

Proceeding Paper

Bootstrap Selector for the Smoothing Parameter of Beran's Estimator [†]

Rebeca Peláez Suárez ^{1,*} , Ricardo Cao Abad ²  and Juan M. Vilar Fernández ² 

¹ Research Group MODES, Department of Mathematics, CITIC, University of A Coruña, 15071 A Coruña, Spain

² Research Group MODES, Department of Mathematics, CITIC, University of A Coruña and ITMATI, 15071 A Coruña, Spain; ricardo.cao@udc.es (R.C.A.); juan.vilar@udc.es (J.M.V.F.)

* Correspondence: rebeca.pelaez@udc.es

† Presented at the 4th XoveTIC Conference, A Coruña, Spain, 7–8 October 2021.

Abstract: This work proposes a resampling technique to approximate the smoothing parameter of Beran's estimator. It is based on resampling by the smoothed bootstrap and minimising the bootstrap approximation of the mean integrated squared error to find the bootstrap bandwidth. The behaviour of this method has been tested by simulation on several models. Bootstrap confidence intervals are also addressed in this research and their performance is analysed in the simulation study.

Keywords: Beran's estimator; survival analysis; bootstrap; bandwidth selector; confidence intervals



Citation: Suárez, R.P.; Abad, R.C.; Fernández, J.M.V. Bootstrap Selector for the Smoothing Parameter of Beran's Estimator. *Eng. Proc.* **2021**, *7*, 28. <https://doi.org/10.3390/engproc2021007028>

Academic Editors: Joaquim de Moura, Marco A. González, Javier Pereira and Manuel G. Penedo

Published: 14 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Let $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$ be a simple random sample of (X, Z, δ) with X being the covariate, $Z = \min\{T, C\}$ the observed variable and $\delta = I_{T \leq C}$ the uncensoring indicator. Usually, T is the time until the occurrence of an event and C is the censoring time. The generalised product-limit estimator of the conditional survival function proposed in [1] is given by

$$\hat{S}_h^B(t|x) = \prod_{i=1}^n \left(1 - \frac{I_{\{Z_i \leq t, \delta_i=1\}} w_{n,i}(x)}{1 - \sum_{j=1}^n I_{\{Z_j < Z_i\}} w_{n,j}(x)} \right) \quad (1)$$

where

$$w_{n,i}(x) = \frac{K((x - X_i)/h)}{\sum_{j=1}^n K((x - X_j)/h)}$$

with $i = 1, \dots, n$ and $h = h_n$ is the bandwidth for the covariable. This estimator depends on a smoothing parameter which is, in practice, unknown. Therefore, finding a method for automatic selection of this bandwidth is truly interesting and very helpful in the analysis of real data subject to censoring. Bootstrap confidence intervals of $S(t|x)$ are also proposed.

2. Bandwidth Selector

Let $rI_h \subset \mathbb{R}$ be an appropriate pilot bandwidth. The bootstrap resampling algorithm consists of generating $U_i \sim U(0, 1)$ and $V_i \sim K$ and obtaining

$$X_i^* = X_{[nU_i]+1} + rV_i,$$

$$Z_i^* = Z_{[nU_i]+1},$$

$$\delta_i^* = \delta_{[nU_i]+1},$$

for each $i = 1, \dots, n$. The bootstrap sample is formed as $\{(X_i^*, Z_i^*, \delta_i^*)\}_{i=1}^n$.

The optimal smoothing parameter is the bandwidth that minimizes the mean integrated squared error given by:

$$MISE_x(h) = E\left(\int (\widehat{S}_h(t|x) - S(t|x))^2 dt\right).$$

Then, the bootstrap bandwidth is obtained by minimizing the Monte Carlo approximation of the bootstrap MISE defined as follows

$$MISE_x^*(h) \simeq \frac{1}{B} \sum_{j=1}^B \left(\int (\widehat{S}_h^{*(j)}(t|x) - \widehat{S}_r(t|x))^2 dt\right),$$

where $\widehat{S}_r(t|x)$ is the Beran survival estimation with pilot bandwidth r using the original sample $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$, $\widehat{S}_h^{*(j)}(t|x)$ is the Beran survival estimation with bandwidth h using the bootstrap resample $\{(X_i^{*(j)}, Z_i^{*(j)}, \delta_i^{*(j)})\}_{i=1}^n$, and B the number of bootstrap resamples.

3. Bootstrap Confidence Intervals

Let $h \in I_h \subset \mathbb{R}$ be an appropriate smoothing parameter and fixed values $(t, x) \in [a, b] \times I$, the bootstrap confidence interval for a confidence level of $1 - \alpha$ is given by

$$\left(\widehat{S}_r(t|x) - \frac{\rho_{1-\alpha/2}}{\sqrt{nh}}, \widehat{S}_r(t|x) - \frac{\rho_{\alpha/2}}{\sqrt{nh}}\right),$$

where $\widehat{S}_r(t|x)$ is the Beran estimation with the pilot bandwidth r that is used in the bootstrap resampling, and $\rho_{\alpha/2}$ and $\rho_{1-\alpha/2}$ are the $100\alpha/2$ and $100(1 - \alpha/2)$ percentiles of the resampling distribution of $\sqrt{nh}(\widehat{S}_h^*(t|x) - \widehat{S}_r(t|x))$, being $\widehat{S}_h^*(t|x)$ the Beran survival estimation of the bootstrap resample.

4. Simulation Study

A simulation study is carried out to analyse the behaviour of the bootstrap algorithm previously described. Several models with different conditional probabilities of censoring were considered. Figure 1 shows the bootstrap estimations of the conditional survival function in two of these scenarios: Model 1 considers the Weibull distribution for life and censoring times and Model 2 considers exponential life and censoring times. Both models have a conditional probability of censoring equal to 0.5. Figure 2 shows the bootstrap confidence intervals in one sample from Models 1 and 2.

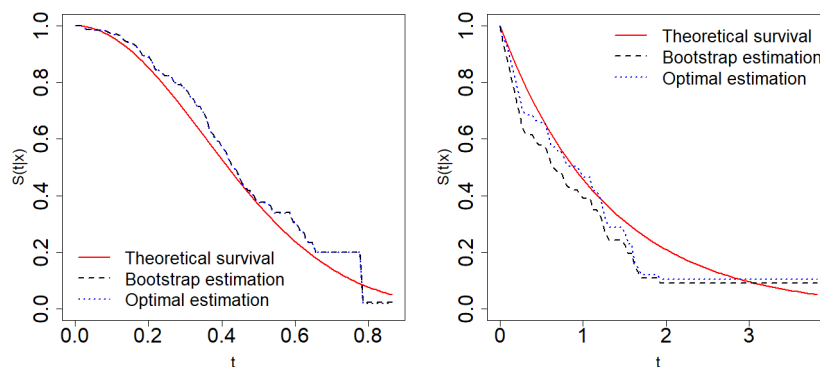


Figure 1. Theoretical survival function $S(t|x)$ (solid line), Beran’s estimation with optimal bandwidth (dotted line) and Beran’s estimation with bootstrap bandwidth (dashed line) for Model 1 (left) and Model 2 (right).

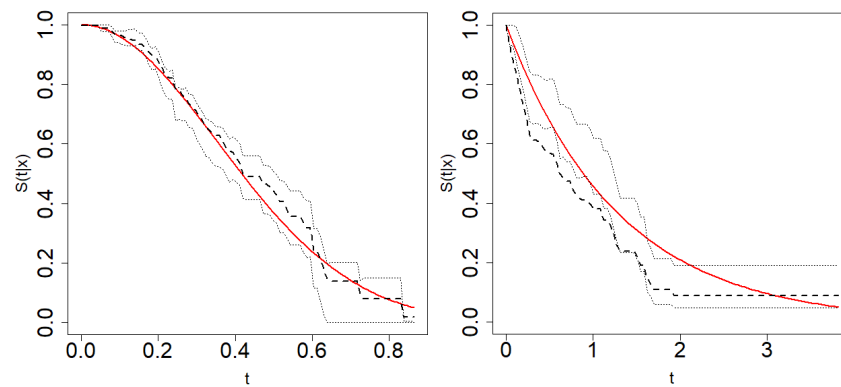


Figure 2. Theoretical survival function (solid line), Beran's estimator with bootstrap bandwidth (dashed line) and the bootstrap confidence intervals (dotted line) for each t in a grid of size $n_t = 100$ in Model 1 (left) and Model 2 (right).

5. Conclusions

The results of the simulations show that this bootstrap algorithm provides adequate smoothing parameters to estimate the survival function in this context. The bootstrap bandwidths obtained are similar to the optimal ones and the estimation errors of both are quite similar. Bootstrap confidence intervals have a reasonable behaviour.

Future lines of work focus on developing a method for choosing the bidimensional smoothing parameters involved in the doubly smoothed Beran estimator presented in [2]. In addition, we deal with the construction of confidence intervals for the conditional survival function based on the doubly smoothed Beran estimator.

Funding: This research has been supported by MINECO Grant MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the ERDF.

References

1. Beran, R. *Nonparametric Regression with Randomly Censored Survival Data*; Technical Report; University of California: Los Angeles, CA, USA, 1981.
2. Peláez Suárez, R.; Cao Abad, R.; Vilar Fernández, J.M. Nonparametric Estimation of the Conditional Survival Function with Double Smoothing. *SORT* **2021**, *45*.