L_1-Statistical Procedures and Related Topics
IMS Lecture Notes – Monograph Series (1997) Volume 31

Bootstrap selection of the smoothing parameter in density estimation under the Koziol-Green model

Jacobo de Uña-Álvarez

University of Vigo, Spain

Wenceslao González-Manteiga and Carmen Cadarso-Suárez

University of Santiago de Compostela, Spain

Abstract: An asymptotic representation of the mean weighted integrated squared error for the kernel estimator of the density under the Koziol-Green model of proportional censorship is obtained for a bootstrap resampling method. A new bandwidth selector based on the bootstrap is consequently proposed. Simulation results for different models using WARPed versions of the estimators show how the bootstrap selector behaves appreciably better than the classical cross-validation method. Finally a real example is analyzed.

Key words: Bandwidth selection, bootstrap, Koziol-Green model, mean integrated squared error, warping.

AMS subject classification: 62G07.

1 Introduction

A typical situation in survival analysis: let Y be the variable of interest with continuous distribution function F, let C be the right-censoring variable with continuous distribution function G, and let (Z, δ) be the observed pair, *i.e.*, $Z = \min(Y, C)$ and $\delta = 1_{\{Y \leq C\}}$. The general random censorship model assumes the independence between Y and C. Hence $E(\delta) = \int (1 - G) dF$ and the (continuous) distribution function H of Z satisfies $1 - H = (1 - F)(1 - G)$.

The Koziol-Green model of proportional censorship (Koziol and Green,

1976) is an interesting sub-model of the above one obtained by imposing the additional parametric assumption

$$1 - G = (1 - F)^{\beta} \text{ for some } \beta > 0. \tag{1}$$

A crucial fact about (1) is that the independence between Z and δ characterizes the model (Sethuraman, 1965); this allows to construct hypothesis tests about such a model (see, for example, Herbst, 1992; Henze, 1993).

Under (1) it is true that $1 - F = (1 - H)^{\theta}$, with $\theta = (1 + \beta)^{-1} = E(\delta)$. This relation motivates the ACL (Abdushukurov, 1984; Cheng and Lin, 1984) estimator

$$1 - \widetilde{F}_n = (1 - H_n)^{\theta_n}, \tag{2}$$

where H_n is the empirical distribution function of the Z's and θ_n is the sample mean of the δ's, given the initial sample $\{(Z_1, \delta_1), ..., (Z_n, \delta_n)\}$. The estimator (2) is the maximum-likelihood estimator of the survival function of interest under the model (1).

Here we are interested in the estimation of the density f of Y (assumed to exist). A kernel estimator is defined in Section 2. How to choose the bandwidth for it is our main question. Different procedures have been considered and studied in the uncensored case. See Cao *et al.* (1994) for a comparative study. Least squares cross-validation (LSCV) has been adapted to density estimation under proportional censorship by Ghorai and Pattanaik (1993). These authors have established asymptotic optimality, in the sense

$$\frac{ISE_w(h_{CV})}{\inf_{h \in L_n} ISE_w(h)} \to 1 \text{ a.s.},$$

where the cross-validation bandwidth h_{CV} is the minimizer of the score function (4) defined in Section 2, ISE_w is the integrated squared error $ISE_w(h) = \int (\widetilde{f}_h - f)^2 w$, for a suitable weight function w (with the role of eliminating endpoint effects), $\widetilde{f}_h$ is the estimator defined in (3), and the set L_n follows the usual regularity conditions (that can be found in the mentioned work).

In a recent paper González-Manteiga *et al.* (1996) motivate the search for improved methods of bandwidth selection. Although the quality of the resulting density estimator is the crucial question, the rate of convergence for cross-validation type selectors is known to be very slow. In the referred work "smoothed bootstrap" ideas are the base of a new criterion for choosing the bandwidth in censored hazard rate estimation. Here these considerations are translated to the context of density estimation under the Koziol-Green model. In Section 2 we introduce a bootstrap mechanism to select the parameter h for the estimator $\widetilde{f}_h$.

We discuss some fast algorithms based on WARP ideas (see Härdle, 1991; Fan and Marron, 1994) in Section 3, introducing methods for computing both the estimator (3) and also the two bandwidth selectors considered here. In Section 4 we present our simulation results for different proportional censorship models. As in González-Manteiga *et al.* (1996), the bootstrap selector behaves appreciably better than the classical cross-validation method. We also present a real example that follows the Koziol-Green model.

2 The estimator: LSCV and bootstrap bandwidth selection

A natural kernel estimator for the true parameter of interest f is given by

$$\widetilde{f}_h(y) = \int K_h(y-v)\widetilde{F}_n(dv) \simeq \sum_{i=1}^{n} K_h(y-Z_i)n^{-1}\theta_n(1-H_n(Z_i))^{\theta_n-1} \quad (3)$$

where $K_h(.) = K(./h)/h$ is the rescaled kernel function K with smoothing parameter h satisfying $h \to 0$ and $nh \to \infty$ as $n \to \infty$. This estimator was considered by Csörgo and Mielniczuk (1988). These authors proved results on strong consistency, asymptotic normality and Bickel-Rosenblatt type confidence bands for (3).

The LSCV bandwidth h_{CV} considered here is the minimizer of

$$CV(h) = \int \widetilde{f}_h^2 w - 2\sum_{i=1}^{n} \widetilde{f}_{h,i}(Z_i)n^{-1}\theta_n(1-H_{n,i}(Z_i))^{\theta_n-1}w(Z_i) \quad (4)$$

where $\widetilde{f}_{h,i}$ and $H_{n,i}$ are the "leave-one-out" versions of $\widetilde{f}_h$ and H_n respectively, given by

$$\widetilde{f}_{h,i}(y) = \sum_{j\neq i} K_h(y-Z_j)(n-1)^{-1}\theta_n(1-H_{n,i}(Z_j))^{\theta_n-1}$$

and

$$H_{n,i}(y) = (n-1)^{-1}\sum_{j\neq i} 1_{\{Z_j\leq y\}}.$$

The function $CV(h)$ is an estimator of $MISE_w(h) - \int f^2 w$, where

$$MISE_w(h) = E\int(\widetilde{f}_h - f)^2 w \quad (5)$$

is the mean weighted integrated squared error.

We obtain the next result, giving an asymptotic representation of (5). We make the following assumptions: (i) the density f is continuous, (ii)

the function w has compact support contained in $(0,T)$, where T satisfies $1-H(T)>0$, and (iii) the kernel K is a square integrable probability density function with compact support.

Theorem 1 *Under the assumptions (i)-(iii):*

$$MISE_w(h) = AMISE_w(h) + o((nh)^{-1})$$

where

$$AMISE_w(h) = (nh)^{-1}R(K)\int \theta(1-H)^{\theta-1}fw + \int (K_h * f - f)^2 w$$

and $R(K) = \int K^2$ *(* $*$ *denotes convolution).*

From now on we define the bootstrap bandwidth selector h^* as the minimizer of an estimated $AMISE_w$. "Smoothed bootstraps" are required to approximate the "bias" part of $MISE_w$. We therefore propose the resampling plan:

1. Draw bootstrap resamples $\{Z_1^*,...,Z_n^*\}$ from $H_n * K_g$.
2. Generate independent bootstrap resamples $\{\delta_1^*,...,\delta_n^*\}$ from a Bernoulli distribution with parameter θ_n.

The distribution $H_n * K_g$ denotes the one having density $\widehat{h}_g(y) = \int K_g(y-v)H_n(dv)$. The bootstrap versions of the estimator (3) and the error (5) are respectively

$$\widetilde{f}_h^*(y) = \sum_{i=1}^{n} K_h(y-Z_i^*)n^{-1}\theta_n^*(1-H_n^*(Z_i^*))^{\theta_n^*-1}$$

and

$$MISE_w^*(h) = E^*\int (\widetilde{f}_h^* - \widetilde{f}_g)^2 w$$

where H_n^* is the empirical distribution function of the Z^*'s and θ_n^* is the sample mean of the δ^*'s. Similar arguments to those used in the proof of Theorem 1 lead to:

$$AMISE_w^*(h) = (nh)^{-1}R(K)\int \theta_n(1-H_n*K_g)^{\theta_n-1}\widehat{f}_g w + \int (K_h*\widehat{f}_g - \widetilde{f}_g)^2 w \tag{6}$$

where $\widehat{f}_g = \theta_n(1-H_n*K_g)^{\theta_n-1}\widehat{h}_g$ is another estimator of f under (1). Then h^* is the minimizer of (6). This expresion shows that our bootstrap design mimics the theoretical $AMISE_w$ in Theorem 1.

Remark 1 *We opted to deal with the pilot estimator* $\widetilde{f}_g$ *instead of the theoretical bootstrap density* $\widehat{f}_g$ *in the definition of* $MISE_w^*$. *An analogous expression for the asymptotic* $MISE_w^*$ *can be obtained using* $\widehat{f}_g$.

3 Warping algorithms

Here we introduce fast algorithms to construct the estimator and to select the bandwidth, both for the LSCV and the bootstrap methods. As in Härdle (1991) and Fan and Marron (1994), the idea is to "bin" the data into an equally spaced grid, so that the number of kernel evaluations can be drastically reduced. Our bins are $B_z = \left[\frac{zh}{M}, \frac{(z+1)h}{M}\right)$, $z \in Z$, where every interval has length $\delta = h/M$ with $M \in Z^+$.We summarize the observed data by the $n_z = \sum_{i=1}^{n} 1_{B_z}(Z_i)$, $z \in Z^+$, the number of observations in the bin B_z. The large number of kernel differences $K_h(y - Z_i)$ is approximated by the much smaller set of values $w_M(k) = K(k/M)$, for $k = 1-M, ..., M-1$, when K is supported on $[-1, 1]$. Fan and Marron (1994) showed that this results in computational speed savings of factors up to 100.

3.1 WARPing the kernel estimator

A WARPing approximation of $\widetilde{f}_h$ at B_z is given by

$$\widetilde{f}_M(z) = (n\delta M)^{-1} \sum_{k=1-M}^{M-1} w_M(k)\theta_n(1 - H_M(z+k))^{\theta_n - 1} n_{z+k} \qquad (7)$$

where $1 - H_M(z) = n^{-1} \sum_{k>z} n_k$. For a fixed h and letting $M \to \infty$, we have $\widetilde{f}_M(y) \to \widetilde{f}_h(y)$ (the WARPing approximation error decreases with the rounding error δ). This point is illustrated in Figure 1.

3.2 WARPing cross-validation

A WARPed version of the score function (4) in Section 2 involves essentially replacing the conventional kernel estimators by their WARPed versions, although additional work is required to get a rapidly computable version. We define the "leave out bin counts" $n_z^{-i} = \sum_{j \neq i} 1_{B_z}(Z_j)$, $z \in Z$, and we define $1 - H_M^{-i}(z) = (n-1)^{-1} \sum_{k>z} n_k^{-i}$.

Now we use the approximations

$$\int \widetilde{f}_h^2 w \simeq \delta \sum_z \widetilde{f}_M^2(z) w_{M,h}(z)$$

and

$$-2n^{-1} \sum_{i=1}^{n} \widetilde{f}_{h,i}(Z_i)\theta_n(1 - H_{n,i}(Z_i))^{\theta_n - 1} w(Z_i)$$

$$\simeq -\frac{2}{n-1} \sum_z \widetilde{f}_M(z)\theta_n(1 - H_M(z))^{\theta_n - 1} n_z w_{M,h}(z)$$

$$+\frac{2w_M(0)}{n(n-1)\delta M} \sum_z n_z w_{M,h}(z)(\theta_n(1 - H_M(z))^{\theta_n - 1})^2$$

where $w_{M,h}(z)$ denotes the value of w at the lower limit of B_z. Our approximated score function becomes

$$\overline{CV}(M) = \delta \sum_z \widetilde{f}_M^2(z) w_{M,h}(z) - \frac{2}{n-1} \sum_z \widetilde{f}_M(z) \theta_n (1 - H_M(z))^{\theta_n - 1} n_z w_{M,h}(z) + \frac{2 w_M(0)}{n(n-1)\delta M} \sum_z n_z w_{M,h}(z) (\theta_n (1 - H_M(z))^{\theta_n - 1})^2. \tag{8}$$

Our proposal is to fix the rounding error δ and then to minimize the function (8) in M.

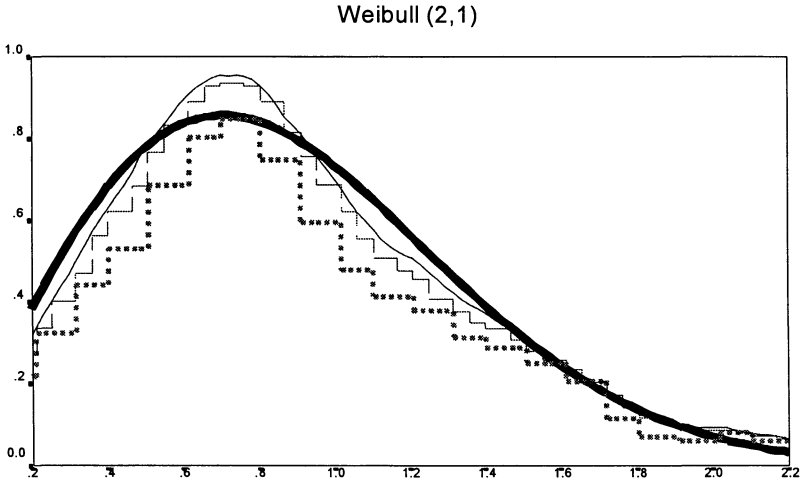

Figure 1: True density function (thick solid line) for a Weibull(2,1) model, the kernel estimator (thin solid line) (computed directly on a sample of 100 uncensored observations of such a model), and two warping approximations of this curve: δ=0.1 (dotted line) and δ=0.01 (dashed line). The role of the rounding error becomes clear; the approximation to the kernel estimator gets better as δ decreases to zero.

3.3 WARPing the smoothed bootstrap

A WARPed version of the function (6) is obtained similarly by replacing the conventional kernel estimates with their corresponding WARPed versions. An approximation of $AMISE_w^*(h)$ is given by

$$\overline{AMISE^*}(M) = \delta \sum_z (M^{-1} \sum_{k=1-M}^{M-1} w_M(k) \widehat{f}_{M_1}(z+k) - \widetilde{f}_{M_1}(z))^2 w_{M,h}(z) + (nh)^{-1} R(K) \delta \sum_z \theta_n (1 - \widehat{H}_{M_1}(z))^{\theta_n - 1} \widehat{f}_{M_1}(z) w_{M,h}(z) \tag{9}$$

where $\widehat{f}_{M_1}$ and $\widehat{H}_{M_1}$ are the WARPed approximations of $\widehat{f}_g$ and $H_n * K_g$ with parameter $M_1 = g/\delta$ at B_z, given respectively by

$$\widehat{H}_{M_1}(z) = (nM_1)^{-1} \sum_{k=1-M}^{M-1} w_{M_1}(k)\overline{n}_{z+k}, \qquad \text{with } \overline{n}_z = \sum_{j=-\infty}^{z} n_j, \text{ and}$$

$$\widehat{f}_{M_1}(z) = \theta_n(1 - \widehat{H}_{M_1}(z))^{\theta_n - 1}\widehat{h}_{M_1}(z), \qquad \text{where}$$

$$\widehat{h}_{M_1}(z) = (n\delta M)^{-1} \sum_{k=1-M}^{M-1} w_{M_1}(k) n_{z+k}.$$

4 Simulation study. Example

4.1 Simulation study

In this subsection we compare LSCV and bootstrap bandwidth selectors for moderate sample sizes. We considered three underlying densities:

Weibull (α, λ). The density of interest is taken as $f = f_{\alpha,\lambda}$, satisfying $f(x) = f_{\alpha,\lambda}(x) = \alpha\lambda(\lambda x)^{\alpha-1}e^{-(\lambda x)^\alpha}1_{[0,\infty)}(x)$ with $\alpha, \lambda > 0$.

Gumbel (α, λ). Its density is $f(x) = f_{\alpha,\lambda}(x) = \alpha\lambda e^{\lambda x - \alpha(e^{\lambda x}-1)}1_{[0.\infty)}(x)$ with $\alpha, \lambda > 0$.

Truncated normal model. We consider the density of the random variable $Y = X \mid X \geq 0$,where $X \in N(\mu, \sigma)$; that is, $f(x) = \frac{\phi_{\mu,\sigma}(x)}{1-\Phi_{\mu,\sigma}(0)}1_{[0,\infty)}(x)$, where $\phi_{\mu,\sigma}(x) = \Phi'_{\mu,\sigma}(x)$ and $\Phi_{\mu,\sigma}$ is the distribution of X.

This simulation study was carried out as in González-Manteiga *et al.* (1996) (see this article for details).

Table 1 contains the results of 1.000 trials of sample size 100 corresponding to the following models:

- Weibull models with parameters λ=1 and α=1,2,3 without censoring (denoted by W(1,1), W(2,1) and W(3,1)) and also with 25% of censoring (denoted by CW(1,1), CW(2,1) and CW(3,1)).
- Gumbel models with parameters λ=1 and α=1,2,3 without censoring (G(1,1), G(2,1) and G(3,1)) and with 25% of censoring (CG(1,1), CG(2,1) and CG(3,1)).
- Truncated normal distributions with parameters μ=1 and σ=0.5 for an uncensored situation and also with a censoring of 25% (N(1,0.5), CN(1, 0.5)).

Although the only sample size in the simulations presented here is n=100, similar results were observed for n=50 and n=200. The triangular kernel was used.

	Mean		Std.Dev.	
MODEL	hcv	h*	hcv	h*
W(1,1)	20.96	15.21	14.54	8.50
CW(1,1)	21.57	18.27	13.55	9.69
W(2,1)	16.64	10.75	19.14	9.21
CW(2,1)	16.01	11.38	20.19	8.61
W(3,1)	21.62	14.39	26.47	12.32
CW(3,1)	21.45	17.31	26.64	13.75
G(1,1)	22.84	15.78	22.97	11.15
CG(1,1)	23.46	16.36	22.44	9.72
G(2,1)	42.58	30.63	33.01	19.02
CG(2,1)	44.12	34.59	31.60	19.39
G(3,1)	60.83	45.94	42.03	27.14
CG(3,1)	64.29	53.52	42.63	29.25
N(1,0.5)	14.82	10.41	16.78	8.81
CN(1,0.5)	19.16	13.17	25.12	10.38

Table 1: Mean and standard deviation of the integrated squared error (L_2-norm) of the kernel density estimator with h_{CV} and h^* bandwidths along 1.000 trials of size 100.

The values in Table 1 are the mean and the standard deviation of the ISE_w (L_2-norm) $\int(\widetilde{f}_{\widehat{h}} - f)^2 w$ along the 1.000 samples of size 100. The columns headed h_{CV} report the ISE_w for estimates based on cross-validation bandwidths (resulting from the minimization of $\overline{CV}(M)$ in expression (8)) and the columns headed h^* report the ISE_w for estimates based on bootstrap bandwidths (resulting from the minimization of $\overline{AMISE^*}(M)$ in expression (9)). Both minimizers were taken as the global minimizer over a fine grid. In both cases the rounding error in the WARP approximation was δ=0.01 and the weighting function was $w(u) = 1_{[H^{-1}(0.05),H^{-1}(0.95))}(u)$. Finally, the pilot bandwidth g used is given by $g = \left[\int(K'')^2\mu_2(K)^{-1}\widehat{a}^{-1}n^{-1}\right]^{1/7}$, where $\widehat{a}$ is the estimator of $\int(h_Z^{(3)})^2$ (h_Z the density of Z) and $\mu_2(K) = \int t^2K(t)dt$. For a sample from a normal density, the pilot bandwidth g is asymptotically optimal for the purpose of estimating the density curvature by the curvature of the kernel estimator (for details see Cao *et al.*, 1994).

MODEL	Mean		Std. Dev.	
	hcv	h*	hcv	h*
W(1,1)	146.55	128.70	49.91	38.19
CW(1,1)	135.03	117.97	45.48	33.95
W(2,1)	105.86	89.47	51.91	38.91
CW(2,1)	96.92	89.27	50.09	35.47
W(3,1)	102.39	87.54	50.99	38.23
CW(3,1)	96.80	93.93	48.93	39.09
G(1,1)	113.94	101.83	50.01	38.84
CG(1,1)	107.99	91.20	47.25	30.21
G(2,1)	124.45	112.14	47.86	37.88
CG(2,1)	117.35	101.75	44.32	31.47
G(3,1)	128.52	117.05	45.89	37.47
CG(3,1)	120.48	106.84	42.83	32.25
N(1,0.5)	102.81	89.58	49.98	38.96
CN(1,0.5)	91.17	82.61	48.10	33.43

Table 2: Mean and standard deviation of the integrated absolute error (L_1-norm) of the kernel density estimator with h_{CV} and h^* bandwidths along 1.000 trials of size 100.

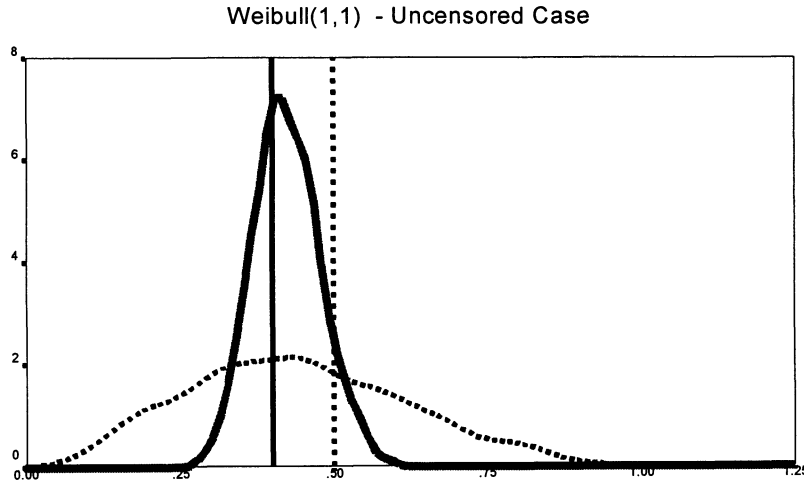

Figure 2: Kernel estimator of the densities of h_{CV} (dashed line) and h^* (solid line) using Gaussian kernel and smooth cross-validation bandwidth, based on 1.000 trials of size 100 of a Weibull(1,1) model (uncensored case). The vertical lines represent the values of h_{l2} (solid line) and h_{l1} (dotted line).

Table 2 is devoted to the L_1-norm $\int \mid \widetilde{f}_{\widehat{h}} - f \mid w$. Tables 1 and 2 show that the L_1 and L_2 norms are more concentrated around their means for the bootstrap selector than for the cross-validation bandwidth. The L_1 and

L_2 norms tend to be smaller with the bootstrap bandwidth than that with cross-validation. To show the accuracy of both bandwidth selectors figures 2 (uncensored case) and 3 (censored) present the kernel estimators of the densities of the two selectors h_{CV} and h^*, using Gaussian kernel and smooth cross-validation bandwidth, based on a sample of 1.000 different values from the models W(1,1) and CW(1,1), as well as some approximations of the $h_{L_2,w}$ and $h_{L_1,w}$ bandwidths. These bandwidths were computed by minimizing the Monte Carlo approximation, based on 1.000 trials of the $MISE_w$, $E\int(\widetilde{f}_h - f)^2 w$, and the $MIAE_w$, $E\int \mid \widetilde{f}_h - f \mid w$, over a grid of h values, ranging from 0.1 to 1.1 with a step of 0.05. We conclude that the performance of h^* is far superior to that of h_{CV}. This is what we expected, because as noted in Cao *et al.* (1994) at least in the uncensored case we have

$$\frac{h_{CV} - h_f}{h_f} = O_P(n^{-1/10}) \qquad \frac{h^* - h_f}{h_f} = O_P(n^{-5/14})$$

where h_f denotes the optimum bandwidth which minimizes $MISE$.

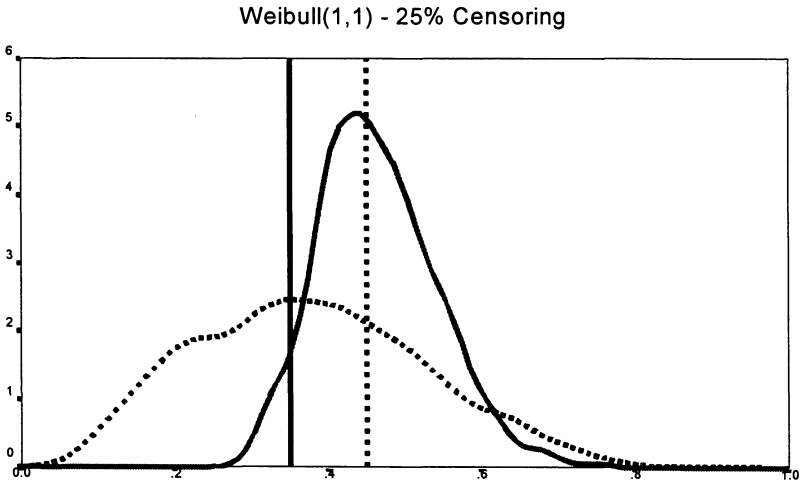

Figure 3: Kernel estimator of the densities of h_{CV} (dashed line) and h^* (solid line) using Gaussian kernel and smooth cross-validation bandwidth, based on 1.000 trials of size 100 of a Weibull(1,1) model (25% censoring). The vertical lines represent the values of h_{l2} (solid line) and h_{l1} (dotted line).

To illustrate the computational cost of the WARPing approximation when used in the algorithms for finding h^* and h_{CV}, the CPU times (relative to the minimum of all of them) are summarized in Table 3. While CPU

time increases slowly with the sample size, n, a sharp increase occurs when the rounding error, δ, gets small. We suggest the choice δ=0.01 in practice. This seems to give a good approximation of the kernel estimator (see Figure 1) at a reasonable computational cost. (See Fan and Marron, 1994) for a comparative study of the CPU time of the WARPing approximation and the kernel estimator).

Sample	δ=0.05		δ=0.01		δ=0.005	
Size	hCV	h*	hCV	h*	hCV	h*
n=50	1.33	1.00	23.78	14.08	97.82	52.69
n=100	5.46	1.17	91.88	21.12	373.83	81.70
n=500	5.65	1.35	107.68	20.54	433.19	107.95

Table 3: Relative CPU times of the computations of LSCV and bootstrap bandwidths for samples of sizes n=50, 100 and 500 and rounding errors δ=0.05, 0.01 and 0.005 in the WARPing approximation.

4.2 Example: PCB-Liver data

Between January, 1974, and May, 1984, the Mayo Clinic conducted a double-blinded randomized trial in Primary Biliary Cirrhosis (PCB) of the liver. A total of n=312 patients agreed to participate in this clinical trial. The data were analyzed in 1986 for presentation in clinical literature (see Fleming *et al.*, 1991). By July, 1986, 125 of the 312 patients had died, resulting in a high proportion of censoring data (60%).

	Sample Size	Deaths	Censored (%)	OR (95% CI)	Median	Mean
GROUP I	163	32	131 (80.37%)	1 Reference	*	10.41
GROUP II	63	31	32 (50.79%)	3.5 (2.2,5.8)	7.66	7.22
GROUP III	50	31	19 (38.00%)	6.45 (3.9,10.7)	4.63	5.47
GROUP IV	36	31	5 (13.89%)	18.9 (11.0,32.5)	2.35	2.60
TOTAL	312	125	187 (59.44%)	* More than 50% people aliving at the closing date		

Table 4: Sample size, number of deaths, percentage of censoring, Odds-Ratio and estimated median and mean for the four groups of patients considered in the PCB-Liver data example according to their prognosis bilirubin levels (<1.45, 1.45-3.25, 3.25-6.75 and >6.75).

One of the most important risk factor for the survivorship of PCB is the serum bilirubin level (Fleming *et al.*, 1991). As in the referred work, we shall consider four groups of patients, according to their prognosis bilirubin levels: Group I (bilirubin<1.45), Group II (1.45-3.25), Group III (3.25-6.75) and Group IV (>6.75). Descriptive statistics appear in Table 4, revealing that the survival time decreases as albumin level increases. Whereas

Group I presents an 80.3% of censored data, Group II has an amount around 50.6%, Group III 38% and Group IV, only 14%. The four groups were tested to follow the Koziol-Green model using the test proposed by Henze (1993). The approximated p-values were 0.1627, 0.1552, 0.1765 and 0.46764, respectively, failing to reject the proportional censoring model for each group. The estimator (3) of each density function can therefore be safely used. In all cases, a gaussian kernel was chosen. For each group, the weighting function considered here was $w(u) = 1_{[a,b]}(u)$ where a and b are, respectively, the 5% and 95% percentiles of the corresponding observed survival time. The selected bootstrap bandwidth minimizing their respective function $\overline{AMISE^*}(M)$ (with rounding error δ=0.1) were h_I^*=3.9, h_{II}^*=4.01, h_{III}^*=3.3 and h_{IV}^*=1.73. The density functions estimates are plotted together in Figure 4. Looking at this figure, one observes that the density estimates reflect the behaviour of the survival time, revealing a great amount of probability around the median values, in agreement with results presented above.

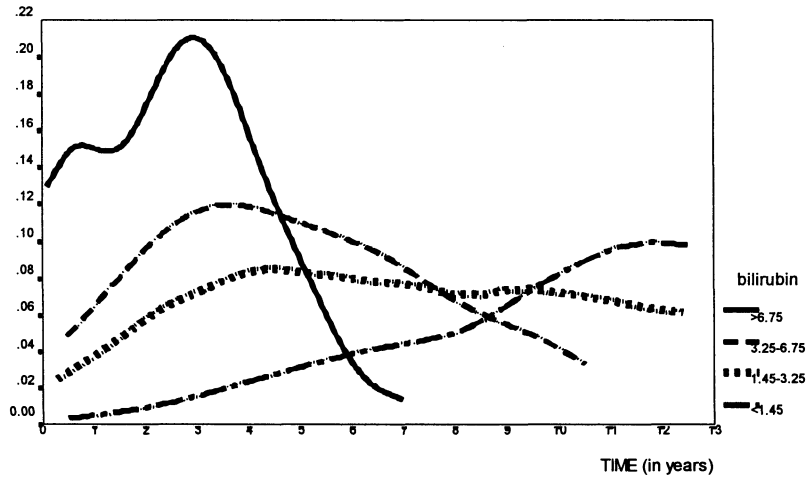

Figure 4: Kernel density estimators for the groups in the PCB-Liver data example: bilirubin > 6.75 (solid line), 3.25-6.75 (dashed line), 1.45-3.25 (dotted line) and <1.45 (dashed-dotted line). The figure supports the numerical results in Table 4.

5 Conclusions

We proposed a smoothed bootstrap selector of the smoothing parameter in kernel density estimation when the Koziol-Green model of proportional

censorship holds. We presented a mathematical analysis supported by simulation. It turned out that our proposal behaves convincingly better than the cross-validation selector. We gave fast implementation using WARPing methods. We showed the practical interest of the introduced techniques through the analysis of real medical data sets.

Acknowledgements

The work of the second and the third author was partially supported by the project XUGA20701B96 of the Xunta de Galicia and by the DGES grant PB95-0826.

References

[1] Abdushukurov, A. A. (1984). On some estimates of the distribution function under random censorship. Conference of Young Scientists, Math. Inst. Acad. Sci. Uzbek SSR, Tashkent. VINITI n^o 8756-V (in Russian).

[2] Cao, R., Cuevas, A. and González-Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation. *Comput. Statist. Data Anal.* **17** 153-176.

[3] Cheng, P. E. and Lin, G.D. (1984). Maximum likelihood estimation of a survival function under the Koziol-Green proportional hazard model. Technical Report, Institute of Statistics, Academia Sinica.

[4] Csörgo, S. and Mielniczuk, J. (1988). Density estimation in the simple proportional hazards model. *Statistics and Probability Letters* **6** 419-426.

[5] Fan, J. and Marron, J. S. (1994). Fast implementations of nonparametric curve estimators. *Journal of Computational and Graphical Statistics* **3** 35-56.

[6] Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes & Survival Analysis.* New York: Wiley-Interscience.

[7] Ghorai, J. K. and Pattanaik, L. M. (1993). Asymptotically optimal bandwidth selection of the kernel density estimator under the proportional hazards model. *Commun. Statist. Theory and Methods* **22** 1383-1401.

[8] González-Manteiga, W., Cao, R. and Marron, J. S. (1996). Bootstrap selection of the smoothing parameter in nonparametric hazard rate estimation. *J.Am. Statist.Assoc.* **91** 1130-1140.

[9] Härdle, W. (1991). *Smoothing Techniques with Implementation in S.* Springer-Verlag.

[10] Henze, N. (1993). A quick omnibus test for the proportional hazards

model of random censorship. *Statistics* **24** 253-263.

[11] Herbst, T. (1992). Test of fit with the Koziol-Green model for random censorship. *Statistics & Decisions* **10** 163-171.

[12] Koziol, J. A. and Green, S. B. (1976). A Cramér-von Mises statistic for randomly censored data. *Biometrika* **63** 465-474.

[13] Sethuraman, J. (1965). On a characterization of the three limiting types of the extreme. *Sankhyā* A **27**, 357-364.