1}$, $W_{i,k} = \max\{U_{i,k}, \min(X_{i,k}, C_{i,k})\}$, $\delta_{1,i,k} = 1_{\{X_{i,k} > U_{i,k}\}}$ and $\delta_{2,i,k} = 1_{\{X_{i,k} < C_{i,k}\}}$, and $X_{i,k}$ is observed only if $\delta_{1,i,k}\delta_{2,i,k} = 1$. We assume that the variables $U_{i,k}$ and $C_{i,k}$ are independent of $X_{i,k}$, with continuous d.f. H_j and G_j such that $\tau_{1,j} = \inf\{u; H_j(u) > 0\} = 0$ and $\tau_{2,j} = \sup\{u; S_j(u)\bar{G}_j(u) > 0\} = \infty$. For $x \geq \tau_{1,n,j}$, the notations of section 3 are modified as

$$
\begin{aligned}
Y^{nc}(x,j,j',n) &= \sum_{i=1}^{n}\sum_{k=1}^{K_i} \delta_{1,i,k}\delta_{2,i,k} 1_{\{J_{i,k}=j'\}} 1_{\{J_{i,k-1}=j\}} 1_{\{X_{i,k}\geq x\}},\\
Y^{c,1}(x,j,n) &= \sum_{i=1}^{n}\sum_{k=1}^{K_i} (1-\delta_{1,i,k}) 1_{\{J_{i,k-1}=j\}} 1_{\{U_{i,k}\leq x\}},\\
Y^{c,2}(x,j,n) &= \sum_{i=1}^{n}\sum_{k=1}^{K_i} (1-\delta_{2,i,k}) 1_{\{J_{i,k-1}=j\}} 1_{\{C_{i,k}\geq x\}},\\
\widehat{Y}^{c,1}(x,j,j',n) &= \int_0^x \frac{1-\widehat{S}_{n,j'|j}(y)}{1-\widehat{S}_{n,j}(y)}\, dY^{c,1}(y,j,n),\\
\widehat{Y}^{c,2}(x,j,j',n) &= -\int_x^\infty \frac{\widehat{S}_{n,j'|j}(y)}{\widehat{S}_{n,j}(y)}\, dY^{c,2}(y,j,n),
\end{aligned}
$$

$\widehat{Y}(x,j,j',n) = Y^{nc}(x,j,j',n) + \widehat{Y}^{c,1}(x,j,j',n) + \widehat{Y}^{c,2}(x,j,j',n)$ and $Y(x,j,n) = \sum_{j'} Y^{nc}(x,j,j',n) + Y^{c,1}(x,j,n) + Y^{c,2}(x,j,n)$.
The self-consistency equation for the estimator $\widehat{S}_{n,j'|j}$ of $S_{j'|j}$ is written as

$$
\begin{aligned}
N(j,n)\widehat{S}_{n,j'|j}(x) = \widehat{Y}(x^+,j,j',n) &- \widehat{S}_{n,j'|j}(x)\int_0^x \frac{dY^{c,2}(j,n)}{\widehat{S}_{n,j}} \\
&+\{1-\widehat{S}_{n,j'|j}(x)\}\int_x^\infty \frac{dY^{c,1}(j,n)}{1-\widehat{S}_{n,j}}
\end{aligned} \tag{8}
$$

A sum over index j' gives an equation for $\widehat{S}_{n,j}$

$$
\begin{aligned}
\widehat{S}_{n,j}(x) = \frac{1}{N(j,n)}[Y(x^+,j,n) &- \widehat{S}_{n,j}(x)\int_0^x \frac{dY^{c,2}(j,n)}{\widehat{S}_{n,j}} \\
&+\{1-\widehat{S}_{n,j}(x)\}\int_x^\infty \frac{dY^{c,1}(j,n)}{1-\widehat{S}_{n,j}}],
\end{aligned}
$$

with $\widehat{S}_{n,j}(0) = 1$. This equation provides an algorithm for a decreasing estimator $\widehat{S}_{n,j}$ starting from $\widehat{S}_{n,j}(0) = 1$ and with jumps at the uncensored transitions times. Let $(W_{(1)} < W_{(2)} < \ldots)$ the ordered sample of the variables $W_{i,k}$, $k = 1,\ldots,K_i$, $i = 1,\ldots,n$ and let $\delta_{(l)}$, $\delta_{1,(l)}$ and $\delta_{2,(l)}$ the indicators related to $X_{(l)}$, then

$$
\begin{aligned}
\widehat{S}_{n,j}(W_{(l)}) &= \widehat{S}_{n,j}(W_{(l-1)}) - \frac{Y^{nc}(X_{(l-1)},j,n) - Y^{nc}(X_{(l)},j,n)}{d_{n,j,(l)}}, \text{ with}\\
d_{n,j,(l)} &= N(j,n) + \int_0^{W_{(l-1)}} \widehat{S}_{n,j}^{-1}(y)\, dY^{c,2}(y^+,j,n)\\
&\quad + \int_{W_{(l)}}^\infty (1-\widehat{S}_{n,j}(y))^{-1}\, dY^{c,1}(y^+,j,n).
\end{aligned}
$$

Since $U_{i,k} \leq C_{i,k}$, boundary constraints are $\widehat{H}_{n,j}(\infty) = 1 \geq \widehat{G}_{n,j}(\infty)$, $\widehat{G}_{n,j}(0) = 0 \leq \widehat{H}_{n,j}(0)$ and (8) uniquely defines estimators of $S_{j'|j}$, H_j, $\bar{G}_j$ and $p_{jj'}$,

$$\widehat{p}_{n,jj'} = \{N(j,n) + \int_0^\infty \frac{dY^{c,1}(j,n)}{1-\widehat{S}_{n,j}}\}^{-1}\{\widehat{Y}(0,j,j',n) + \int_0^\infty \frac{dY^{c,1}(j,n)}{1-\widehat{S}_{n,j}}\},$$

and, starting from $\widehat{S}_{n,j'|j}(0) = \widehat{p}_{n,jj'}$, $\widehat{S}_{n,j'|j}$ is a decreasing step-function with jumps at the uncensored transitions times,

$$\widehat{S}_{n,j'|j}(W_{(l)}) = \widehat{S}_{n,j'|j}(W_{(l-1)}) - \frac{Y^{nc}(X_{(l-1)},j,j',n) - Y^{nc}(X_{(l)},j,j',n)}{d_{n,j,(l)}}.$$

4.2 Left-truncated and right-censored observations

The k-th transition $X_{i,k}$ of the process for an individual i is now observed on an interval $[U_{i,k}, C_{i,k}]$, conditionally on $X_{i,k} \wedge C_{i,k} > U_{i,k}$. The variables $U_{i,k}$ and $C_{i,k}$ are only supposed to be independent and independent of $X_{i,k}$ but without $U_{i,k} < C_{i,k}$ and all the observations of the states and the duration times are missing for the transitions with $X_{i,k} \wedge C_{i,k} \leq U_{i,k}$. The nonparametric estimators of the survival functions are now defined from the counting processes

$$\begin{aligned}
Y^{nc,nt}(x,j,j',n) &= \sum_{i=1}^n \sum_{k=1}^{K_i} \delta_{i,k} 1_{\{J_{i,k}=j'\}} 1_{\{J_{i,k-1}=j\}} 1_{\{U_{i,k}<x\leq X_{i,k}\}}, \\
Y^{c,nt}(x,j,n) &= \sum_{i=1}^n \sum_{k=1}^{K_i} (1-\delta_{i,k}) 1_{\{J_{i,k-1}=j\}} 1_{\{U_{i,k}<x\leq C_{i,k}\}}, \\
N^{c,nt}(x,j,n) &= \sum_{i=1}^n \sum_{k=1}^{K_i} (1-\delta_{i,k}) 1_{\{J_{i,k-1}=j\}} 1_{\{U_{i,k}<C_{i,k}\leq x\}}, \\
\widehat{Y}^{c,nt}(x,j,j',n) &= -\int_x^\infty \frac{\widehat{S}_{n,j'|j}(y)}{\widehat{S}_{n,j}(y)}\, dY^{c,nt}(y,j,n), \\
Y_n(x,j) &= \sum_{j'} Y^{nc,nt}(x,j,j',n) + Y^{c,nt}(x,j,n).
\end{aligned}$$

Self-consistency equations may be written for $\widehat{H}_{n,j}$, $\widehat{H}_{n,j}\widehat{S}_{n,j}$ and $\widehat{H}_{n,j}\widehat{S}_{n,j'|j}$,

$$\widehat{H}_{n,j}(x) = n^{-1}\{Y_n(x^+, j) + \sum_{i=1}^{n} 1_{\{U_{i,k}<X_{i,k}\leq x\}} 1\{J_{i,k-1} = j\} \frac{\widehat{H}_{n,j}(x)}{\widehat{H}_{n,j}(X_{i,k})}\},$$

$$\widehat{H}_{n,j}(x)\widehat{S}_{n,j}(x) = \frac{1}{N(j,n)}\{Y_n(x^+, j) + \widehat{H}_{n,j}(x)\widehat{S}_{n,j}(x)\int_0^x \frac{dN^{c,nt}(j,n)}{\widehat{S}_{n,j}\widehat{H}_{n,j}}\},$$

$$\widehat{H}_{n,j}(x)\widehat{S}_{n,j'|j}(x) = \frac{1}{N(j,n)}\{Y^{nc,nt}(x,j,j',n) + \widehat{Y}^{c,nt}(x,j,j',n) + \widehat{H}_{n,j}(x)\widehat{S}_{n,j'|j}(x)\int_0^x \frac{dN^{c,nt}(j,n)}{\widehat{S}_{n,j}\widehat{H}_{n,j}}\}.$$

Let $(U_{(1)} < U_{(2)} < \ldots)$ and respectively $(X_{(1)} < X_{(2)} < \ldots)$ be the ordered sample of the variables $U_{i,k}$ and $X_{i,k}$, $k = 1, \ldots, K_i$, $i = 1, \ldots, n$ and let $\delta_{(l)}$ be the indicator related to $X_{(l)}$, $R_U(l)$ and $R_X(l)$ be the ranks of $U(l)$ and $X(l)$. The nonparametric estimator of $H_j(t) = \exp\{-\int_t^\infty H_j^{-1}\, dH_j\}$ may be defined as an increasing step function with jumps at the observations $U_{i,k}$ such that $J_{i,k-1} = j$, with $\widehat{H}_{n,j}(0) = 0$ and

$$\widehat{H}_{n,j}(U_{(l+1)}) = \widehat{H}_{n,j}(U_{(l)}) + \frac{1\{U_{(l+1)} \leq X_{R_U(l+1)}\}1\{J_{R_U(l+1)} = j\}}{n - \sum_{l'} 1\{X_{l'} < U_{l'} \leq U_{(l)}\}\widehat{H}_{n,j}^{-1}(U_{l'})}.$$

$\widehat{S}_{n,j}$ is deduced as a step function with jumps at the $X_{i,k}$ such that $J_{i,k-1} = j$, with $\widehat{S}_{n,j}(0) = 1$ and satisfying

$$(\widehat{H}_{n,j}\widehat{S}_{n,j})(X_{(l)}) = (\widehat{H}_{n,j}\widehat{S}_{n,j})(X_{(l-1)}) - \frac{\delta_{(l)}1\{J_{R_X(l)} = j\}1\{U_{R_X(l)} < X_{(l)}\}}{n - \sum_{l'=1}^{n}(1-\delta_{l'})1\{U_{l'} < X_{l'} \leq X_{(l-1)}\}(\widehat{H}_{n,j}\widehat{S}_{n,j})^{-1}(X_{l'})}.$$

An explicit expression of $\widehat{S}_{n,j'|j}$ is similar.

All the proposed estimators are all uniformly consistent on compact sets included in the support of the survival functions.

References

[Chang and Yang, 1987]N.M. Chang and G.L. Yang. Strong consistency of a nonparametric estimator of the survival function with doubly censored data. *Ann. Statist.*, pages 1536–1547, 1987.

[Gill, 1980]R. Gill. Nonparametric estimation based on censored observations of a markov renewal process. *Z. Wahrsch. verw. Gebiete*, pages 97–116, 1980.

[Woodroof, 1985]M. Woodroof. Estimating a distribution function with truncated data. *Ann. Statist.*, pages 163–177, 1985.