Estimating parameters in the single-index model under censoring: a comparative study

Ewa Strzalkowska-Kominiak
Ricardo Cao

Abstract

In this paper we propose a new method for estimating parameters in a single-index model under censoring based on the Beran estimator for conditional distribution function. This, likelihood based, method is also used for the bandwidth selection. Hence the proposed method is a useful and simple tool for selecting the bandwidth also for the Beran estimator with one dimentional covariate. Additionally, we recall an another method which base on Kaplan-Meier integrals and compare the two approaches in a simulation study. We apply both methods to primary biliary cirrhosis data set and propose the bootstrap test for the parameters.

1 Introduction

The single-index model is a useful tool to incorporate a vector of covariates $X \in \mathbb{R}^d$ into a regression model avoiding the so called "curse of dimensionality". By assuming that there exists a vector of parameters θ_0 so that the response variable depends only on the projection $\theta_0' X$, we avoid a multivariate regression. This assumption is a reasonable compromise between a fully parametric and a fully nonparametric model. Additionally, in medical or economic studies, the large number of explanatory variables is not the only problem. Very often the response variable is only partly observed and so censored from the right. Hence, our goal is to estimate the vector θ_0 using appropriate methods for censored data under the single-index model assumption. In this paper we estimate θ_0 using two different methods for estimating the conditional distribution function (d.f.) and the density under censoring. The first one is based on Kaplan-Meier integrals in the presence of the covariates. See Bouaziz and Lopez (2010), Strzalkowska-Kominiak and Cao (2012) and Stute (1996), for details. The second one is based on the Beran estimator for the conditional d.f. defined in Beran (1981) and studied by González-Manteiga and Cadarso-Suárez (1993). In both cases, the estimation of θ_0 and the choice of bandwidth base on maximum likelihood method.

To describe our model, let Z be a random variable dependent on a vector of covariates $X = (X_1, ..., X_d)'$ and $f(z|\mathbf{x})$ the density function of Z given $X = \mathbf{x}$. Moreover, let

$$\theta_0 = (\theta_1, ..., \theta_d)'$$

be a vector of parameters with the property:

$$f_{\theta_0}(z|\theta_0'\mathbf{x}) = f(z|\mathbf{x}), \tag{1}$$

where $f_{\theta_0}(z|\theta_0'\mathbf{x})$ is the conditional density of Z given $\theta_0'X = \theta_0'\mathbf{x}$. Furthermore, let $F(z|X = \mathbf{x})$ and $F_{\theta_0}(z|\theta_0'X = \theta_0'\mathbf{x})$ be the conditional distribution functions of Z, given X and $\theta_0'X$, respectively. As a consequence of (1) we have

$$F_{\theta_0}(z|\theta_0'X = \theta_0'\mathbf{x}) = F(z|X = \mathbf{x}).$$

Moreover, the random variable Z may be censored from the right by $C \sim G$. Hence, we observe

$$Y = \min(Z, C) \sim H$$

together with

$$\delta = 1_{\{Z \leq C\}}.$$

The paper is organized as follow. In Section 2 we recall the estimator of θ_0 presented in Strzalkowska-Kominiak and Cao (2012) and propose a new method for estimating θ_0 using Beran's estimator. Both methods are compared in Section 3 via simulation study. Finally, the two approaches are illustrated in Section 4 by applying them to a primary biliary cirrhosis data set.

2 Proposed methods

Let us consider the observed censored sample $\{(X_1, Y_1, \delta_1), ..., (X_n, Y_n, \delta_n)\}$. In the next two subsections we will present the two alternative methods to be compared.

2.1 The Kaplan-Meier based approach

In this section we recall the method presented by Strzalkowska-Kominiak and Cao (2012). Let us define,

$$\hat{F}_\theta(y|\theta'\mathbf{x}) = \frac{\frac{1}{nh_1}\sum_{i=1}^n 1_{\{Y_i \le a_n\}} \frac{\delta_i}{1-G_n(Y_i-)} K\left(\frac{\theta'\mathbf{x}-\theta' X_i}{h_1}\right) \mathbb{K}\left(\frac{y-Y_i}{h_2}\right)}{\frac{1}{nh_1}\sum_{i=1}^n 1_{\{Y_i \le a_n\}} \frac{\delta_i}{1-G_n(Y_i-)} K\left(\frac{\theta'\mathbf{x}-\theta' X_i}{h_1}\right)}$$

and

$$\hat{f}_\theta(y|\theta'\mathbf{x}) = \frac{\frac{1}{nh_1h_2}\sum_{i=1}^n 1_{\{Y_i \le a_n\}} \frac{\delta_i}{1-G_n(Y_i-)} K\left(\frac{\theta'\mathbf{x}-\theta' X_i}{h_1}\right) K\left(\frac{y-Y_i}{h_2}\right)}{\frac{1}{nh_1}\sum_{i=1}^n 1_{\{Y_i \le a_n\}} \frac{\delta_i}{1-G_n(Y_i-)} K\left(\frac{\theta'\mathbf{x}-\theta' X_i}{h_1}\right)},$$

where K is a Kernel function (typically $K(u) \geq 0 \; \forall u$ and $\int K(u) = 1$), $\mathbb{K}(y) = \int_{-\infty}^{y} K(z)dz$ and $a_n \to \tau_H$ when $n \to \infty$. The $G_n(t)$ denotes here the Kaplan-Meier estimator for $G(t) = \mathbb{P}(C \le t)$, while $\tau_H = \inf\{t : H(t) = 1\}$.

Moreover, let

$$\hat{l}_n(\theta, h_1, h_2) = \frac{1}{n}\sum_{i=1}^n \Bigg(\delta_i \log \hat{f}_\theta^{-i}(Y_i|\theta' X_i) + (1-\delta_i)\log(1-\hat{F}_\theta^{-i}(Y_i|\theta' X_i))\Bigg) 1_{\{Y_i \le a_n, X_i \in A^{c_n}\}},$$

be the log-likelihood function under censoring, where A^{c_n} is defined so that $\mathbb{P}(X_i \in A^{c_n}) \to 1$ for every $i = 1, ..., n$ when $n \to \infty$. See Strzalkowska-Kominiak and Cao (2012) for the possible choices of a_n and A^{c_n}. Finally, set

$$(\hat{\theta}_n, \hat{h}_1, \hat{h}_2) = \arg\max_{\theta, h_1, h_2} \hat{l}_n(\theta, h_1, h_2). \tag{2}$$

Additionally, we propose the conditional distribution function estimator.

$$\hat{F}^*_{\hat{\theta}_n}(y|\hat{\theta}'_n\mathbf{x}) = \frac{\frac{1}{nh_1}\sum_{i=1}^n \frac{\delta_i}{1-G_n(Y_i-)} K\left(\frac{\hat{\theta}'_n\mathbf{x}-\hat{\theta}'_n X_i}{h_1}\right) 1_{\{Y_i \le y\}}}{\frac{1}{nh_1}\sum_{i=1}^n \frac{\delta_i}{1-G_n(Y_i-)} K\left(\frac{\hat{\theta}'_n\mathbf{x}-\hat{\theta}'_n X_i}{h_1}\right)}. \tag{3}$$

Although we may also use the smoothed version given by

$$\hat{F}_{\hat{\theta}_n}(y|\hat{\theta}_n'\mathbf{x}) = \frac{\frac{1}{nh_1}\sum_{i=1}^n 1_{\{Y_i \le a_n\}} \frac{\delta_i}{1-G_n(Y_i-)} K\left(\frac{\hat{\theta}_n'\mathbf{x}-\hat{\theta}_n' X_i}{h_1}\right) \mathbb{K}\left(\frac{y-Y_i}{h_2}\right)}{\frac{1}{nh_1}\sum_{i=1}^n 1_{\{Y_i \le a_n\}} \frac{\delta_i}{1-G_n(Y_i-)} K\left(\frac{\hat{\theta}_n'\mathbf{x}-\hat{\theta}_n' X_i}{h_1}\right)}.$$

Remark 1 *A similar, Kaplan-Meier based, approach was already proposed by Bouaziz and Lopez (2010). Nevertheless, they estimated the quantity $E(\log f_\theta^\tau(Z|\theta' X)J(X)1_{\{Z\in A_\tau\}})$, where f_θ^τ is the density of Z given $\theta' X$ and $Z \in A^\tau$ while $J(X)1_{\{Z\in A_\tau\}}$ is a trimming function. This kind of likelihood, in case $J(X)1_{\{Z\in A_\tau\}} \equiv 1$, is mostly used in the complete data setup. On the other hand, our likelihood function takes the censoring mechanism into account and outperform the method presented by Bouaziz and Lopez (2010). Both methods differ additionally in the bandwidth selection. See Strzalkowska-Kominiak and Cao (2012), for details.*

Remark 2 *The trimming functions, $1_{\{Y_i \le a_n\}}$ and $1_{\{X_i \in A^{c_n}\}}$, are needed for proving the $\sqrt{n}$-consistency of the estimator $\hat{\theta}_n$. Nevertheless, we showed in Strzalkowska-Kominiak and Cao (2012) that both trimmings can be asymptotically negligible. This means, we may work with sequences a_n and c_n so that $\mathbb{P}(Y_i \le a_n, X_i \in A^{c_n}) \to 1$, when $n \to \infty$. Hence, in the simulation study as well as in the real data example, we set $1_{\{Y_i \le a_n\}} \equiv 1$ and $1_{\{X_i \in A^{c_n}\}} \equiv 1$.*

2.2 The Beran-based approach

In this section we use the Beran conditional d.f. estimator defined in Beran (1981)

$$\tilde{F}_{n\theta}(y|\theta'\mathbf{x}) = 1 - \prod_{i=1}^n \left[1 - \frac{B_{in}(\theta'\mathbf{x})1_{\{Y_i \le y\}}\delta_i}{\sum_{j=1}^n 1_{\{Y_j \ge Y_i\}} B_{jn}(\theta'\mathbf{x})}\right],$$

where

$$B_{in}(\theta'\mathbf{x}) = \frac{K\left(\frac{\theta'\mathbf{x}-\theta' X_i}{h_1}\right)}{\sum_{j=1}^n K\left(\frac{\theta'\mathbf{x}-\theta' X_j}{h_1}\right)}.$$

Moreover, let us define the conditional density and the smoothed d.f. as follows

$$\tilde{f}_\theta(y|\theta'\mathbf{x}) = \frac{1}{h_2}\sum_{j=1}^n W_{jn}(\theta'\mathbf{x}) K\left(\frac{y-Y_j}{h_2}\right) \qquad (4)$$

and

$$\tilde{F}_{\theta}^{S}(y|\theta'\mathbf{x}) = \sum_{j=1}^{n} W_{jn}(\theta'\mathbf{x})\mathbb{K}\left(\frac{y-Y_j}{h_2}\right), \tag{5}$$

where

$$W_{jn}(\theta'\mathbf{x}) = \tilde{F}_{n\theta}(Y_j|\theta'\mathbf{x}) - \tilde{F}_{n\theta}(Y_j - |\theta'\mathbf{x}).$$

Similarly to the previous subsection, let us define the log-likelihood function

$$\tilde{l}_n(\theta, h_1, h_2) = \frac{1}{n}\sum_{i=1}^{n}\bigg(\delta_i \log \tilde{f}_{\theta}^{-i}(Y_i|\theta' X_i) + (1-\delta_i)\log(1-\tilde{F}_{\theta}^{S,-i}(Y_i|\theta' X_i))\bigg),$$

where $\tilde{f}_{\theta}^{-i}(Y_i|\theta' X_i)$ and $\tilde{F}_{\theta}^{S,-i}(Y_i|\theta' X_i)$ are leave-one-out estimators from (4) and (5).

Then, as before, we set

$$(\tilde{\theta}_n, \tilde{h}_1, \tilde{h}_2) = \arg\max_{\theta, h_1, h_2} \tilde{l}_n(\theta, h_1, h_2). \tag{6}$$

Finally, we estimate the conditional distribution function with the well-known Beran estimator

$$\tilde{F}_{n\tilde{\theta}_n}(y|\tilde{\theta}_n'\mathbf{x}) = 1 - \prod_{i=1}^{n}\left[1 - \frac{B_{in}(\tilde{\theta}_n'\mathbf{x})1_{\{Y_i \le y\}}\delta_i}{\sum_{j=1}^{n} 1_{\{Y_j \ge Y_i\}} B_{jn}(\tilde{\theta}_n'\mathbf{x})}\right] \tag{7}$$

or with its smoothed version

$$\tilde{F}_{\tilde{\theta}_n}^{S}(y|\tilde{\theta}_n'\mathbf{x}) = \sum_{j=1}^{n} W_{jn}(\tilde{\theta}_n'\mathbf{x})\mathbb{K}\left(\frac{y-Y_j}{h_2}\right).$$

Remark 3 *A simulation study showed that smoothing the Beran estimator is crucial. If we would define the likelihood function, in terms of the unsmoothed Beran estimator $\tilde{F}_{n\theta}$:*

$$\tilde{l}^*_n(\theta, h_1) = \frac{1}{n}\sum_{i=1}^{n}\Bigg(\delta_i \log(\tilde{F}_{n\theta}(Y_i|\theta' X_i) - \tilde{F}_{n\theta}(Y_i - |\theta' X_i)) + (1-\delta_i)\log(1 - \tilde{F}^{-i}_{n\theta}(Y_i|\theta' X_i))\Bigg), \qquad (8)$$

the selected bandwidth, $\hat{h}_1$, would tend to be very small and the estimator of θ_0 would give a huge mean squared error.

Since $\lim_{x\to\pm\infty} K(x) = 0$, we have

$$\lim_{h_1\to 0^+} K\left(\frac{\theta' X_i - \theta' X_j}{h_1}\right) = \begin{cases} 0, & \text{if } i \neq j \\ K(0). & \text{if } i = j \end{cases}$$

Hence

$$\lim_{h_1\to 0^+} \tilde{F}_{n\theta}(Y_i|\theta' X_i) = \delta_i \quad \text{and} \quad \lim_{h_1\to 0^+} \tilde{F}_{n\theta}(Y_i - |\theta' X_i) = 0.$$

Finally,

$$\lim_{h_1\to 0^+} \tilde{l}^*_n(\theta, h_1) = \frac{1}{n}\sum_{i=1}^{n}\left(\delta_i \log(\delta_i) + (1-\delta_i)\log(1-\delta_i)\right) = 0.$$

This fact together with $\tilde{l}^*_n(\theta, h_1) \leq 0$ $\forall\theta$, $\forall h_1 > 0$ implies that $\tilde{l}^*_n$ attains its maximum for $h_1 \to 0^+$. If, additionally, K has a compact support, we can find a small h_1^* that $\tilde{l}_n(\theta, h_1) = 0$ for all $h_1 \leq h_1^*$.

Remark 4 *The proposed likelihood function $\tilde{l}^*_n$ is a useful tool to select the bandwidth for a Beran estimator also without the single-index model assumption. For this we set $d = 1$ and $\theta = 1$ and maximize $\tilde{l}^*_n$ as a function of h_1 and h_2. Remark, that only the first bandwidth h_1 will be used to compute the conditional Beran estimator $\tilde{F}_n(y|x)$ with one dimensional covariate.*

Remark 5 *In the simulation study, using three different models, the Beran based approach gives better results than Kaplan-Meier method presented in the previous subsection. Nevertheless, on the contrary to Kaplan-Meier approach, proving the asymptotic properties of the parameter vector $\tilde{\theta}_n$ in the Beran case is still an open problem. The difficulty in the proofs is caused by a complicated structure of the weights $W_{jn}(\theta'\mathbf{x})$ in the definition of the conditional density, given by (4), which depend on the parameter θ.*

3 Simulation study

In this section we compare the methods based on Kaplan-Meier and Beran estimator through the simulation study. We consider three different scenarios. In the first and the second one, similarly to Bouaziz and Lopez (2010), the variable of interest Z follows a linear regression, while the censoring variable is generated from the exponential distribution with constant parameter (Model 1) and a parameter dependent on the covariates (Model 2). In the third scenario, we generate Z from Cox proportional hazard model with weibull baseline hazard (Model 3).

Moreover, we select the bandwidths h_1 and h_2 by maximizing $\hat{l}_n$ or $\tilde{l}_n$, respectively. For this we consider two possible strategies: (a) Optimizing the likelihood function over two different bandwidths h_1 and h_2, (b) optimizing the likelihood function over h by setting $h_1 = h\hat{\sigma}(\theta' X)$ and $h_2 = h\hat{\sigma}(Z)$, where $\hat{\sigma}$ is the estimated standard deviation. More precisely, the algorithm follows the steps:

1. Choose some preliminary bandwidths, e.g. $h_1 = n^{-1/7}$, $h_2 = n^{-1/3}$ in case (a) and $h = n^{-1/5}$ in case (b).
2. Maximize $\hat{l}_n$ or $\tilde{l}_n$ with respect to θ.
3. Use the estimated θ to compute $\hat{\sigma}(\theta' X)$ in case (b) and maximize the likelihood $\hat{l}_n$ or $\tilde{l}_n$ with respect to (h_1, h_2) in case (a) and h in case (b).
4. Repeat 1-3 until convergence.

In the following subsection we compare the Kaplan-Meier and Beran based methods for estimating θ_0 in Models 1-3 and summarize the results. In the subsection 3.2 we present results regarding conditional distribution function estimation using both approaches.

3.1 Comparison of the estimators for θ_0

3.1.1 Model 1

Let us consider the model used by Bouaziz and Lopez (2010). Let

$$Z = \theta_0' X + \varepsilon,$$

where $\theta_0 = (1, 0.5, 1.4, 0.2)'$, $X = (X_1, X_2, X_3, X_4)'$,

$$X_i \sim \begin{cases} \mathcal{N}(0,1), & \text{with probability } 0.2 \\ \mathcal{N}(0.25,2), & \text{with probability } 0.8 \end{cases}$$

is a normal mixture for $i = 1,2,3,4$ and $\varepsilon \sim \mathcal{N}(0, |\theta_0' X|)$. Moreover, let $C \sim \exp(\lambda)$, so that we observe only $Y = \min(Z, C)$.

The goal is to estimate θ_0 with $\hat\theta_n = (1, \hat\theta_{n1}, \hat\theta_{n2}, \hat\theta_{n3})'$ using the Kaplan-Meier method and $\tilde\theta_n = (1, \tilde\theta_{n1}, \tilde\theta_{n2}, \tilde\theta_{n3})'$ using the Beran method. Setting $\theta_1 = 1$ guarantees the identifiability of the model.

In this subsection we consider constant λ, so that C is independent of Z.

The following results show the estimated bias, variance and mean squared error (MSE) for different sample sizes and censoring rates. More precisely, we take $n = 100$ and $n = 200$ together with $\lambda = 0.3$ (25% of censoring) and $\lambda = 0.85$ (40 % of censoring). The results are based on 500 trials.

Table 1: Estimated bias, variance and MSE for the Kaplan-Meier method for $\lambda = 0.3$, $n = 100/200$ and 500 trials.

	$n = 100$			$n = 200$		
	$\hat\theta_{n1}$	$\hat\theta_{n2}$	$\hat\theta_{n3}$	$\hat\theta_{n1}$	$\hat\theta_{n2}$	$\hat\theta_{n3}$
	Different bandwidths h_1 and h_2					
Bias	0.0071	0.0152	-0.0032	0.0041	0.0068	0.0045
Variance	0.0295	0.0738	0.0249	0.0097	0.0278	0.0086
MSE	0.1285036			0.0462986		
	Bandwidths $h_1 = h\hat\sigma(\theta' X)$ and $h_2 = h\hat\sigma(Z)$					
Bias	0.0049	0.0229	0.0035	0.0053	0.0124	0.0060
Variance	0.0201	0.0551	0.0175	0.00798	0.0219	0.0066
MSE	0.0932777			0.0367803		

Table 2: Estimated bias, variance and MSE for the Beran method for $\lambda = 0.3$, $n = 100/200$ and 500 trials.

	$n = 100$			$n = 200$		
	$\tilde{\theta}_{n1}$	$\tilde{\theta}_{n2}$	$\tilde{\theta}_{n3}$	$\tilde{\theta}_{n1}$	$\tilde{\theta}_{n2}$	$\tilde{\theta}_{n3}$
	Different bandwidths h_1 and h_2					
Bias	0.0079	0.0169	0.0068	0.0033	0.0012	0.0039
Variance	0.0204	0.0502	0.0182	0.0075	0.0192	0.0066
MSE	0.0891526			0.0333587		
	Bandwidths $h_1 = h\hat{\sigma}(\theta' X)$ and $h_2 = h\hat{\sigma}(Z)$					
Bias	0.0086	0.0240	0.0080	0.0043	0.0044	0.0045
Variance	0.0213	0.0523	0.0183	0.0077	0.0197	0.0066
MSE	0.0927401			0.0340391		

Table 3: Estimated bias, variance and MSE for the Kaplan-Meier method for $\lambda = 0.85$, $n = 100/200$ and 500 trials.

	$n = 100$			$n = 200$		
	$\hat{\theta}_{n1}$	$\hat{\theta}_{n2}$	$\hat{\theta}_{n3}$	$\hat{\theta}_{n1}$	$\hat{\theta}_{n2}$	$\hat{\theta}_{n3}$
	Different bandwidths h_1 and h_2					
Bias	-0.0134	0.0138	-0.0163	0.0034	-0.00003	0.0131
Variance	0.0431	0.1135	0.0407	0.0198	0.0471	0.0172
MSE	0.1980503			0.0842094		
	Bandwidths $h_1 = h\hat{\sigma}(\theta' X)$ and $h_2 = h\hat{\sigma}(Z)$					
Bias	-0.0042	0.0104	-0.0136	0.0081	0.0152	0.0145
Variance	0.0262	0.0701	0.0246	0.0167	0.0371	0.0136
MSE	0.1211611			0.0679582		

3.1.2 Model 2

In this subsection we consider the same model as before, but with $\lambda = \lambda(X)$. Hence there is some dependence between C and Z. More precisely, we take

$$\lambda(X) = \lambda_1 |\theta_0' X|,$$

Table 4: Estimated bias, variance and MSE for the Beran method for $\lambda = 0.85$, $n = 100/200$ and 500 trials.

	$n = 100$			$n = 200$		
	$\tilde{\theta}_{n1}$	$\tilde{\theta}_{n2}$	$\tilde{\theta}_{n3}$	$\tilde{\theta}_{n1}$	$\tilde{\theta}_{n2}$	$\tilde{\theta}_{n3}$
	Different bandwidths h_1 and h_2					
Bias	0.0041	0.0152	-0.0041	-0.0008	0.00002	0.00398
Variance	0.0227	0.0579	0.0198	0.0076	0.0214	0.0082
MSE	0.1007357			0.0372533		
	Bandwidths $h_1 = h\hat{\sigma}(\theta' X)$ and $h_2 = h\hat{\sigma}(Z)$					
Bias	0.0071	0.0163	-0.0058	-0.0004	0.0027	0.0037
Variance	0.0205	0.051	0.0175	0.0075	0.0209	0.0078
MSE	0.0893034			0.0362851		

with $\lambda_1 = 0.15$ and $\lambda_1 = 0.65$, which gives, as above, 25% and 40% of censoring.

Table 5: Estimated bias, variance and MSE for the Kaplan-Meier method for $\lambda_1 = 0.15$, $n = 100/200$ and 500 trials.

	$n = 100$			$n = 200$		
	$\hat{\theta}_{n1}$	$\hat{\theta}_{n2}$	$\hat{\theta}_{n3}$	$\hat{\theta}_{n1}$	$\hat{\theta}_{n2}$	$\hat{\theta}_{n3}$
	Different bandwidths h_1 and h_2					
Bias	-0.0081	0.0075	0.0006	0.0013	0.0037	-0.0068
Variance	0.0504	0.2399	0.0437	0.0141	0.0379	0.0132
MSE	0.3341622			0.0652351		
	Bandwidths $h_1 = h\hat{\sigma}(\theta' X)$ and $h_2 = h\hat{\sigma}(Z)$					
Bias	0.0134	0.0393	0.0075	0.0036	0.0102	-0.0046
Variance	0.0271	0.0799	0.0218	0.0109	0.0314	0.0103
MSE	0.1305688			0.0527837		

Table 6: Estimated bias, variance and MSE for the Beran method for $\lambda_1 = 0.15$, $n = 100/200$ and 500 trials.

	$n = 100$			$n = 200$		
	$\tilde{\theta}_{n1}$	$\tilde{\theta}_{n2}$	$\tilde{\theta}_{n3}$	$\tilde{\theta}_{n1}$	$\tilde{\theta}_{n2}$	$\tilde{\theta}_{n3}$
	Different bandwidths h_1 and h_2					
Bias	0.0047	0.0241	0.0053	-0.0025	0.0005	-0.0027
Variance	0.0188	0.0498	0.0151	0.0067	0.0168	0.0054
MSE	0.0843058			0.0290116		
	Bandwidths $h_1 = h\hat{\sigma}(\theta' X)$ and $h_2 = h\hat{\sigma}(Z)$					
Bias	0.0080	0.0292	0.0058	-0.0018	0.0009	-0.0033
Variance	0.0191	0.0527	0.0142	0.0067	0.0168	0.0055
MSE	0.0868858			0.0289839		

Table 7: Estimated bias, variance and MSE for the Kaplan-Meier method for $\lambda_1 = 0.65$, $n = 100/200$ and 500 trials.

	$n = 100$			$n = 200$		
	$\hat{\theta}_{n1}$	$\hat{\theta}_{n2}$	$\hat{\theta}_{n3}$	$\hat{\theta}_{n1}$	$\hat{\theta}_{n2}$	$\hat{\theta}_{n3}$
	Different bandwidths h_1 and h_2					
Bias	-0.0113	-0.0169	0.0179	0.0134	0.0300	-0.0021
Variance	0.0996	0.3588	0.0753	0.0503	0.1324	0.0400
MSE	0.5344124			0.2239124		
	Bandwidths $h_1 = h\hat{\sigma}(\theta' X)$ and $h_2 = h\hat{\sigma}(Z)$					
Bias	0.0378	0.0969	0.0435	0.02568	0.0620	0.0063
Variance	0.0595	0.2863	0.0588	0.0441	0.1010	0.0373
MSE	0.4173347			0.1869656		

3.1.3 Model 3

In this section we consider the proportional hazard model given by

$$h(t|x) = h_0(t)e^{\theta' X},$$

where the baseline hazard $h_0(t) = 2t$ and hence corresponds to weibull distribution with scale parameter equals 1 and shape parameter equals 2. Moreover, the vector of covariates equals $X = (X_1, X_2, X_3)$, where $X_1 \sim U[0, 10]$, $X_2 \sim \mathcal{N}(0, 2)$, $X_3 \sim \exp(1)$. Additionally, we have a

Table 8: Estimated bias, variance and MSE for the Beran method for $\lambda_1 = 0.65$, $n = 100/200$ and 500 trials.

	$n = 100$			$n = 200$		
	$\tilde{\theta}_{n1}$	$\tilde{\theta}_{n2}$	$\tilde{\theta}_{n3}$	$\tilde{\theta}_{n1}$	$\tilde{\theta}_{n2}$	$\tilde{\theta}_{n3}$
	Different bandwidths h_1 and h_2					
Bias	0.0005	0.0106	0.0152	-0.0014	0.0180	0.0003
Variance	0.0206	0.0523	0.0175	0.0086	0.0219	0.0065
MSE	0.0907719			0.0373466		
	Bandwidths $h_1 = h\hat{\sigma}(\theta' X)$ and $h_2 = h\hat{\sigma}(Z)$					
Bias	0.0046	0.0180	0.0139	0.0010	0.0198	0.0002
Variance	0.016	0.0477	0.0162	0.0083	0.0196	0.0060
MSE	0.0804382			0.0343129		

censoring variable $C \sim \exp(\lambda)$, where $\lambda = 1$ gives us approximately 25% of censoring and $\lambda = 2.5$ corresponds to 36% of censoring. As in the previous sections, our goal is to estimate θ_0 with $\hat{\theta}_n = (1, \hat{\theta}_{n1}, \hat{\theta}_{n2})'$ using the Kaplan-Meier method and $\tilde{\theta}_n = (1, \tilde{\theta}_{n1}, \tilde{\theta}_{n2})'$ using the Beran method.

Table 9: Estimated bias, variance and MSE for the Kaplan-Meier method for $\lambda = 1$, $n = 100/200$ and 500 trials.

	$n = 100$		$n = 200$	
	$\hat{\theta}_{n1}$	$\hat{\theta}_{n2}$	$\hat{\theta}_{n1}$	$\hat{\theta}_{n2}$
	Different bandwidths h_1 and h_2			
Bias	-0.0205	0.0167	0.1286	-0.0286
Variance	0.2405	0.2847	0.6463	0.2041
MSE	0.5258934		0.8678262	
	Bandwidths $h_1 = h\hat{\sigma}(\theta' X)$ and $h_2 = h\hat{\sigma}(Z)$			
Bias	-0.0286	0.0341	-0.0044	-0.0008
Variance	0.0753	0.1207	0.0392	0.0851
MSE	0.1979335		0.1243878	

Table 10: Estimated bias, variance and MSE for the Beran method for $\lambda = 1$, $n = 100/200$ and 500 trials.

	$n = 100$		$n = 200$	
	$\tilde{\theta}_{n1}$	$\tilde{\theta}_{n2}$	$\tilde{\theta}_{n1}$	$\tilde{\theta}_{n2}$
	Different bandwidths h_1 and h_2			
Bias	-0.0319	0.0139	0.0033	-0.0045
Variance	0.0665	0.0613	0.0322	0.0313
MSE	0.1289889		0.0635717	
	Bandwidths $h_1 = h\hat{\sigma}(\theta' X)$ and $h_2 = h\hat{\sigma}(Z)$			
Bias	-0.0323	0.0335	-0.0155	-0.0062
Variance	0.0647	0.1080	0.0297	0.0577
MSE	0.1749557		0.0876373	

Table 11: Estimated bias, variance and MSE for the Kaplan-Meier method for $\lambda = 2.5$, $n = 100/200$ and 500 trials.

	$n = 100$		$n = 200$	
	$\hat{\theta}_{n1}$	$\hat{\theta}_{n2}$	$\hat{\theta}_{n1}$	$\hat{\theta}_{n2}$
	Different bandwidths h_1 and h_2			
Bias	-0.0830	0.0229	-0.0200	0.0177
Variance	0.6268	0.3516	0.6294	0.2146
MSE	0.9858353		0.8447778	
	Bandwidths $h_1 = h\hat{\sigma}(\theta' X)$ and $h_2 = h\hat{\sigma}(Z)$			
Bias	-0.0251	0.0393	-0.0244	0.0431
Variance	0.1002	0.1553	0.0567	0.0934
MSE	0.2576651		0.1525090	

Table 12: Estimated bias, variance and MSE for the Beran method for $\lambda = 2.5$, $n = 100/200$ and 500 trials.

	$n = 100$		$n = 200$	
	$\tilde{\theta}_{n1}$	$\tilde{\theta}_{n2}$	$\tilde{\theta}_{n1}$	$\tilde{\theta}_{n2}$
	Different bandwidths h_1 and h_2			
Bias	-0.0080	0.0257	-0.0166	-0.0065
Variance	0.0577	0.0746	0.2223	0.0340
MSE	0.1330798		0.2566678	
	Bandwidths $h_1 = h\hat{\sigma}(\theta' X)$ and $h_2 = h\hat{\sigma}(Z)$			
Bias	0.0003	0.0209	-0.0037	0.0049
Variance	0.0574	0.1089	0.0283	0.0531
MSE	0.1667591		0.0814847	

3.1.4 Summary of Models 1-3

Summarizing the results given in Tables 1-12, when Z and C are independent (Model 1 and 3) and the censoring equals 25%, the Beran and the Kaplan-Meier estimators are giving similar results if $h_1 = h\hat{\sigma}(\theta' X)$ and $h_2 = h\hat{\sigma}(Z)$. In both models the Beran estimator is better when the censoring is heavy. Moreover, recall that Model 1 was previously considered by Bouaziz and Lopez (2010). Our both methods are mostly better than the one presented there. Finally, if Z and C are independent given X but dependent in general (Model 2), the Beran estimator is always much better than the one based on Kaplan-Meier integrals. The difference becomes larger when the censoring is heavier.

3.2 Comparison of the conditional distribution function estimators.

We compare the Kaplan-Meier and Beran-based estimators for the conditional d.f. through a simulation study using the Kolmogorov-Smirnov (KS) distance. For this, let

$$KS_j^{KM}(\mathbf{x}) = \sup_{y \in \mathbb{R}} |\hat{F}^{*(j)}_{\hat{\theta}_n}(y|\hat{\theta}_n' \mathbf{x}) - F_{\theta_0}(y|\theta_0' \mathbf{x})|$$

and

$$KS_j^{Ber}(\mathbf{x}) = \sup_{y \in \mathbb{R}} |\tilde{F}^{(j)}_{n\tilde{\theta}_n}(y|\tilde{\theta}_n' \mathbf{x}) - F_{\theta_0}(y|\theta_0' \mathbf{x})|$$

be a Kolmogorov-Smirnov distance between the estimated conditional d.f. obtained using the jth trial and the true one for a fixed $\mathbf{x}$ in the Kaplan-Meier and the Beran case, respectively. Set

$$KS_{KM}(\mathbf{x}) = \frac{1}{m}\sum_{j=1}^{m} KS_j^{KM}(\mathbf{x})$$

$$KS_{Beran}(\mathbf{x}) = \frac{1}{m}\sum_{j=1}^{m} KS_j^{Ber}(\mathbf{x})$$

be the estimated mean KS distance based on m trials.

In the following tables we present the KS distance for Models 1-3 (used in the previous subsection) for the Kaplan-Meier and the Beran estimators. We set $\mathbf{x} = (EX_1, EX_2, EX_3, EX_4)' = (0.2, 0.2, 0.2, 0.2)'$ in Models 1 and 2. In Model 3 we take $\mathbf{x} = (EX_1, EX_2, EX_3)' = (5, 0, 1)'$. For a sake of brevity we use only the bandwidths $h_1 = h\hat{\sigma}(\theta' X)$ and $h_2 = h\hat{\sigma}(Z)$.

Table 13: Estimated KS distance for Model 1 with $\lambda = 0.3/0.85$, $n = 100/200$ and 500 trials.

	$\lambda = 0.3$		$\lambda = 0.85$	
	$n = 100$	$n = 200$	$n = 100$	$n = 200$
$KS_{KM}(\mathbf{x})$	0.20615	0.16331	0.25795	0.20675
$KS_{Beran}(\mathbf{x})$	0.19653	0.15883	0.22030	0.18031

Table 14: Estimated KS distance for Model 2 with $\lambda = 0.15/0.65$, $n = 100/200$ and 500 trials.

	$\lambda = 0.15$		$\lambda = 0.65$	
	$n = 100$	$n = 200$	$n = 100$	$n = 200$
$KS_{KM}(\mathbf{x})$	0.19605	0.15921	0.22237	0.18198
$KS_{Beran}(\mathbf{x})$	0.19242	0.15653	0.19941	0.16253

Table 15: Estimated KS distance for Model 3 with $\lambda = 1/2.5$, $n = 100/200$ and 500 trials.

	$\lambda = 1$		$\lambda = 2.5$	
	$n = 100$	$n = 200$	$n = 100$	$n = 200$
$KS_{KM}(\mathbf{x})$	0.2375	0.1926	0.2578	0.2072
$KS_{Beran}(\mathbf{x})$	0.2122	0.1643	0.2229	0.1773

The performance of the Beran-based estimator is always better than that of the Kaplan-Meier one. The difference is more remable when the censoring is heavy. Nevertheless, the computation of the Beran based estimator is much more time consuming.

4 Real data example

To compare our models we analyze a data set on patients with a chronic liver disease, called primary biliary cirrhosis (PBC) from Mayo Clinic. The data can be found in the book by Fleming and Harrington (1991) who analyzed it using Cox regression model.

For our study we use the data set called `pdcRandomSurvivalForest` from the open-source software R. This set includes 424 patients from which 312 were randomized to the treatment with D-penicillamine (DPCA) and placebo. The response variable Z, which is the time to death, may be censored from the right and depends on several explanatory variables. Here we consider the same transformed covariates as in Fleming and Harrington (1991):

$$\begin{aligned} X_1 &= \text{log(bilirubin)} \\ X_2 &= \text{log(algumin)} \\ X_3 &= \text{age/365 \ \ (in years)} \\ X_4 &= \text{edema} \\ X_5 &= \text{log(prothrombin \ \ time)} \\ X_6 &= \text{treatment (1=DPCA, 2=placebo)} \end{aligned}$$

In the following we estimate θ_0 in the single-index model based on 312 data of patients randomized to placebo and DPCA. Since Fleming and Harrington (1991) have shown that the treatment has a negligible effect on the prognosis, we analyze the data first with the treatment variable and then without it. More precisely, we estimate the vector θ_0 using all covariates $X_1, ..., X_6$ (Model I), and then removing the treatment variable, X_6, and repeating the analysis based on $X_1, ..., X_5$ (Model II).

Let $X = (X_1, X_2, X_3, X_4, X_5, X_6)'$ be the vector of covariates. We define $\theta_0 = (1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6)'$ in Model I and $\theta_0 = (1, \theta_2, \theta_3, \theta_4, \theta_5)'$ in Model II.

The following table presents the result of the Kaplan-Meier and Beran-based estimators for θ_0 defined in (2) and (6), respectively.

Table 16: Estimated parameters in Model I

Kaplan-Meier					
$\hat{\theta}_{n1}$	$\hat{\theta}_{n2}$	$\hat{\theta}_{n3}$	$\hat{\theta}_{n4}$	$\hat{\theta}_{n5}$	$\hat{\theta}_{n6}$
1	-5.1456	0.0432	1.8707	3.0280	0.3934
Beran					
$\tilde{\theta}_{n1}$	$\tilde{\theta}_{n2}$	$\tilde{\theta}_{n3}$	$\tilde{\theta}_{n4}$	$\tilde{\theta}_{n5}$	$\tilde{\theta}_{n6}$
1	-3.3569	0.0422	1.0854	4.2438	0.2706

Table 17: Estimated parameters in Model II

Kaplan-Meier				
$\hat{\theta}_{n1}$	$\hat{\theta}_{n2}$	$\hat{\theta}_{n3}$	$\hat{\theta}_{n4}$	$\hat{\theta}_{n5}$
1	-5.7731	0.0408	1.9961	3.1193
Beran				
$\tilde{\theta}_{n1}$	$\tilde{\theta}_{n2}$	$\tilde{\theta}_{n3}$	$\tilde{\theta}_{n4}$	$\tilde{\theta}_{n5}$
1	-3.2684	0.0398	0.9928	4.2105

Within Tables 16 and 17 the results of the Kaplan-Meier and the Beran estimator for θ_0 are similar. Moreover, by comparing Tables 16 and 17, the estimated θ_0 is not changing much when we remove the treatment variable X_6. This would agree with the result in Fleming and Harrington (1991), that the patient's lifetime does not depend on the treatment. In the following, we propose a bootstrap based procedure to confirm this result. We apply the bootstrap method proposed by Iglesias-Pérez and González-Manteiga (2003) for censored data. In order to check the hypothesis $\mathbb{H}_0 : \theta_6 = 0$ we propose, for Beran based method, the following bootstrap test:

1. We choose B and for $k = 1, ..., B$ we repeat the steps:

   a) Basing on $\tilde{\theta}_n = (\tilde{\theta}_{n1}, ..., \tilde{\theta}_{n5})$ from Table 17 and $\tilde{X}_i = (X_{1i}, ..., X_{5i})'$ we generate, for $i = 1, ..., n$, T_i^* from $\tilde{F}_{n\tilde{\theta}_n}(y|\tilde{\theta}_n' \tilde{X}_i)$.

   b) We generate C_i^* from $G_n(y)$ for $i = 1, ..., n$.

   c) Set $Y_i^* = \min(T_i^*, C_i^*)$ and $\delta_i^* = 1_{\{T_i^* \leq C_i^*\}}$ for $i = 1, ..., n$.

   d) Basing on the bootstrap sample (Y_i^*, δ_i^*) and $\tilde{X}_i^* = (X_{1i}, ..., X_{5i}, X_{6i})'$, we compute $\tilde{\theta}_n^* = (\tilde{\theta}_{n1}^*, ..., \tilde{\theta}_{n5}^*, \tilde{\theta}_{n6}^*)$.

e) We set $\tilde{\theta}^*_{n6(k)} = \tilde{\theta}^*_{n6}$.

2. For a given level α we estimate from the sample $\tilde{\theta}^*_{n6(1)}, ..., \tilde{\theta}^*_{n6(B)}$, the $\frac{\alpha}{2}100\%$ and $(1-\frac{\alpha}{2})100\%$ quantiles ($\tilde{q}^*_{\frac{\alpha}{2}}$ and $\tilde{q}^*_{1-\frac{\alpha}{2}}$) of the asymptotic distribution of the $\tilde{\theta}_{n6}$ under $\mathbb{H}_0$.

3. If $\tilde{\theta}_{n6}$, from Table 16, belongs to the confidence interval $[\tilde{q}^*_{\frac{\alpha}{2}}, \tilde{q}^*_{1-\frac{\alpha}{2}}]$ the null hypothesis ($\mathbb{H}_0 : \theta_6 = 0$) will not be rejected.

Similarly, we test H_0 using Kaplan-Meier estimator $\hat{F}_{n\theta}(y|\theta'\mathbf{x})$ with $\hat{\theta}_n$. The confidence interval, using the Kaplan-Meier based estimator, will be denoted by $[\hat{q}^*_{\frac{\alpha}{2}}, \hat{q}^*_{1-\frac{\alpha}{2}}]$. In this case, the null hypothesis ($\mathbb{H}_0 : \theta_6 = 0$) wont be rejected, if $\hat{\theta}_{n6} \in [\hat{q}^*_{\frac{\alpha}{2}}, \hat{q}^*_{1-\frac{\alpha}{2}}]$, where $\hat{\theta}_{n6}$ is given in Table 16.

The results for the bootstrap confidence intervals, for $\alpha = 0.1$ and $B = 1000$, are as follow

$$CI_{KM} = [\hat{q}^*_{\frac{\alpha}{2}}, \hat{q}^*_{1-\frac{\alpha}{2}}] = [-0.49, 0.53] \text{ and } CI_{Beran} = [\tilde{q}^*_{\frac{\alpha}{2}}, \tilde{q}^*_{1-\frac{\alpha}{2}}] = [-0.40, 0.44].$$

Since, $\hat{\theta}_{n6} = 0.3934 \in CI_{KM}$ and $\tilde{\theta}_{n6} = 0.2706 \in CI_{Beran}$, the null hypothesis, $\mathbb{H}_0 : \theta_6 = 0$, is not rejected. Additionally, we may estimate p-values with $\hat{p}_{KM}$ and $\hat{p}_{Beran}$ given by

$$\hat{p}_{KM} = \frac{1}{B}\sum_{k=1}^{B} 1_{\{|\hat{\theta}^*_{n6(k)}| \geq 0.3934\}} = 0.186 \text{ and } \hat{p}_{Beran} = \frac{1}{B}\sum_{k=1}^{B} 1_{\{|\tilde{\theta}^*_{n6(k)}| \geq 0.2706\}} = 0.295.$$

This confirms the result of Fleming and Harrington (1991), that the patient's lifetime T does not depend on the treatment.

Finally, we give some examples of estimated conditional distribution function. For clarity, let $F_{KM}(t|\hat{\theta}'_n\mathbf{x})$ denote the Kaplan-Meier estimator defined in (3) and $F_B(t|\tilde{\theta}'_n\mathbf{x})$ denote the Beran estimator defined in (7).

The following figure presents the estimated conditional distribution functions for the Model I, where the treatment variable is present. Let

$$\mathbf{x}_1 = (0.58, 1.25, age, 0, 2.37, 2)' \text{ where } age \in \{40, 50, 60\}.$$

Here the differences between the $F_{KM}(t|\hat{\theta}'_n\mathbf{x}_1)$ and $F_B(t|\tilde{\theta}'_n\mathbf{x}_1)$ are caused by the heavy censoring of 60%. By definition, the Kaplan-Meier based estimator F_{KM} is jumping to one, at the last order statistic, while Beran's estimator, F_B, is not. Nevertheless, both curves are showing the same tendencies. This means, the probability to fail before time t is increasing with age (Figure 1).

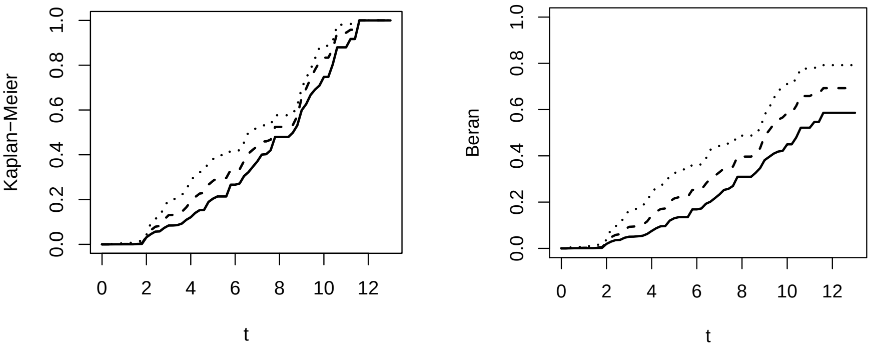

Figure 1: Estimators $F_{KM}(t|\hat{\theta}'_n\mathbf{x}_1)$ (left) and $F_B(t|\tilde{\theta}'_n\mathbf{x}_1)$ (right), for $age = 40$ (solid line), $age = 50$ (dashed line) and $age = 60$ (dotted line)

Acknowledgement

The authors acknowledge financial support from Ministerio de Economía y Competitividad Grants MTM2008-00166 and MTM2011-22392 (EU ERDF support included). Additionally, Ewa Strzalkowska-Kominiak acknowledges financial support from a Juan de la Cierva scholarship, from the Spanish Ministerio de Economía y Competitividad.

References

Beran, R., 1981. Nonparametric regression with randomly censored survival data. Technical Report, University of California, Berkley.

Bouaziz, O., Lopez, O., 2010. Conditional density estimation in a censored single-index model. *Bernoulli* **16**, 514–542.

Fleming, T., Harrington, D., 1991. Counting Processes and Survival Analysis. New York: Wiley.

González-Manteiga, W., Cadarso-Suárez, C., 1993. Asymptotic properties of a generalized Kaplan-Meier estimator with some applications. *Journal of Nonparametric Statistics* **4**, 65–78.

Iglesias-Pérez, M. C., González-Manteiga, W., 2003. Bootstrap for the conditional distribution function with truncated and censored data. *Annals of the Institute of Statistical Mathematics* **55**, 331–357.

Kaplan, E. L., Meier, P., 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457-481.

Strzalkowska-Kominiak, E., Cao, R., 2012. Maximum likelihood estimation for conditional distribution single-index models under censoring. Submitted.

Stute, W., 1996. Distributional convergence under random censorship when covariables are present. *Scandinavian Journal of Statistics* **23**, 461–471.