South African Statistical Journal
Vol. 56, No. 1, pp. 53–67
https://doi.org/10.37920/sasj.2022.56.1.4

Bandwidth choice for the smooth Kaplan–Meier estimator when the censoring variable can be discontinuous

Élie Youndjé

Laboratoire Raphaël Salem UMR 6085 CNRS Université de Rouen Normandie, France

In this paper a new cross-validation method (CVN) for selecting the bandwidth of the smooth Kaplan–Meier estimator is introduced. Its performance is compared to that of the usual cross-validation by simulations. The censoring variables in the experimentations are discrete. A set of ideas explaining how the classical and smooth Kaplan–Meier estimators work on the truncation model (usually encountered in reliability studies) is also given in the paper.

Keywords: Censored data, Bandwidth selection, Kaplan–Meier estimator, Kernel smoothing, Survival analysis, Truncation model.

1. Introduction

Survival analysis is a branch of statistics which studies time to event random variables. For this type of variables, it would be desirable to obtain their mean (life expectancy), their median (median life expectancy), their variances, and their distributions in general. However, in most survival analysis studies, the variable of interest is censored and, if the support of the censoring variable is shorter than that of the variable of interest, some of its summary statistics cannot be estimated from the data.

One of the most important problems with survival data is the modelling of the survival function. A popular tool for doing so is the Kaplan–Meier estimator. It enables one to estimate survival probabilities and some quantiles of the variable of interest. It is an invaluable tool because it helps to compare survival functions between subgroups in the data. Also, in some instances, the median might be estimatable while the mean is not; in this case the Kaplan–Meier estimator gives an estimate of the median and this quantity is the only information regarding life expectancy.

It is shown in the literature that smoothing a discontinuous distribution function estimator can bring some advantages. When the data is complete, Reiss (1981) showed that the smooth empirical distribution improves the empirical distribution in terms of mean squared error; Ghorai and Susarla (1990) extended this result to the setting of censored data. They showed that the smooth Kaplan–Meier estimator has a better mean squared error than the ordinary Kaplan–Meier estimator.

Implementation of the smooth Kaplan–Meier estimator requires a bandwidth and its value is of vital importance for the performance of the estimator (see for instance Ghorai and Susarla, 1990). Youndjé (2016) introduced some criteria for selecting this parameter of the estimator. But, apart

Corresponding author: Élie Youndjé (Elie.Youndje@univ-rouen.fr)
MSC2020 subject classifications: 62G05, 62N86, 62N02.

from the CV method, the other methods were devised to work when the variable of interest and the censoring variable are continuous. In this paper, we will put more emphasis on the case where the censoring distribution is discontinuous. We will introduce a new cross-validation criterion (CVN), give some of its appealing theoretical properties and then compare its performance with that of the CV criterion on models in which the censoring random variable is discrete. The motivation of the CV method is given in details in Youndjé (2016).

This paper is organised as follows. In the next section, we introduce the CVN criterion. Section 3 contains some of its theoretical properties. Simulations are carried out in Section 4. In Section 5 we summarise the findings of the simulation study. All the proofs are gathered in the appendix.

2. The new cross-validation score function

Let X and C be two nonnegative independent random variables having distribution functions F and G respectively. In the sequel, we assume that F is continuous. Let us set

$$T = X \wedge C, \quad \Delta = \mathbb{1}_{[X \leq C]}.$$

In this article we consider the estimation of F with right censored data. We assume that we have independent and identically distributed (for short i.i.d.) observations

$$(T_1, \Delta_1), \ldots, (T_n, \Delta_n)$$

of the random pair (T, Δ) and we want to use them to estimate F. The popular estimator of F in the right censored data model is the Kaplan–Meier estimator given by

$$1 - F_n^{KM}(x) = \prod_{i:T_{(i)} \leq x} \left(1 - \frac{\Delta_{(i)}}{n - i + 1}\right).$$

Here, $(T_{(i)}, \Delta_{(i)})$, $i = 1, \ldots, n$, are the n-pairs (T_i, Δ_i), $i = 1, \ldots, n$, sorted in the lexicographical order of $(T, 1 - \Delta)$. Let K be a continuous and bounded probability function and $h = h(n)$ a positive sequence. The smooth Kaplan–Meier estimator based on censored data can be expressed as (see for example Ghorai and Susarla, 1990)

$$F_h(x) = \int \frac{1}{h} K\left(\frac{x - y}{h}\right) F_n^{KM}(y) \mathrm{d}y.$$

For a distribution function D, let us define the least upper bound of its support as

$$\tau_D = \inf\{x \mid D(x) = 1\}.$$

Let L denote the distribution function of the random variable T. It will be convenient in the sequel to set $\tau = \tau_L$. It is also useful to notice that $\tau = \tau_F \wedge \tau_G$.

For a distribution function D, set

$$D(a^-) = \lim_{t \uparrow a} D(t).$$

Let $G_n^{KM} = G_n$ be the Kaplan–Meier estimator of G based on $(T_1, 1-\Delta_1), \ldots, (T_n, 1-\Delta_n)$. It is shown in Satten and Datta (2001) that F_n^{KM} can be expressed as

$$F_n^{KM}(x) = \frac{1}{n}\sum_{i=1}^{n} W_i \mathbb{1}_{[T_i \le x]}, \tag{1}$$

where

$$W_i = \frac{\Delta_i}{1 - G_n(T_i^-)}.$$

This representation implies that

$$F_h(x) = \frac{1}{n}\sum_{i=1}^{n} W_i H\left(\frac{x - T_i}{h}\right), \tag{2}$$

with

$$H(x) = \int_{-\infty}^{x} K(t)\mathrm{d}t.$$

Let us denote by $\tilde{F}_n$ the classical empirical distribution (which is not available in the setting of censored data) i.e.

$$\tilde{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[X_i \le x]},$$

and $\hat{F}_h$ the ordinary Nadaraya estimator of F given by

$$\hat{F}_h(x) = \frac{1}{n}\sum_{i=1}^{n} H\left(\frac{x - X_i}{h}\right).$$

It follows from Equations (1) and (2) that, if there are no censored observations (hence no jump in G_n i.e. $G_n \equiv 0$) we have

$$F_n^{KM}(x) = \tilde{F}_n(x) \quad \text{and} \quad F_h(x) = \hat{F}_h(x).$$

Let F_h^{-i} be the kernel estimator of F obtained with the data $(T_1, \Delta_1), \ldots, (T_{i-1}, \Delta_{i-1})$, $(T_{i+1}, \Delta_{i+1}), \ldots, (T_n, \Delta_n)$. Likewise let $\hat{F}_h^{-i}$ be the leave-one-out Nadaraya estimator of F. We are now going to summarise how the authors motivated the $CVB(h)$ criterion (note that CVB is denoted CV in that paper) presented in Bowman et al. (1998). $CVB(h)$ is the cross-validation criterion used to select the bandwidth of $\hat{F}_h$.

Step 1 The error criteria for measuring the global performance of $\hat{F}_h$, $ISE(\hat{F}_h)$ and $MISE(\hat{F}_h)$ are defined by

$$ISE(\hat{F}_h) = \int \left(\hat{F}_h(x) - F(x)\right)^2 \mathrm{d}x \quad \text{and} \quad MISE(\hat{F}_h) = \mathrm{E}\left(ISE(\hat{F}_h)\right).$$

Step 2 The authors introduced the $CVB(h)$ criterion given by

$$CVB(\hat{F}_h) = \frac{1}{n}\sum_{i=1}^{n} \int \left(\hat{F}_h^{-i}(x) - \mathbb{1}_{[X_i \le x]}\right)^2 \mathrm{d}x.$$

Step 3 They showed that there exists a random variable T independent of h such that

$$ED(h) = MISE(\hat{F}_h^{-n}), \quad \text{where } D(h) = CVB(h) - T.$$

We will use similar ideas to propose a bandwidth selector for the estimator F_h.

Step 1 Measures of global performance for F_h are given by

$$ISE(F_h) = \int_0^\tau (F_h(x) - F(x))^2 \, \mathrm{d}x \quad \text{and} \quad MISE(F_h) = \mathrm{E}\left(ISE(F_h)\right),$$

see Youndjé (2016) for the motivations.

Step 2 We introduce the following criterion (to be ready for use, this criterion will be modified in step 4):

$$CVN^*(F_h) = \frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_i}{1 - G(T_i^-)} \int_0^\tau \left(F_h^{-i}(x) - F(\tau)\mathbb{1}_{[T_i \le x]}\right)^2 \mathrm{d}x.$$

Step 3 Set

$$T = \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \left(F^2(\tau)\mathbb{1}_{[X_i \le x]} - F(\tau)F^2(x)\right) \mathrm{d}x,$$

and note that T is independent of h. We will prove in the Appendix that

$$ED(h) = F(\tau)MISE(F_h^{-n}), \quad \text{where } D(h) = CVN^*(h) - T. \tag{3}$$

Step 4 This extra step compared to the approach of Bowman et al. (1998) is due to the fact that our CVN^* involves the unknown functions F, G and the quantity τ. In this last step, we are going to plug-in the estimates of these quantities. Our final CVN criterion is given by

$$CVN(F_h) = \frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_i}{1 - G_n(T_i^-)} \int_0^{T_{(n)}} \left(F_h^{-i}(x) - F_n^{KM}(T_{(n)})\mathbb{1}_{[T_i \le x]}\right)^2 \mathrm{d}x.$$

Recall that F_n^{KM} is the Kaplan–Meier estimator of F, $G_n = G_n^{KM}$ that of G, and $T_{(n)} = \max(T_i, i = 1, \ldots, n)$. It was shown in Youndjé (2016) that $T_{(n)}$ is a consistent estimator of τ.

In this paper, we will compare the performance of the CVN criterion to that of the CV method introduced in Youndjé (2016). This CV criterion is given by

$$CV(F_h) = \frac{1}{n}\sum_{i=1}^{n} \int_0^{T_{(n)}} \left(F_h^{-i}(x) - \frac{\Delta_i}{1 - G_n(T_i^-)}\mathbb{1}_{[T_i \le x]}\right)^2 \mathrm{d}x.$$

Note that when there is no censoring, we have: $\Delta_i = 1$, $G_n \equiv 0$, $F_n^{KM}(T_{(n)}) = 1$ so that

$$CVN(F_h) = \frac{1}{n}\sum_{i=1}^{n} \int_0^{T_{(n)}} \left(\hat{F}_h^{-i}(x) - \mathbb{1}_{[X_i \le x]}\right)^2 \mathrm{d}x;$$

and therefore $CV = CVN$. Moreover, if the kernel function K is supported over the interval $[-1, 1]$, we have

$$CVB(\hat{F}_h) = \frac{1}{n}\sum_{i=1}^{n} \int_0^{T_{(n)}+h} \left(\hat{F}_h^{-i}(x) - \mathbb{1}_{[X_i \le x]}\right)^2 \mathrm{d}x.$$

It follows that the three criteria are very similar.

3. Some theoretical properties of CVN

In this paper we will assume that the distribution function F, the kernel K and the bandwidth h satisfy the following conditions:

(A.1) K is nonnegative and $\int K(u)\mathrm{d}u = 1$;

(A.2) K is compactly supported on $[-1,\ 1]$;

(A.3) K is symmetric and $0 < \int u^2 K(u)\mathrm{d}u < +\infty$;

(A.4) F is twice continuously differentiable, and $F' = f$, $F'' = f'$ are bounded;

(A.5) $\lim_{n\to+\infty} h = \lim_{n\to+\infty} h(n) = 0$.

Actually, we are going to establish an "optimality" result for a "truncated" version of CVN. We are going to introduce a weight function (see for example Härdle and Marron (1985) for the case of a regression function estimation) to deal with the random denominators in CVN. Let $R > 0$ such that

$$R < \tau. \tag{4}$$

We will consider the following quantities:

$$CVN1(h) = \frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_i}{1 - G_n(T_i^-)} \mathbb{1}_{[T_i \le R]} \int_0^R \left(F_h^{-i}(x) - F(R)\mathbb{1}_{[T_i \le x]}\right)^2 \mathrm{d}x,$$

$$CVN1^*(h) = \frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_i}{1 - G(T_i^-)} \mathbb{1}_{[T_i \le R]} \int_0^R \left(F_h^{-i}(x) - F(R)\mathbb{1}_{[T_i \le x]}\right)^2 \mathrm{d}x,$$

$$T1 = \frac{1}{n}\sum_{i=1}^{n} \int_0^R \left(F^2(R)\mathbb{1}_{[X_i \le x]} - F(R)F^2(x)\right) \mathrm{d}x.$$

Some comments are in order here. Setting

$$We(u) = \mathbb{1}_{[u \le R]},$$

we see that $CVN1$ (for instance) can be written as

$$CVN1(h) = \frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_i}{1 - G_n(T_i^-)} We(T_i) \int_0^{+\infty} \left(F_h^{-i}(x) - F(R)\mathbb{1}_{[T_i \le x]}\right)^2 We(x)\mathrm{d}x.$$

Therefore, the quantity R helps us in setting a weight function. Note also, the presence of the weight function inside and outside the integrals. We believe that this is inherent to the cross-validation approach to distribution function estimation.

Condition (4) is technical and will help to cope with the random denominators in $CVN1$ and $CVN1^*$. For instance, it follows from that condition that

$$\frac{\Delta_i}{1 - G(T_i^-)} \mathbb{1}_{[T_i \le R]} \le \frac{1}{1 - G(R)}.$$

The fact that $R < \tau$ is essential to obtain that $G(R) < 1$, as is evidenced by the case where the censoring variable is uniform over $]0, 1[$ and $G(1^-) = G(1) = 1$.

We have the following theorem:

Theorem 1. *If Condition* (4) *holds and if (A.1)–(A.5) hold, then we have*

(i) $E(CVN1^*(h) - T1) = F(R)\int_0^R E\left(F_h^{-n}(u) - F(u)\right)^2 \mathrm{d}u,$

(ii) $CVN1(h) \overset{a.s.}{=} CVN1^*(h) + o(CVN1^*(h)).$

The proof of Theorem 1 (i) is similar to that of Equation (3) and is thus omitted. The proof of Theorem 1 (ii) is in the Appendix.

Remarks

(i) Theorem 1 (i) suggests that $CVN1^*$ is potentially a good bandwidth selector. Theorem 1 (ii) says that $CVN1$ is equivalent ($CVN1/CVN1^* \xrightarrow{a.s.} 1$) to $CVN1^*$, thus $CVN1$ is likely also a good selector.

(ii) Theorem 1 (ii) would have been more appealing (for practical situations) if the term $F(R)$ in $CVN1$ was $F_n^{KM}(R)$. Unfortunately, we were unable to prove Theorem 1 (ii) with $F(R)$ replaced by $F_n^{KM}(R)$.

(iii) What is the connection between CVN (introduced in Section 2) and Theorem 1?

(a) Assume that τ and $F(\tau)$ are known. Since Theorem 1 is true for all R such that $R < \tau$, we can extrapolate by setting $R = \tau$ and get

$$CVN^{\tau}(F_h) = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{1 - G_n(T_i^-)}\int_0^{\tau}\left(F_h^{-i}(x) - F(\tau)\mathbb{1}_{[T_i \le x]}\right)^2 \mathrm{d}x,$$

which is very similar to CVN.

(b) To analyse the case where $F(\tau)$ and τ are unknown, set

$$J_n = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{1 - G_n(T_i^-)}.$$

Since the kernel function K is supported on $[-1,\ 1]$, we have

$$F_h(x) = J_n \quad \text{if} \quad x > T_{(n)} + h.$$

This equality is of course true when $x > \tau + h$. It can easily be verified that

$$J_n = F_n^{KM}(T_{(n)}) = F_n^{KM}(\tau).$$

Therefore

$$J_n \xrightarrow{a.s.} F(\tau).$$

See for example Stute and Wang (1993). It follows from (2) that

$$J_n = \max_x(F_h(x)). \tag{5}$$

We have chosen to use J_n as estimator of $F(\tau)$ throughout this paper. The main inconvenience of this choice is that J_n is greater than or equal to (see Equation (5)) any kernel estimate $F_h(\tau)$ of $F(\tau)$ regardless of the kernel and the bandwidth. But, this choice has also its advantages, since we do not need to choose a bandwidth and J_n is a strongly consistent estimator of $F(\tau)$. Besides, when $F(\tau) = 1$, J_n is better than any kernel estimate $F_h(\tau)$ of $F(\tau)$ (see Equation (5)).

Let us now turn to the estimation of τ. In Youndjé (2016), it was shown that $T_{(n)}$ is a consistent estimator of τ. In summary, if R is replaced by $T_{(n)}$ and $F(R)$ by $F_n^{KM}(T_{(n)})$ in $CVN1$, then we obtain the CVN criterion introduced in Section 2.

4. Simulations

This section is organised according to the distribution of the censoring random variable C. Section 4.1 presents simulations results when the censoring variable C is constant, that is we are dealing with truncation models. The models where the censoring random variables are geometric are presented in Section 4.2.

Let $\mathrm{W}(\alpha,\ \lambda)$ be the Weibull random variable with parameters $\alpha > 0$ and $\lambda > 0$ whose distribution function is given by

$$F_W(x) = 1 - e^{-(\lambda x)^\alpha}, \quad x > 0.$$

In our simulation study the target distribution function will be $F = F_{W_0}$, where $W_0 = W(\alpha_0, \lambda_0)$, $\alpha_0 = 1.6$ and $\lambda_0 = \frac{1}{16}$. Throughout our experiments, the kernel function used is the Epanechnikov kernel defined by

$$K(x) = \frac{3}{4}\mathbb{1}_{[-1,\ 1]}(x)(1 - x^2).$$

Sample sizes $n = 100$, $n = 300$ and $n = 500$ are considered throughout the experiments.

4.1 Models where C is constant: right truncation models

Recall that in this case

$$G(x) = \begin{cases} 0 & \text{for} \quad x < t_0, \\ 1 & \text{otherwise,} \end{cases}$$

and $G_n = G$ if there is at least one censored observation. Therefore, we have

$$F_n^{KM}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_i}{1 - G(T_i^-)}\mathbb{1}_{[T_i \le x]}.$$

This formula of $F_n^{KM}(x)$ and Proposition 2 in Youndjé (2016) can be used to prove that

$$\mathrm{E}\left(F_n^{KM}(u)\right) = \begin{cases} F(u) & \text{for} \quad u < t_0, \\ F(t_0) & \text{otherwise.} \end{cases}$$

In this setting, another representation of the Kaplan–Meier estimator, which helps to gain more insights and understanding is also available. Observe that in this framework $G(T_i^-) = 0$, so we have

$$F_n^{KM}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[X_i \le t_0]}\mathbb{1}_{[T_i \le x]} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[X_i \le x \wedge t_0]}.$$

This representation shows how powerful the Kaplan–Meier estimator is. It captures all the appealing characteristics of the empirical distribution function $\tilde{F}_n(x)$ when $x \leq t_0$ and is equal to $\tilde{F}_n(t_0)$ when $x > t_0$.

A legitimate question is whether it is possible to deduce some of the properties of $F_h(x)$ in this setting from those of the ordinary kernel estimator $\hat{F}_h$ of F? The answer is affirmative and is given by the following equations:

$$F_h(x) = \hat{F}_h(x) \quad \text{when} \quad x < t_0 - h, \tag{6}$$

$$F_h(x) = \tilde{F}_n(t_0) \quad \text{when} \quad x > t_0 + h. \tag{7}$$

To prove Equation (6), observe that in this setting we have

$$\begin{aligned} F_h(x) =& \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[X_i \leq t_0]} H\left(\frac{x - T_i}{h}\right) \\ =& \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[X_i \leq t_0]} H\left(\frac{x - X_i}{h}\right) \\ =& \frac{1}{n}\sum_{i=1}^{n} H\left(\frac{x - X_i}{h}\right) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[X_i > t_0]} H\left(\frac{x - X_i}{h}\right). \end{aligned}$$

Since the term on the right hand side in the last equality is equal to zero when $x < t_0 - h$, the proof of Equation (6) is established. To prove Equation (7), note that $F_h(x)$ can also be written as

$$F_h(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[X_i \leq t_0]} H\left(\frac{x - T_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} H\left(\frac{x - T_i}{h}\right) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[X_i > t_0]} H\left(\frac{x - t_0}{h}\right).$$

For $x > t_0 + h$, we have:

$$\frac{x - t_0}{h} > 1, \quad \text{and} \quad \frac{x - T_i}{h} \geq \frac{x - t_0}{h} > 1.$$

Since K is supported on $[-1,\ 1]$, we get

$$F_h(x) = 1 - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[X_i > t_0]} = \tilde{F}_n(t_0),$$

and the proof of Equation (7) is complete.

In this subsection, we have $\tau < +\infty$. The measure of accuracy used to assess the performance of a bandwidth h of F_h is given by

$$ISE(h) = ISE(F_h) = \int_0^{\tau} (F_h(x) - F(x))^2 \, \mathrm{d}x.$$

This measure will be slightly modified in Subsection 4.2 to take into account the fact that $\tau = +\infty$. Two models are used to assess the finite sample behavior of CV and CVN. The first model is the Weibull-Truncation-50% (W-T-50%) one given by

$$\text{Model W-T-50\%:} \quad X \sim \mathrm{W}(\alpha_0,\ \lambda_0), \quad C \sim \mathrm{Trunc}(t_1),$$

Table 1. Integrated squared error for $\hat{h}_{cvn}$ and $\hat{h}_0$; mean, median and standard deviation over 100 replications.

CVN RESULTS							
		$ISE(\hat{h}_{cvn})$			$ISE(\hat{h}_0)$		
Model	n	Mean	Median	Standard deviation	Mean	Median	Standard deviation
W-T-50%	200	0.00945	0.00675	0.00878	0.00854	0.00538	0.00861
	400	0.00394	0.00251	0.00470	0.00346	0.00197	0.00447
	800	0.00221	0.00120	0.00266	0.00196	0.00101	0.00249
W-T-25%	200	0.01519	0.01120	0.01332	0.01384	0.01040	0.01286
	400	0.00819	0.00475	0.00860	0.00711	0.00409	0.00806
	800	0.00430	0.00279	0.00500	0.00387	0.00233	0.00466

where $\mathrm{W}(\alpha_0, \lambda_0)$ is the Weibull random variable with parameters $\alpha_0 = 1.6, \lambda_0 = \frac{1}{16}$ and $C \sim \mathrm{Trunc}(t_1)$ (i.e. $C = t_1$), $t_1 = 12.724$. With this choice of the parameters $\mathrm{P}(X > C) = 0.5$ so the censoring level is 50% when estimating F. The second model is the Weibull-Truncation-25% (W-T-25%) defined by

$$\text{Model W-T-25\%:} \quad X \sim \mathrm{W}(\alpha_0,\ \lambda_0), \quad C \sim \mathrm{Trunc}(t_2),$$

where $\mathrm{W}(\alpha_0, \lambda_0)$ is the Weibull random variable with parameters $\alpha_0 = 1.6$, $\lambda_0 = \frac{1}{16}$ and $C \sim \mathrm{Trunc}(t_2)$, $t_2 = 19.624$. With this choice of the parameters, the censoring rate when estimating F is 25%. The following notations are used in the simulations below (see the tables):

$\hat{h}_0$ is the minimiser with respect to h of $ISE(h)$;

$\hat{h}_{cv}$ is the minimiser with respect to h of $CV(h)$;

$\hat{h}_{cvn}$ is the minimiser with respect to h of $CVN(h)$.

In Table 1 below are summarised the results of our computations regarding the CVN method on the truncation models. The values of $ISE(\hat{h}_{cvn})$ and $ISE(\hat{h}_0)$ are of comparable magnitude both in terms of mean and standard deviation. This demonstrates that the CVN method works well on truncation models.

Table 2 displays the results of the computations regarding the CV method for truncation models. We can draw the same conclusions as in Table 1. A curious thing to notice in Tables 1 and 2 is that the results do not improve with the censoring rate. This is merely because τ (domain of integration in ISE) is large when the censoring rate is small (25%). Equations (6) and (7) can help to explain this phenomenon.

4.2 Models where C is geometric

The distribution of the random variable C is geometric with parameter $p \in]0, 1[$ (and one writes $C \sim \mathrm{Ge}(p)$) if its mass probability function is given by

$$\mathrm{P}(C = 0) = 0, \quad \mathrm{P}(C = k) = p(1-p)^{k-1}, \quad k = 1, 2, 3, \ldots$$

Table 2. Integrated squared error for $\hat{h}_{cv}$ and $\hat{h}_0$; mean, median and standard deviation over 100 replications.

CV RESULTS							
		$ISE(\hat{h}_{cv})$			$ISE(\hat{h}_0)$		
Model	n	Mean	Median	Standard deviation	Mean	Median	Standard deviation
W-T-50%	200	0.00948	0.00693	0.00879	0.00854	0.00538	0.00861
	400	0.00402	0.00255	0.00470	0.00346	0.00197	0.00447
	800	0.00222	0.00126	0.00269	0.00196	0.00101	0.00249
W-T-25%	200	0.01531	0.01124	0.01342	0.01384	0.01040	0.01286
	400	0.00827	0.00481	0.00883	0.00711	0.00409	0.00806
	800	0.00429	0.00279	0.00499	0.00387	0.00233	0.00466

If the censoring random variable C is geometric, then $\tau = +\infty$ and some care needs to be taken to define $ISE(F_h)$. Because, for β large enough

$$\int_\beta^{+\infty} (F_h(x) - F(x))^2 \, \mathrm{d}x = \int_\beta^{+\infty} (J_n - F(x))^2 \, \mathrm{d}x,$$

where

$$J_n = \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i}{1 - G_n(T_i^-)}.$$

Thus we have

$$\int_\beta^{+\infty} (F_h(x) - F(x))^2 \, \mathrm{d}x = \begin{cases} +\infty & \text{when} \quad \Delta_{(n)} = 0, \\ \int_\beta^{+\infty} (1 - F(x))^2 \, \mathrm{d}x & \text{otherwise.} \end{cases}$$

In order to avoid having $ISE(F_h) = +\infty$, we will use a modified definition of $ISE(h)$. Let $\epsilon > 0$ and choose τ_ϵ such that

$$\int_{\tau_\epsilon}^{\infty} (1 - F(x))^2 \, \mathrm{d}x < \epsilon.$$

Our ISE in this subsection is defined by

$$ISE(h) = ISE(F_h) = \int_0^{\tau_\epsilon} (F_h(x) - F(x))^2 \, \mathrm{d}x.$$

Two models are considered in this subsection. The model (W-G-50%) defined by

$$\text{Model W-G-50\%:} \quad X \sim \mathrm{W}(\alpha_0, \, \lambda_0), \quad C \sim \mathrm{Ge}(p_1),$$

where $\mathrm{W}(\alpha_0, \lambda_0)$ is the Weibull random variable with parameters $\alpha_0 = 1.6$, $\lambda_0 = \frac{1}{16}$, and $C \sim \mathrm{Ge}(p_1)$, $p_1 = 0.0574$. In this model, $\mathrm{P}(X > C) = 0.5$, so that the censoring level is 50% when estimating F. The second model is the Weibull-Geometric-25% (W-G-25%) defined by

$$\text{Model W-G-25\%:} \quad X \sim \mathrm{W}(\alpha_0, \, \lambda_0), \quad C \sim \mathrm{Ge}(p_2),$$

Table 3. Integrated squared error for $\hat{h}_{cvn}$ and $\hat{h}_0$; mean, median and standard deviation over 100 replications.

CVN RESULTS							
		$ISE(\hat{h}_{cvn})$			$ISE(\hat{h}_0)$		
Model	n	Mean	Median	Standard deviation	Mean	Median	Standard deviation
W-G-50%	200	0.05218	0.03602	0.05558	0.04490	0.03428	0.05148
	400	0.02294	0.01731	0.01850	0.02038	0.01525	0.01710
	800	0.01049	0.00857	0.00804	0.00953	0.00776	0.00752
W-G-25%	200	0.02925	0.02215	0.02460	0.02579	0.01837	0.02330
	400	0.01502	0.01091	0.01328	0.01334	0.00842	0.01248
	800	0.00728	0.00484	0.00611	0.00643	0.00432	0.00562

Table 4. Integrated squared error for $\hat{h}_{cv}$ and $\hat{h}_0$; mean, median and standard deviation over 100 replications.

CV RESULTS							
		$ISE(\hat{h}_{cv})$			$ISE(\hat{h}_0)$		
Model	n	Mean	Median	Standard deviation	Mean	Median	Standard deviation
W-G-50%	200	0.05191	0.03648	0.05515	0.04490	0.03428	0.05148
	400	0.02295	0.01678	0.01809	0.02038	0.01525	0.01710
	800	0.01052	0.00856	0.00780	0.00953	0.00776	0.00752
W-G-25%	200	0.02931	0.02203	0.02470	0.02579	0.01837	0.02330
	400	0.01512	0.01101	0.01331	0.01334	0.00842	0.01248
	800	0.00734	0.00488	0.00621	0.00643	0.00432	0.00562

where $\mathrm{W}(\alpha_0, \lambda_0)$ is the Weibull random variable with parameters $\alpha_0 = 1.6$, $\lambda_0 = \frac{1}{16}$, and $C \sim \mathrm{Ge}(p_2)$, $p_2 = 0.0219$. With this choice of the parameters, the censoring rate when estimating F is 25%.

Tables 3 and 4 summarise the results of the computations done to assess the performance of the criterion CVN and CV when the censoring distribution is geometric. The quantities $\hat{h}_0$, $\hat{h}_{cvn}$ and $\hat{h}_{cv}$ have the same meaning as in Subsection 4.1. These tables show that both methods work well even when $\tau = +\infty$. In contrast with Tables 1 and 2 the results improve when the censoring rate decreases.

4.3 Execution time

In the two subsections above, we have compared the CV and CVN criteria statistically. To be a bit more precise, if $\hat{h}_0$ is the minimiser of $ISE(F_h)$, we have compared their performances in estimating $ISE(F_{\hat{h}_0})$. In this subsection we will examine if one bandwidth selector is computationally more

Table 5. Ratios of execution time for heavily censored models.

Model	n	Time(CV)/Time(CVN)
W-T-50%	200	2.457521
	400	2.407998
	800	2.661548
W-G-50%	200	2.117334
	400	2.160770
	800	2.353984

appealing than the other. We have

$$CV(F_h) = \frac{1}{n}\sum_{i=1}^{n} U_i \quad \text{and} \quad CVN(F_h) = \frac{1}{n}\sum_{i=1}^{n} V_i,$$

with

$$U_i = \int_0^{T_{(n)}} \left(F_h^{-i}(x) - \frac{\Delta_i}{1 - G_n(T_i^-)} \mathbb{1}_{[T_i \le x]} \right)^2 \mathrm{d}x,$$

$$V_i = \frac{\Delta_i}{1 - G_n(T_i^-)} \int_0^{T_{(n)}} \left(F_h^{-i}(x) - F_n^{KM}(T_{(n)}) \mathbb{1}_{[T_i \le x]} \right)^2 \mathrm{d}x.$$

From the second equation above, it is obvious that $V_i = 0$ if $\Delta_i = 0$. Thus, when $\Delta_i = 0$, there is no need to compute the integral contained in the formula of V_i. On the other hand, the integral in U_i has to be computed whatever the value of $\Delta_i = 0$. This observation suggest that CVN might be computationally superior to CV. We have checked this assumption on models W-T-50% and W-G-60%. We have computed the time it takes to calculate the results contained in Tables 1-4 for the models W-T-50% and W-G-50%. In these models, censoring is heavy, thus CVN is more likely to outperform CV in terms of computation time. Because a faster computer takes less time to execute a program compared to a slower one, we show below not the time taken by each method, but the time taken by CV over the time taken by CVN. We believe this ratio to be less computer dependent. In Table 5 the results regarding computation time are shown.

5. Conclusions

All our computations in this paper were done when the censoring variable is discrete. However, CV and CVN are devised to work when the censoring random variable is completely arbitrary. To the best of our knowledge CV and CVN are the only methods which can be used to select the bandwidth of the smooth Kaplan–Meier estimator when the censoring variable is discontinuous and there exists no paper containing simulations (on this problem) with discontinuous censoring variable.

It is clear from Tables 1 and 2 that CVN is superior to CV on the models studied (truncation models). When the censoring random variable is geometric the situation is less clear. For the model W-G-50%, the CV method obtains better results than CVN. For the model W-G-25%, CVN

is superior to CV. It is important to point out that the example analysed, that is $X \sim \mathrm{W}(\alpha,\ \lambda)$ (with $\alpha = 1.6$ and $\lambda = \frac{1}{16}$) was chosen so as to make CVN outperform CV when the censoring mechanism is truncation. In fact, experimentally we have found that for any fixed α, if one chooses λ small enough, CVN would outperform CV when the censoring variable is constant. The main thing to point out from this simulation study is that no method is superior to the other one on all models.

The CVN method has a huge advantage over CV (particularly when censoring is heavy), it is computationally more efficient. When $\Delta_i = 0$, there is no need to compute the integral involving F_h^{-i} in the CVN score function, while in CV one has to. The results presented in Table 5 show that CVN is computationally superior to CV when the observations are heavily censored.

A. Proofs

A.1 Proof of Equation (3)

In this proof we are going to use the following equalities proved in Youndjé (2016):

$$\mathrm{E}\left[\Delta \mid X\right] = 1 - G(X^-),$$

$$\mathrm{E}\left[\frac{\Delta \mathbb{1}_{[T\le u]}}{1-G(T^-)}\right] = \begin{cases} F(u), & \text{for} \quad 0 \le u < \tau \\ F(\tau) & \text{for} \quad u \ge \tau. \end{cases}$$

Set

$$E_i = \frac{\Delta_i}{1-G(T_i^-)} \int_0^\tau \left(F_h^{-i}(x) - F(\tau)\mathbb{1}_{[T_i\le x]}\right)^2 \mathrm{d}x,$$

$$T_i = \int_0^\tau \left(F^2(\tau)\mathbb{1}_{[X_i\le x]} - F(\tau)F^2(x)\right) \mathrm{d}x,$$

$$Z_i = \frac{\Delta_i}{1-G(T_i^-)}.$$

We have

$$Z_i = \frac{\Delta_i \mathbb{1}_{[T_i\le\tau]}}{1-G(T_i^-)} = \frac{\Delta_i \mathbb{1}_{[X_i\le\tau]}}{1-G(X_i^-)}.$$

It follows that

$$\mathrm{E}\left[Z_i \mid X_i\right] = \frac{\mathbb{1}_{[X_i\le\tau]}}{1-G(X_i^-)} \mathrm{E}\left[\Delta_i \mid X_i\right] = \mathbb{1}_{[X_i\le\tau]}. \tag{8}$$

To continue, let

$$\overline{E}_i = \mathrm{E}\left[E_i \mid (X_1,\Delta_1),\ldots,(X_{i-1},\Delta_{i-1}), X_i, (X_{i+1},\Delta_{i+1}),\ldots,(X_n,\Delta_n)\right].$$

Observe that

$$E_i = \frac{\Delta_i}{1-G(X_i^-)} \int_0^\tau \left(F_h^{-i}(x) - F(\tau)\mathbb{1}_{[T_i\le x]}\right)^2 \mathrm{d}x.$$

It follows from (8) that

$$\overline{E}_i = \frac{\mathrm{E}\left[\Delta_i \mid X_i\right]}{1-G(X_i^-)} \int_0^\tau \left(F_h^{-i}(x) - F(\tau)\mathbb{1}_{[T_i\le x]}\right)^2 \mathrm{d}x = \mathbb{1}_{[X_i\le\tau]} \int_0^\tau \left(F_h^{-i}(x) - F(\tau)\mathbb{1}_{[T_i\le x]}\right)^2 \mathrm{d}x.$$

We have

$$\begin{aligned}\overline{E}_i - T_i &= \int_0^\tau \Big[\left(F_h^{-i}(x)\right)^2 \mathbb{1}_{[X_i \le \tau]} - 2F_h^{-i}(x)F(\tau)\mathbb{1}_{[X_i \le \tau]}\mathbb{1}_{[X_i \le x]} \\ &\quad + F^2(\tau)\mathbb{1}_{[X_i \le x]}\mathbb{1}_{[X_i \le \tau]} - F^2(\tau)\mathbb{1}_{[X_i \le x]} + F(\tau)F^2(x)\Big]\,\mathrm{d}x \\ &= \int_0^\tau \Big[\left(F_h^{-i}(x)\right)^2 \mathbb{1}_{[X_i \le \tau]} - 2F_h^{-i}(x)F(\tau)\mathbb{1}_{[X_i \le x]} + F(\tau)F^2(x)\Big]\,\mathrm{d}x.\end{aligned}$$

It follows that

$$\begin{aligned}\mathrm{E}(E_i - T_i) &= \mathrm{E}(E_i) - \mathrm{E}(T_i) \\ &= \mathrm{E}(\overline{E}_i) - \mathrm{E}(T_i) \\ &= \mathrm{E}(\overline{E}_i - T_i) \\ &= \int_0^\tau \Big[F(\tau)\mathrm{E}\left(F_h^{-i}(x)\right)^2 - 2F(\tau)\mathrm{E}\left(F_h^{-i}(x)\right)F(x) + F(\tau)F^2(x)\Big]\,\mathrm{d}x \\ &= \int_0^\tau F(\tau)\mathrm{E}\left(F_h^{-i}(x) - F(x)\right)^2\,\mathrm{d}x.\end{aligned}$$

Hence we have

$$\begin{aligned}\mathrm{E}(D(h)) &= \mathrm{E}(CVN^*(h) - T) \\ &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^n (E_i - T_i)\right) \\ &= \frac{1}{n}\sum_{i=1}^n \int_0^\tau F(\tau)\mathrm{E}\left(F_h^{-i}(x) - F(x)\right)^2\,\mathrm{d}x \\ &= \int_0^\tau F(\tau)\mathrm{E}\left(F_h^{-n}(x) - F(x)\right)^2\,\mathrm{d}x.\end{aligned}$$

This last equality is the desired result.

A.2 Proof of Theorem 1 (ii)

We have

$$\begin{aligned}&CVN1(h) - CVN1^*(h) \\ &= \frac{1}{n}\sum_{i=1}^n \left(\frac{\Delta_i}{1 - G_n(T_i^-)} - \frac{\Delta_i}{1 - G(T_i^-)}\right)\mathbb{1}_{[T_i \le R]} \int_0^R \left(F_h^{-i}(x) - F(R)\mathbb{1}_{[T_i \le x]}\right)^2\,\mathrm{d}x \\ &= \frac{1}{n}\sum_{i=1}^n \frac{\Delta_i}{1 - G(T_i^-)}\frac{(G_n(T_i^-) - G(T_i^-))}{1 - G_n(T_i^-)}\mathbb{1}_{[T_i \le R]} \int_0^R \left(F_h^{-i}(x) - F(R)\mathbb{1}_{[T_i \le x]}\right)^2\,\mathrm{d}x.\end{aligned}$$

It follows that

$$\begin{aligned}|CVN1(h) - CVN1^*(h)| = \frac{1}{n}\sum_{i=1}^n \frac{\Delta_i}{1 - G(T_i^-)}\frac{|G_n(T_i^-) - G(T_i^-)|}{1 - G_n(R)}\mathbb{1}_{[T_i \le R]} \\ \times \int_0^R \left(F_h^{-i}(x) - F(R)\mathbb{1}_{[T_i \le x]}\right)^2\,\mathrm{d}x.\end{aligned}$$

Since G_n converges uniformly to G on $[0,\ \tau]$ (see Stute and Wang, 1993), there exists an $n_0 \in \mathbb{N}$, such that $n \geq n_0$ implies

$$|CVN1(h) - CVN1^*(h)| \overset{\text{a.s.}}{\leq} \frac{2}{\gamma}\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{1-G(T_i^-)}|G_n(T_i^-) - G(T_i^-)|\mathbb{1}_{[T_i\leq R]} \times \int_0^R \left(F_h^{-i}(x) - F(R)\mathbb{1}_{[T_i\leq x]}\right)^2 \mathrm{d}x,$$

where $\gamma = 1 - G(R)$. Therefore we have

$$\begin{aligned} &|CVN1(h) - CVN1^*(h)| \\ &\overset{\text{a.s.}}{\leq} \frac{2}{\gamma}\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{1-G(T_i^-)}\mathbb{1}_{[T_i\leq R]}\int_0^R \left(F_h^{-i}(x) - F(R)\mathbb{1}_{[T_i\leq x]}\right)^2 \mathrm{d}x \sup_{0\leq u\leq R}|G_n(u) - G(u)| \\ &= o(CVN1^*(h)). \end{aligned}$$

Conflict of interest. The corresponding author states that there is no conflict of interest.

Acknowledgments. The excellent referees have helped me to improve the presentation, the content and language of this paper. I am profoundly grateful.

References

Bowman, A., Hall, P., and Prvan, T. (1998). Bandwidth selection for the smoothing of distribution functions. *Biometrika*, **85**, 799–808.

Ghorai, J. K. and Susarla, V. (1990). Kernel estimation of a smooth distribution function based on censored data. *Metrika*, **37**, 71–86.

Härdle, W. and Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Annals of Statistics*, **13**, 1465–1481.

Reiss, R.-D. (1981). Nonparametric estimation of smooth distribution functions. *Scandinavian Journal of Statistics*, **8**, 116–119.

Satten, G. A. and Datta, S. (2001). The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average. *American Statistician*, **55**, 207–210.

Stute, W. and Wang, J.-L. (1993). The strong law under random censorship. *Annals of Statistics*, **21**, 1591–1607.

Youndjé, E. (2016). Some heuristics about bandwidth selection for the smooth Kaplan-Meier estimator. *Journal of the Korean Statistical Society*, **45**, 568–580.

Manuscript received 2020-06-03, revised 2021-02-08, accepted 2022-01-10.