Product-limit estimators of the survival function with left or right censored data

Valentin Patilea[1] and Jean-Marie Rolin[2]

[1] CREST-ENSAI
Campus de Ker-Lann
Rue Blaise Pascal - BP 37203
35172 Bruz cedex, France
(e-mail: patilea@ensai.fr)
[2] Institut de Statistique
Université Catholique de Louvain
20, voie du Roman Pays
1348 Louvain-la-Neuve, Belgique
(e-mail: rolin@stat.ucl.ac.be)

Abstract. The problem of estimating the distribution of a lifetime when data may be left or right censored is considered. Two models are introduced and the corresponding product-limit estimators are derived. Strong uniform convergence and asymptotic normality are proved for the product-limit estimators on the whole range of the observations. A bootstrap procedure that can be applied to confidence intervals construction is proposed.
Keywords: bootstrap, delta-method, martingales, left and right censoring, strong convergence, weak convergence.

1 Introduction

A great deal of recent attention in survival analysis has focused on estimating the survivor distributions in the presence of various and complex censoring mechanisms. The goal of this paper is to analyze simple models for lifetime data that may be left or right censored. Typically, a lifetime T is left or right censored if, instead of observing T we observe a finite nonnegative random variable Y, and a discrete random variable with values 0, 1 or 2. By definition, when $A = 0$, $Y = T$, when $A = 1$, $Y < T$ and, when $A = 2$, $Y > T$. Models for left or right censored data were proposed by [Turnbull, 1974], [Sampath and Chandra, 1990] and [Huang, 1999]. See also [Gu and Zhang, 1993], [Kim, 1994].

Assume that the sample consists of n independent copies of (Y, A) and let F_T be the distribution of the lifetime of interest T. Using the plug-in (or substitution) principle, the nonparametric estimation of F_T is straight as soon as F_T can be expressed as an explicit function of the distribution of (Y, A). The existence of such a function requires a precise description of the censoring mechanism that is generally achieved by introducing 'latent' variables and by making assumptions on their distributions. In this paper,

two latent models allowing for explicit inversion formula, that is closed-form function relating F_T to the distribution of (Y, A), are proposed.

In some sense, our first latent model lies between the classical right-censorship model and the current-status data model. It may be applied to the following framework. Consider a study where T the age at onset for a disease is analyzed. The individuals are examined only one time and they belong to one of the following categories: (i) evidence of the disease is present and the age at onset is known (from medical records, interviews with the patient or family members, ...); (ii) the disease is diagnosed but the age at onset is unknown or the accuracy of the information about this is questionable; and (iii) the disease is not diagnosed at the examination time. Let C denote the age of the individual at the examination time. In the first case the exact failure time T (age at onset) is observed, that is $Y = T$. In case (ii) the failure time T is left-censored by C and thus $Y = C$, $A = 2$. Finally, the onset time T is right-censored by C for the individuals who have not yet developed the disease; in this case $Y = C$, $A = 1$. If no observation as in (ii) occurs, we are in the classical right-censorship framework, while if no uncensored observation is recorded we have current-status data. Our first latent model can be applied, for instance, with the data sets analyzed by [Turnbull and Weiss, 1978].

The second latent model proposed is closely related to the first one. It lies between the left-censorship model and the current-status data model. Consider the example of a reliability experiment where the failure time of a type of device is analyzed. A sample of devices is considered and a single inspection for each device in the sample is undertaken. Some of them already failed without knowing when (left censored observations). To increase the precision of the estimates, a proportion of the devices still working is selected randomly and followed until failure (uncensored observations). For the remaining working devices the failure time is right censored by the inspection time.

Let us point out that, without any model assumption, given a distribution for the observed variables (Y, A) with $Y \geq 0$ and $A \in \{0, 1, 2\}$, we can always apply our two inversion formulae. In this way we build two pseudo-true distribution functions of the lifetime of interest which are functionals of the observed distribution. If the experiment under observation is compatible with the hypothesis of one of our latent models, the true F_T can be exactly recovered from the observed distribution. Otherwise, we can only approximate the true lifetime distribution.

The paper is organized as follows. Section 2 introduces our two latent models through the equations relating the distribution of the observations to those of the latent variables. Solving these equations for F_T we deduce the inversion formulae. The product-limit estimators are obtained by applying the inversion formulae to the empirical distribution. Section 2 is ended with some remarks and comments on related models. Section 3 contains the

asymptotic results for the first latent model (similar arguments apply for the second model). We prove strong uniform convergence for the product-limit estimator on the whole range of the observations. Our proof extends and simplifies the results of [Stute and Wang, 1994] and [Gill, 1994] provided in the case of the Kaplan-Meier estimator. Next, the asymptotic normality of our product-limit estimator is obtained. The variance of the limit Gaussian process being complicated, a bootstrap procedure for which the asymptotic validity is a direct consequence of the delta-method is proposed.

2 The latent models

2.1 Model 1

The survival time of interest is T (e.g., the age at onset). Let C be a censoring time (e.g., the age of the individual at the examination time) and Δ be a Bernoulli random variable. Assume that the latent variables T, C and Δ are independent. The observations are independent copies of the variables (Y, A), with $Y \geq 0$ and $A \in \{0, 1, 2\}$. These variables are defined as

$$Y = \min(T, C) + (1 - \Delta) \max(C - T, 0) = C + \Delta \min(T - C, 0)$$

and $A = 2(1-\Delta)\mathbf{1}_{\{T \leq C\}} + \mathbf{1}_{\{C<T\}}$, where $\mathbf{1}_A$ denotes the indicator function of the set A. With this censoring mechanism the lifetime T is observed, right censored or left censored. In view of the definitions of Y and A, note that if Δ is constant and equal to one (resp. zero), we obtain right censored (resp. current status) data.

Let F_T and F_C denote the distributions of T and C, respectively. Let $p = P(\Delta = 1)$. Define the observed subdistributions of Y as

$$H_k(B) = P(Y \in B,\, A = k), \qquad k = 0, 1, 2, \tag{1}$$

for any B Borel subset of $[0, \infty]$. As usually in survival analysis, the censoring mechanism defines a map Φ between the distributions of the latent variables and the observed distributions. For the censoring mechanism we consider, the relationship $(H_0, H_1, H_2) = \Phi(F_T, F_C, p)$ between the subdistributions of Y and the distributions of the latent variables T, C and Δ is the following:

$$\begin{cases} H_0(dt) = p\, F_C([t, \infty])\, F_T(dt) \\ H_1(dt) = F_T((t, \infty])\, F_C(dt) \\ H_2(dt) = (1 - p)\, F_T([0, t])\, F_C(dt) \end{cases}. \tag{2}$$

Remark that when $p = 1$ (resp. $p = 0$) the equations (2) boil down to the equations of the classical independent right-censoring (resp. current status) model.

By plug-in applied with the empirical distribution, the nonparametric estimation of the distribution of T is straight as soon as the map Φ is invertible

and F_T can be written as an explicit function of the observed subdistributions H_k, $k = 0, 1, 2$. The model considered allows us an explicit inversion formula for F_T. In order to derive this inversion formula, integrate the first and the second equation in (2) on $[t, \infty]$ and deduce

$$H_0([t,\infty]) + pH_1([t,\infty]) = pF_T\left([t,\infty]\right) F_C\left([t,\infty]\right). \tag{3}$$

For $t = 0$, it follows that

$$p = \frac{H_0([0,\infty])}{1 - H_1([0,\infty])} = \frac{H_0([0,\infty])}{H_0([0,\infty]) + H_2([0,\infty])}. \tag{4}$$

Recall that the hazard measure associated to a distribution F is $\Lambda(dt) = F(dt)/F([t,\infty])$. In our case, use (2)-(3) to deduce that the hazard function corresponding to F_T can be written as

$$\Lambda_T(dt) = \frac{H_0(dt)}{H_0([t,\infty]) + pH_1([t,\infty])}. \tag{5}$$

Finally, the distribution F_T can be expressed as

$$F_T\left((t,\infty]\right) = \prod_{[0,t]} (1 - \Lambda_T(ds)), \tag{6}$$

where $\prod_{[0,t]}$ is the product-integral on $[0, t]$. Note that there is no explicit formula for F_T if $p = 0$ in equations (2), that is with current status data.

Given the explicit relationship between the distribution of T and the observed subdistributions, to obtain the product-limit estimator of F_T, we simply replace H_k, $k = 0, 1, 2$ by their empirical counterparts. Let $\widehat{F}_T$ denote the product-limit estimator of F_T.

2.2 Model 2

As in Model 1, assume that T, C and Δ are independent. The observations are independent copies of the variables (Y, A), with $Y \geq 0$ and $A \in \{0, 1, 2\}$ where

$$\begin{cases} Y = T, \quad A = 0 \;\; if \quad 0 \leq C \leq T \quad and \quad \Delta = 1; \\ Y = C, \quad A = 1 \;\; if \quad 0 \leq C \leq T \quad and \quad \Delta = 0; \\ Y = C, \quad A = 2 \;\; if \quad 0 \leq T < C. \end{cases} \tag{7}$$

The equations of this model are

$$\begin{cases} H_0(dt) = p\, F_C\left([0,t]\right) F_T(dt) \\ H_1(dt) = (1-p)\, F_T\left([t,\infty]\right) F_C(dt) \\ H_2(dt) = F_T\left([0,t)\right) F_C(dt) \end{cases}. \tag{8}$$

Remark that when $p = 1$ (resp. $p = 0$) the equations (8) boil down to the equations of the classical independent left-censoring (resp. current status)

model. This model also allows for an explicit inversion formula for F_T. By integration in the first and the third equation in (8), $H_0([0,t])+pH_2([0,t]) = pF_T([0,t])F_C([0,t])$. Deduce

$$p = \frac{H_0([0,\infty])}{1-H_2([0,\infty])}.$$

Recall that given a distribution F, the associated reverse hazard measure is $M(dt) = F(dt)/F([0,t])$. By equations (8) deduce that the reverse hazard function M_T associated to F_T can be written as

$$M_T(dt) = \frac{H_0(dt)}{H_0([0,t]) + pH_2([0,t])}.$$

Finally, the distribution F_T can be expressed as

$$F_T\left([0,t]\right) = \prod_{(t,\infty]} (1 - M_T(ds)).$$

Applying the inversion formula with the empirical subdistributions, we get the product-limit estimator of F_T in Model 2.

Note that if $\widetilde{T} = h(T)$ and $\widetilde{C} = h(C)$, with $h \geq 0$ a decreasing transformation, then $\widetilde{T}$, $\widetilde{C}$ and Δ are the variables of Model 1 applied to the left or right censored lifetime $h(Y)$. In other words, Model 2 is equivalent to Model 1, up to a time reversal transformation.

2.3 Related models

[Huang, 1999] introduced a model for the so-called partly interval-censored data, Case 1; see also [Kim, 1994]. In such data, for some subjects, the exact failure time of interest T is observed. For the remaining subjects, only the information on their current status at the examination time is available. [Huang, 1999] considered the nonparametric maximum likelihood estimator (NPMLE) of F_T. Unfortunately, NPMLE does not have an explicit form and therefore Huang needs strong assumptions for deriving its asymptotic properties and a numerical algorithm for the applications. Let us point out that, on contrary to our Model 1 (resp. Model 2), in Huang's model one may observe exact failure times even if failure occurs after (resp. before) the examination time. Moreover, in Huang's model one may still obtain a $\sqrt{n}-$consistent estimator of the distribution F_T if one simply considers the empirical distribution of the uncensored lifetimes. This is no longer true in our models.

Perhaps, the most popular model for left or right-censored data is the one introduced by [Turnbull, 1974]; see also [Gu and Zhang, 1993]. In Turnbull's model there are three latent lifetimes L (left-censoring), T (lifetime of interest) and R (right-censoring) with $L \leq R$. The observed variables

are $Y = \max(L, \min(T, R)) = \min(\max(L, T), R)$ and A defines as follows: $A = 0$ if $L < T \leq R$; $A = 1$ if $R < T$; and $A = 2$ if $T \leq L$. The equations of this model are

$$\begin{cases} H_0(dt) = \{F_R([t, \infty]) - F_L([t, \infty])\}\, F_T(dt) \\ H_1(dt) = F_T((t, \infty])\, F_R(dt) \\ H_2(dt) = F_T([0, t])\, F_L(dt) \end{cases},$$

where H_k, $k = 0, 1, 2$ are defined as in (1) and F_T, F_L and F_R are the distributions of T, L and R, respectively. The NPMLE of the distribution of the failure time T is not explicit but it can be computed, for instance, by iterations based on the so-called self-consistency equation. Note that imposing $F_C(dt) = (1-p)^{-1}F_L(dt) = F_R(dt)$, one recovers the equations of Model 1. However, for the applications we have in mind, there is no natural interpretation for such a constraint in Turnbull's model. Moreover, we derive a product-limit estimator for our Model 1. Finally, the asymptotic results below are much simpler and they are obtained under weaker conditions than in Turnbull's model.

3 Asymptotic results

In this section the strong uniform convergence and the asymptotic normality for the estimator of the distribution F_T in Model 1 are derived. Moreover, we propose a bootstrap procedure that can be used to build confidence intervals for F_T. As in the previous sections, the distributions F_T and F_C need not be continuous. For simpler notation, hereafter, the subscript T is suppressed when there is no possible confusion. We write $\widehat{F}$ (resp. F, $\widehat{\Lambda}$ and Λ) instead of $\widehat{F}_T$ (resp. F_T, $\widehat{\Lambda}_T$ and Λ_T).

3.1 Strong uniform convergence

Let H_{nk} be the empirical counterparts of the subdistributions H_k, $k = 0, 1, 2,$, that is

$$H_{nk}([0, t]) = \sum_{i=1}^{n} \mathbf{1}_{\{Y_i \leq t,\ A_i = k\}}, \qquad k = 0, 1, 2.$$

Clearly, $\sup_{t \geq 0} |H_{nk}([0, t]) - H_k([0, t])| \to 0$, almost surely. We want to prove the strong uniform convergence of the distribution $\widehat{F}$, that is

$$\sup_{t \in I} \left| \widehat{F}([0, t]) - F([0, t]) \right| \to 0, \quad asn \to \infty, \qquad almost\ surely,$$

where $I = \{t : H_0([t, \infty]) + pH_1([t, \infty]) > 0\}$. For this purpose, first we prove the almost sure convergence of the hazard function.

Theorem 1 *Assume that $p \in (0,1]$ and let $t_* = \sup I$. For any $\sigma \in I$,*

$$\sup_{0 \le t \le \sigma} \left| \widehat{\Lambda}([0,t]) - \Lambda([0,t]) \right| \to 0, \qquad \text{as } n \to \infty, \qquad \text{almost surely.}$$

Moreover, if $t_* \notin I$ and $\Lambda([0,t_*)) < \infty$, then $\widehat{\Lambda}([0,t_*)) \to \Lambda([0,t_*))$, almost surely.

The strong uniform convergence of the distribution $\widehat{F}$ follows without any additional assumption.

Theorem 2 *Assume that $p \in (0,1]$. Then*

$$\sup_{t \in I} \left| \widehat{F}([0,t]) - F([0,t]) \right| \to 0, \qquad \text{as} \quad n \to \infty, \quad \text{almost surely.}$$

With $p = 1$ one recovers the strong uniform convergence result for the Kaplan-Meier estimator obtained by [Stute and Wang, 1994], [Gill, 1994]. Our alternative proof is simpler.

3.2 Asymptotic normality

Here we study the weak convergence of the process $\sqrt{n}(\widehat{F} - F)$ where $\widehat{F}$ is the product-limit estimator of Model 1. In this case, $\widehat{\Lambda}$ does no longer have a martingale structure (in t) as in the case of the Nelson-Aalen estimator, that is when $p = 1$. However, a continuous time *submartingale* property for $\widehat{\Lambda}$ can be obtained. This suffices us to extend the techniques of Gill (1983) and to use them in combination with the functional delta-method in order to establish the weak convergence of $\sqrt{n}(\widehat{F} - F)$ to a Gaussian process. Here, the weak convergence is denoted by $\Rightarrow$. The space $D[a,b]$ of càdlàg functions defined on $[a,b]$ is endowed with the supremum norm and the ball $\sigma-$field.

Theorem 3 *Assume that $p \in (0,1]$ and define $U(t) = \sqrt{n}(\widehat{F}([0,t]) - F([0,t]))$, $t \ge 0$. Let $t_* = \sup I$.*

a) Let τ be a point in I. Then, $U \Rightarrow \mathbf{G}$ in $D[0,\tau]$, where $\mathbf{G}$ is a Gaussian process.

b) If $t_* \notin I$, but

$$\int_{[0,t_*)} \frac{H_0(dt)}{\{H_0([t,\infty]) + pH_1([t,\infty])\}^2} < \infty, \tag{1}$$

then $\mathbf{G}$ can be extended to a Gaussian process on $[0,t_*]$ and $U \Rightarrow \mathbf{G}$ in $D[0,t_*]$.

The proof of the weak convergence is postponed to the appendix. Note that when $t_* \notin I$, condition (1) is equivalent to

$$F_T([t_*,\infty]) > 0 \qquad \text{and} \qquad \int_{[0,t_*)} \frac{F_T(dt)}{F_C([t,\infty])} < \infty. \tag{2}$$

3.3 Bootstrapping the product-limit estimator

Theorem 3 may be used to obtain confidence intervals and confidence bands for F. However, the law of the process $\widetilde{\mathbf{G}}(t) = \mathbf{G}(t)/F((t,\infty])$ being complicated, one may prefer a bootstrap method in order to avoid handling this process in practical applications. Here, a bootstrap sample is obtained by simple random sample with replacement from the set of observations $\{(Y_i, A_i) : 1 \leq i \leq n\}$. Let $\{(Y_i^*, A_i^*) : 1 \leq i \leq n\}$ denote a bootstrap sample and let H_k^* be the corresponding subdistributions. Apply equations (4) to (6) to obtain the bootstrap estimator $\widehat{F}^*$. The following theorem state that the bootstrap works almost surely for our product-limit estimator on any interval $[0,\tau]$ such that $H_0([\tau,\infty]) + pH_1([\tau,\infty]) > 0$. This result is a simple corollary of Theorem 3.9.13 of [Van der Vaart and Wellner, 1996] and it is based on the uniform Hadamard differentiability of the maps involved in the inversion formula of Model 1.

Theorem 4 *Let $\tau \in I$ and let $\widetilde{\mathbf{G}}(t)$ be the limit of $U(t)/F((t,\infty])$ in $D[0,\tau]$, as obtained from Theorem 3. Then, the process*

$$\sqrt{n}\{\widehat{F}^*([0,t]) - \widehat{F}([0,t])\}/\widehat{F}((t,\infty])$$

converges to $\widetilde{\mathbf{G}}$ in $D[0,\tau]$, almost surely.

References

[Gill, 1994]R. Gill. Lectures on survival analysis. *Lecture Notes in Mathematics (Ecole d'été de Probabilités de Saint-Flour XXII 1992)*, pages 115–241, 1994.

[Gu and Zhang, 1993]M.G. Gu and C.-H. Zhang. Asymptotic properties of self-consistent estimators based on doubly censored data. *Ann. Statist.*, pages 611–624, 1993.

[Huang, 1999]J. Huang. Asymptotic properties of nonparametric estimation based on partly interval-censored data. *Statist. Sinica*, pages 501–519, 1999.

[Kim, 2003]J.S. Kim. Maximum likelihood estimation for the proportional hazards models with partly interval-censored data. *J. R. Stat. Soc. Ser B*, pages 489–502, 2003.

[Samuelsen, 1989]S.O. Samuelsen. Asymptotic theory for non-parametric estimators from doubly censored data. *Scand. J. Statist.*, pages 1–21, 1989.

[Stute and Wang, 1994]W. Stute and J. Wang. The strong law under random censorship. *Ann. Statist.*, pages 1591–1607, 1994.

[Turnbull and Weiss, 1978]B.W. Turnbull and L. Weiss. A likelihood ratio statistics for testing goodness of fit with randomly censored data. *Biometrics*, pages 367–375, 1978.

[Turnbull, 1974]B.W. Turnbull. Nonparametric estimation of a survivorship function with doubly censored data. *J. Amer. Statist. Assoc.*, pages 169–173, 1974.

[Van der Vaart and Wellner, 1996]A.W. Van der Vaart and J.A. Wellner. *Weak convergence and empirical processes.* Springer Verlag, New-York, 1996.