# Simultaneous marginal survival estimators when doubly-censored data is present

**Guadalupe Gómez**
Departament d'Estadística i I.O.
Universitat Politècnica de Catalunya
Jordi Girona 1–3, 08034 Barcelona, Spain
*lupe.gomez@upc.edu*

**Olga Julià**
Departament de Probabilitat, Lògica i Estadística.
Universitat de Barcelona.
Gran Via 585, 08007 Barcelona, Spain.
*olgajulia@ub.edu*.

## Abstract

A doubly censoring scheme occurs when the lifetimes being measured are censored either from above or below. This scheme is found both in reliability and in survival studies. Denote by $T$ the variable of interest and by $L$ and $R$, with $L < R$, the censoring variables. We propose new nonparametric simultaneous marginal survival estimators for the laws of $T$, $L$ and $R$, based on an inverse-probability-of-censoring approach.

## 1 Introduction

When the lifetimes being measured are censored either from above or below, and as a result, left- and right-censored observations as well as exact values are present for the same data set, we are in a doubly-censored setting. An instance of such data, which is analyzed at the end of the paper, is found when we are interested in the time period from the start of IV drug use to AIDS diagnosis in a cohort of HIV-infected drug users (Langohr, Gómez and Muga (2004)). Instances of this scheme might also be found in reliability studies where several components are to be assembled to have a system operating and some of them have to be redundant. If we are interested in the marginal lifetimes $T_A$, $T_B$ and $T_C$ of components A, B and C which are assembled such that A and B are in parallel and C is in series with this parallel system, then $T_A$ is exactly observed if A is the cause of failure, left-censored by $\min\{T_B, T_C\}$ and right-censored by $T_C$ if C fails before A.

Most of the non-parametric estimators for $S_T(t)$ when doubly censored data is present are based on Turnbull's pioneering paper (1974) who develops a self-consistent estimator which is proved to be the non-parametric maximum likelihood estimator for the underlying survival, assuming that data have been grouped. Several authors have studied the asymptotic properties of the self-consistent estimator, among others, Chang and Yang (1987) and Chang (1990). Satten and Datta (2001) use an inverse-probability-of-censoring representation for the Kaplan-Meier estimator and has inspired our method.

Let $T$ be a positive random variable whose survival function is $S_T(t) = P(T > t)$. The doubly censoring situation could be formally expressed with the aid of two positive random variables, $L$ and $R$, with $L < R$ a.s., such that $T$ would be exactly observed only if $L \leq T \leq R$. Random interval $[L, R]$ plays an important role in this problem, not only because it is the window within which $T$ is exactly observed, but also because reasonable estimators for $S_T(t)$, would heavily depend on the survival functions of $R$ and $L$, $S_L(t)$ and $S_R(t)$, respectively. If we assume that $[L, R]$ is independent of $T$, but let the joint distribution of $L$ and $R$ be arbitrary, it follows that the probability of $T$ being observed at time $t$ will be given by $S_R(t^-) - S_L(t)$. Analogously, the probability of $L$ and $R$ being observed at times $l$ and $r$ will be given by $1 - S_T(l^-)$ and $S_T(r)$, respectively. We assume that $S_T(0) = S_L(0) = 1$ and $S_T(\infty) = S_L(\infty) = S_R(\infty) = 0$.

Given $n$ individuals, the triplet of independent and identically random variables $\{(T_i, L_i, R_i), \quad i = 1 \ldots, n\}$ forms the complete data set. The observable data consists of the pairs $\{(U_i, \delta_i), \quad i = 1 \ldots, n\}$, where $U_i = \min\{\max\{L_i, T_i\}, R_i\} = (L_i \vee T_i) \wedge R_i$ and $\delta_i = \mathbb{1}_{\{T_i > R_i\}} - \mathbb{1}_{\{T_i < L_i\}}$. Let $u_{(1)} \leq u_{(2)} \leq \cdots \leq u_{(n)}$ be the order statistic corresponding to the observed sample $u_1, u_2, \cdots u_n$ and define $\delta_{(1)}, \delta_{(2)}, \cdots, \delta_{(n)}$ as the corresponding censoring values. Whenever ties are present, we consider left-censored observations as preceding exact observations, which in turn precede right-censored observations. Let $o_j, \ j = 1 \ldots r$ be the distinct ordered failure and censoring times among $u_{(1)} \leq u_{(2)} \leq \cdots \leq u_{(n)}$ and use $d_j$, $\mu_j$ and $\lambda_j$ to denote the number of exact, left-censored and right-censored at $o_j$.

We propose simultaneous marginal nonparametric inverse-weighted survival estimators (IWSE) for $S_T$, $S_L$ and $S_R$. The novelty of our approach is twofold. On one hand, it provides a new characterization of the self-consistent survival estimator in terms of the sizes of its jumps. On the other, it allows to estimate the marginal laws of the censoring random variables. The IWSE we propose are self-consistent estimators, strongly consistent and asymptotically normal, generalize the empirical survival function and reduce to the Kaplan-Meier estimator if the observed data are only right-censored. Since the characterization we propose for the estimators is in terms of the sizes of their jumps, it allows to construct the estimator recursively. Our estimator is computationally straightforward and not computationally intensive.

## 2 Inverse weighted joint survival estimators for $S_T$, $S_L$ and $S_R$.

**Definition 1.** We define the inverse weighted survival estimators, IWSE, of $S_T(t)$, $S_L(t)$ and $S_R(t)$ as the estimators $\hat{S}_T(t)$, $\hat{S}_L(t)$ and $\hat{S}_R(t)$ which take values between 0 and 1 and are the solution of the following system:

1. For all $t \in [o_1, o_r)$,
$$\left.\begin{array}{rcl} \hat{S}_T(t) &=& \frac{1}{n}\sum_{j:o_j>t}^{r} \frac{d_j}{\hat{S}_R(o_j^-)-\hat{S}_L(o_j)} + \hat{S}_T(o_r), \\ \hat{S}_L(t) &=& \frac{1}{n}\sum_{j:o_j>t}^{r} \frac{\mu_j}{1-\hat{S}_T(o_j^-)}, \\ \hat{S}_R(t) &=& 1 - \frac{1}{n}\sum_{j:o_j\le t}^{r} \frac{\lambda_j}{\hat{S}_T(o_j)}. \end{array}\right\} \quad (1)$$

2. For all $t \in (0, o_1)$, expression (1) is also valid except when the first exact event precedes the first left-censored observation, in which case $\hat{S}_T(t) = 1$, or when the first left-censored observation precedes the first exact event, in which case $\hat{S}_L(t) = 1$.

3. For all $t \in [o_r, \infty)$, expression (1) is also valid except when the last exact event follows the last right-censored observation, in which case $\hat{S}_T(t) = 0$, or when the last right-censored observation follows the last exact event, in which case $\hat{S}_R(t) = 0$.

Given $\hat{S}_T(t)$, we define $f_j = \hat{S}_T(o_j^-) - \hat{S}_T(o_j)$, $1 \le j \le r$, $f_0 = 1 - \hat{S}_T(o_1^-)$ and $f_{r+1} = \hat{S}_T(o_r)$. Furthermore, given $\hat{S}_L(t)$ and $\hat{S}_R(t)$, we take $h_j = \hat{S}_L(o_j^-) - \hat{S}_L(o_j)$ and $g_j = \hat{S}_R(o_j^-) - \hat{S}_R(o_j)$.

**Theorem 1.** *An estimator $\hat{S}_T(t)$ is the IWSE of $\hat{S}_T(t)$ if and only if it is a right-continuous step function for which jumps are possible only at $o_j$ and $\{f_0, f_1, \cdots, f_{r+1}\}$ are determined by the following conditions:*

1. *For all $0 \le j \le r+1$, $0 \le f_j \le 1$ and $\sum_{j=0}^{r+1} f_j = 1$.*
2. *$f_j = 0$ if and only if $d_j = 0$ for all $1 \le j \le r$.*
3. *For any two consecutive death times $o_i < o_j$, that is, when $d_i > 0$, $d_j > 0$, and $d_k = 0$ if $i < k < j$, we have*
$$\frac{d_i}{f_i} - \frac{d_j}{f_j} = \frac{\sum_{i\le k<j} \lambda_k}{\sum_{m=j}^{r+1} f_m} - \frac{\sum_{i<k\le j} \mu_k}{1 - \sum_{m=j}^{r+1} f_m}. \quad (2)$$
4. *(a) If $\delta_{(n)} = 0$ then $f_{r+1} = 0$ and $\sum_1^r g_j = 1 - \frac{d_r}{n f_r}$.*
   *(b) If $\delta_{(n)} = 1$ and $o_\tau$ is the last time of death ($d_\tau > 0$ and $d_k = 0$ for $k > \tau$), then $\sum_1^r g_j = 1$ and*
$$\frac{d_\tau}{f_\tau} = \frac{\sum_{k=\tau}^{r} \lambda_k}{f_{r+1}} - \frac{\sum_{k=\tau+1}^{r} \mu_k}{1 - f_{r+1}}, \quad (3)$$
   *(c) If $\delta_{(n)} = -1$, then we inspect the value of the preceding $\delta_{(k)}$ until we find one whose value is $\neq -1$. We then apply (4a) and (4b).*
5. *(a) If $\delta_{(1)} = 0$, then $f_0 = 0$ and $\sum_1^r h_j = 1 - \frac{d_1}{n f_1}$.*
   *(b) If $\delta_{(1)} = -1$ and $o_\nu$ is the first time of death ($d_\nu > 0$ and $d_k = 0$ for $k < \nu$), then $\sum_1^r h_j = 1$ and*
$$\frac{d_\nu}{f_\nu} = \frac{\sum_{k=1}^{\nu+1} \mu_k}{f_0} - \frac{\sum_{k=1}^{\nu} \lambda_k}{1 - f_0}, \quad (4)$$

(c) *If $\delta_{(1)} = 1$, then we inspect the value of the next $\delta_{(k)}$ until we find one whose value is $\neq 1$. We then apply (5a) and (5b).*

6. *The estimators $\hat{S}_L(t)$ and $\hat{S}_R(t)$ are the inverse-weighted survival estimators if and only if they are right-continuous step functions for which jumps are only possible at $o_j$ and the sizes of these jumps are given by*

$$h_j = \frac{\mu_j}{n(1 - \sum_{k=j}^{r} f_k)} \quad \text{and} \quad g_j = \frac{\lambda_j}{n \sum_{k=j+1}^{r} f_k} \cdot \tag{5}$$

**Corollary 1.** *1. If there are no left-censored data, the inverse weighted survival estimator $\hat{S}_T$ coincides with the Kaplan-Meier estimator for almost all $t < o_r$.*

*2. If there are no right-censored data, the inverse weighted survival estimator $\hat{S}_T$ coincides with the Kaplan-Meier estimator for left-censored data for almost all $t > o_1$ (Gómez, Julià and Utzet, 1992).*

## 3 Algorithm

In order to obtain the survival estimators $\hat{S}_T(t) = \sum_{j:o_j>t} f_j, \hat{S}_L(t) = \sum_{j:o_j>t} h_j, \hat{S}_R(t) = 1 - \sum_{j:o_j \leq t} g_j$, we proceed using Theorem 1 and solving (3) iteratively to obtain $f_j$. Equation (5) is used to get $\hat{S}_L(t)$ and $\hat{S}_R(t)$. Recall that $o_\nu$ is the first exact time and $o_\tau$ is the last exact time.

1. Whenever the first and the last times are exact ($\nu = 1$, $\tau = r$), we use iteratively equation (3) to obtain $f_j$ for $1 \leq j \leq r$.

2. $\nu > 1$ means that prior to an exact time, we have left and/or right-censored observations. In these cases the problem reduces to setting 1.

   (a) If we only have right-censored observations, they can be discarded and reduced to setting 1 with the corresponding reduced sample size.

   (b) If we have either only left or both left and right -censored observations and $\delta_{(1)} = -1$, then we substitute $\delta_{(1)} = -1$ by $\delta_{(1)} = 0$, proceed as in setting 1, compute all the $f_j$ for $1 \leq j \leq r$, and reinterpret $f_1$ as $f_0$, that is, as the Prob$\{T < o_\nu\}$.

   (c) If we have both left and right -censored observations and $\delta_{(1)} = 1$, then we exchange $\delta_{(1)} = 1$ by $\delta_{(1)} = -1$, and the first $\delta = -1$ by $\delta = 1$ and proceed as in 2b.

3. $\tau < r$ means that after the last exact time, we have left and/or right-censored observations. These cases reduce to setting 1 following an analogous argument as in 2.

## 4 Properties of the inverse weighted survival estimator

We prove the unicity of the IWSE and the equivalence between a self-consistent estimator and an IWSE. Since a self-consistent survival estimator always exists for the doubly censored scheme (Tsai and Crowley (1985)), we conclude that the IWSE has the same asymptotic behavior that the self-consistent estimator, that is, the IWSE is strongly consistent and asymptotically normal.

**Theorem 2.** *If $\hat{U}(t)$ and $\hat{S}(t)$ are two inverse weighted survival estimators in the sense of Definition 1, then $\hat{U}(t) = \hat{S}(t)$ for all $t$.*

**Theorem 3.** *A survival estimator $\bar{S}_T^n(t)$ of $S_T(t)$ based on the observable data $\{(U_i, \delta_i) \quad i = 1 \ldots, n\}$ and given by (6) is self-consistent if and only if $\bar{S}_T^n(t)$ is an inverse weighted survival estimator as defined in Definition 2.*

$$\bar{S}_T^n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[u_i>t]} + \frac{1}{n} \sum_{i=1}^{n} \frac{\bar{S}_T^n(t)}{\bar{S}_T^n(u_i)} \mathbb{1}_{[u_i \leq t,\ \delta_i=1]} - \frac{1}{n} \sum_{i=1}^{n} \frac{1 - \bar{S}_T^n(t)}{1 - \bar{S}_T^n(u_i^-)} \mathbb{1}_{[u_i>t, \delta_i=-1]}. \tag{6}$$

## 5 Illustration

The German Trias i Pujol Hospital, located in Badalona (Spain), has been running a detoxification program for intravenous drug users since 1985 and recording the date of several events, such as the day the patients started injecting drugs, the day on which they were diagnosed as having AIDS, if available, the time of death and whether or not they had AIDS when they died. Among the 232 patients who were infected with HIV, 136 (59%) of them left the center AIDS-free, thus contributing right-censored observations; 14 (6%) died with AIDS with no prior recording of the time to AIDS diagnosis (these form the subcohort of left-censored observations); and the remaining 82 (35%) had been diagnosed as having AIDS while following the program. Let $T$ be the number of years elapsed from becoming addicted to IV drugs to being diagnosed as having AIDS, among HIV-infected patients, $L$ stands for the number of years elapsed between the start of drug use and the death with AIDS with no prior AIDS diagnosis, and $R$ stands for the number of years elapsed from the start of drug use to leaving the detoxication program being AIDS-free. The IWSE for the survival for the number of years to AIDS diagnosis is plotted in Figure 1. This curve shows that the median number of years to AIDS diagnosis is 15.44 and that after 10 years 70% of IV drug users are AIDS-free. It can also be observed that because of the right-censored data, percentiles larger than 40% cannot be estimated. It is possible to estimate the survival for $R$ along the entire real line as a consequence of a right-censored observation as the maximum of all the failure and censoring points.

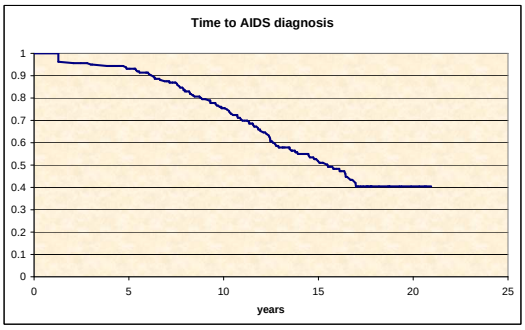


Figure 1: The IWSE of the elapsed number of years from starting the IV-addiction to being AIDS-diagnosed for HIV-infected patients

## Acknowledgements

This work is partially supported by grant MTM2005–08886 from the Ministerio de Ciencia y Tecnología.

## References

[1] Chang, M.N. (1990). Weak convergence of a self-consistent estimator of the survival function with doubly censored data. *Ann. Statist.* **2**, 437–453.

[2] Chang, M.N. and Yang, G.L. (1987). Strong consistency of a non-parametric estimator of the survival function with doubly censored data. *Ann. Statist.* **15**, 1536–1547.

[3] Gómez, G. Julià, O. and Utzet, F. (1992). Survival Analysis for Left Censored Data. *Survival Analysis: State of the Art* J.P. Klein and P.K. Goel (ed.) 269-298, Kluver, Dordretcht.YES

[4] Langohr, K. Gómez, G. and Muga, R. (2004) A parametric survival model with an interval-censored covariate. *Statistics in Medicine* **23**, 3159-3175.

[5] Satten, G.A. and Datta, S. (2001). The Kaplan-Meier Estimator as an Inverse-Probability-of-Censoring Weighted Average. *The American Statistician* **55**, No. 3, 207–210.

[6] Tsai and Crowley (1985). A large sample study of generalized maximum likelihood estimators from incomplete data via self-consistency. *Ann. Statist.* **13**, No. 4, 1317–1334.

[7] Turnbull, B.W. (1974). Non-parametric Estimation of a Survivorship Function with Doubly Censored Data. *Journal of the American Statistical Association* **69**, No. 345, 169–173.