Compstat'2010, Paris, August 22–27

Censored Survival Data: Simulation and Kernel Estimates

Jiří Zelinka

Department of Mathematics and Statistics

Faculty of Science, Masaryk University

Brno, Czech Republic

Supported by MŠMT LC06024

Introduction

- Previous research (Horová, Pospíšil & Zelinka (2008) and Horová, Pospíšil & Zelinka (2009)): combination of kernel smoothing and dynamic model in survival analysis.
- Verification of developed method: simulations $\rightarrow$ **problem**
- The subject of this paper is to solve this problem.

Survival and hazard functions

$T \geq 0$	**survival time**
F	**cumulative distribution function** (c.d.f.) of T
$\bar{F} = 1 - F$	**survival function**
$\lambda = \lambda(x)$	**hazard function**

Hazard function: – intensity of survival probability:

$$\lambda(x) = -\frac{\bar{F}'(x)}{\bar{F}(x)} = -\log'(\bar{F}(x)) = \frac{f(x)}{\bar{F}(x)} \tag{1}$$

if the density f exists. From (1) we have

$$\bar{F}(x) = e^{-\int_0^x \lambda(t)dt}. \tag{2}$$

Random censorship model

$T_1, T_2, \ldots, T_n$ i.i.d. **lifetimes** with c.d.f. F

$C_1, \ldots, C_n$ i.i.d. **censoring times** with c.d.f. G

Censoring times are independent of the lifetimes.

In the **random censorship model** we observe pairs

$$(X_i, \delta_i),\ i = 1, \ldots, n,\ \text{where } X_i = \min(T_i, C_i)$$

$\delta_i = 1_{\{X_i = T_i\}}$ indicates whether the observations is censored or not.

$\{X_i\}$ are i.i.d. with survival function $\bar{L}$: $\bar{L}(x) = \bar{F}(x)\bar{G}(x)$.

Kernel estimates of the hazard function

Let $[0, \tau]$, $\tau > 0$, be an interval for which $L(\tau) < 1$ and $\lambda \in C^2[0, \tau]$ and let K be a continuous and symmetric function on R called a kernel satisfying conditions:

1. $\operatorname{supp} K = [-1, 1]$
2. $K \in \operatorname{Lip}[-1, 1]$
3. $\int_{-1}^{1} x^k K(x) \mathrm{d}x = \begin{cases} 1, & k = 0 \\ 0, & k = 1 \\ \beta_2 \neq 0, & k = 2. \end{cases}$

The well-known kernels:

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

$K(x) = \frac{15}{16}(1 - x^2)^2 1_{[-1,1]}$ quartic kernel

The kernel estimate of the hazard function is given as

$$\hat{\lambda}_{h,K}(x) = \frac{1}{h} \sum_{i=1}^{n} K\left(\frac{x - X_{(i)}}{h}\right) \frac{\delta_{(i)}}{n - i + 1}\,. \tag{3}$$

The parameter h is called bandwidth or smoothing parameter.
Let us denote

$$V(K) = \int_{-1}^{1} K^2(x)\mathrm{d}x, \quad \beta_2 = \int_{-1}^{1} x^2 K(x)\mathrm{d}x,$$

$$\Lambda = \int_0^T \frac{\lambda(x)}{\bar{L}(x)}\mathrm{d}x, \quad D_2 = \int_0^T \left(\lambda^{(2)}(x)\right)^2 \mathrm{d}x.$$

The global quality of the estimate – Mean Integrated Square Error:

$$MISE\left(\hat{\lambda}_{h,K}\right) = \int_0^T MSE\left(\hat{\lambda}_{h,K}(x)\right)\mathrm{d}x = \int_0^T \mathsf{E}\left(\hat{\lambda}_{h,K}(x) - \lambda(x)\right)^2 \mathrm{d}x,$$

The leading term $\overline{MISE}(\hat{\lambda}_{h,K})$ of $MISE(\hat{\lambda}_{h,K})$ takes the form

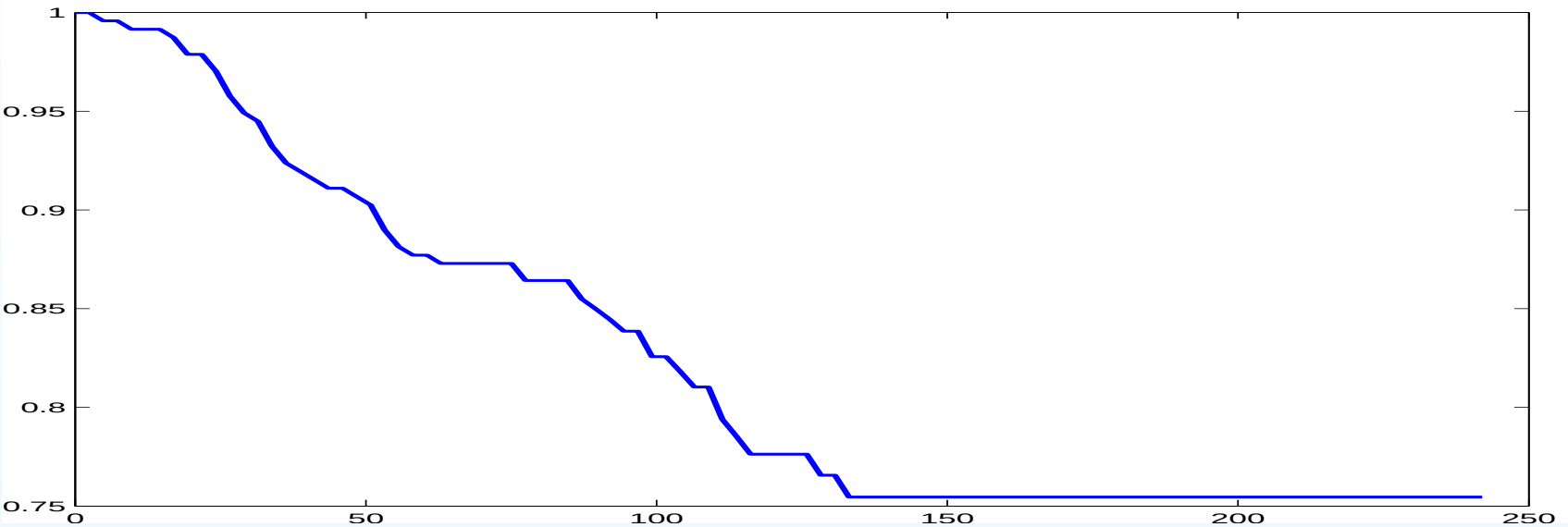
$$\overline{MISE}\left(\hat{\lambda}_{h,K}\right) = \frac{1}{4}h^4\beta_2^2 D_2 + \frac{V(K)\Lambda}{nh}$$

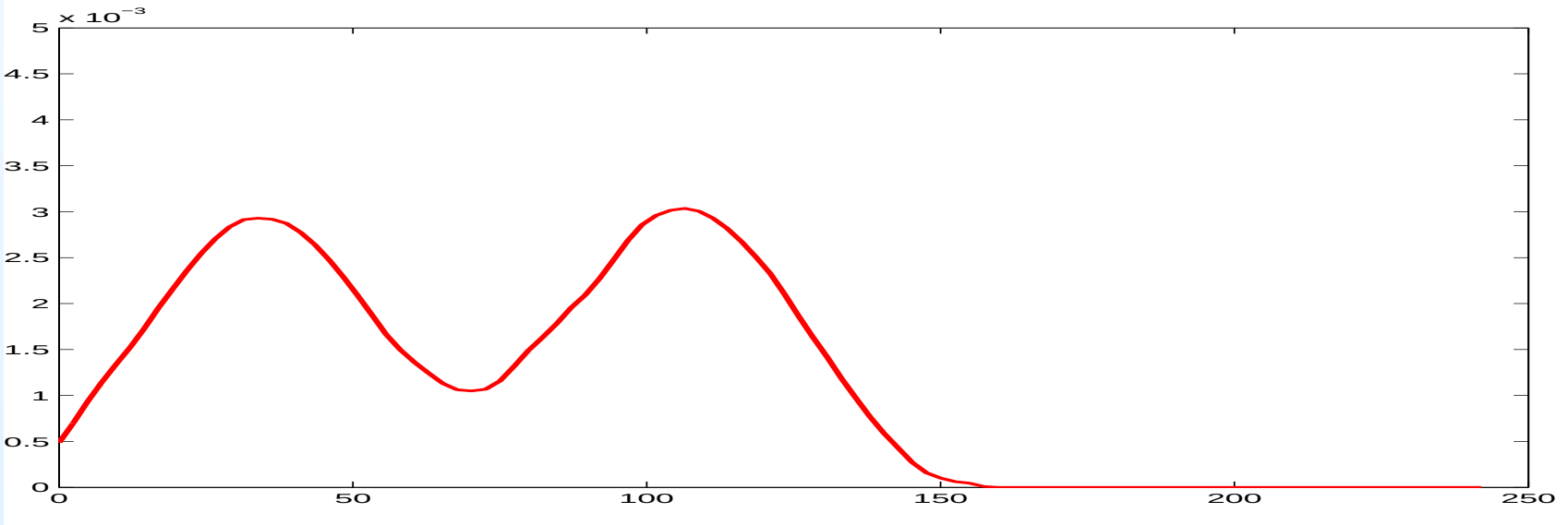
The asymptotically optimal bandwidth minimizing $\overline{MISE}(\hat{\lambda}_{h,K})$ with respect to h is given by the formula

$$h_{opt} = n^{-1/5}\left(\frac{\Lambda V(K)}{\beta_2^2 D_2}\right)^{1/5} \tag{4}$$

The estimate of h_{opt} will be denoted with $\hat{h}_{opt}$. See Horová & Zelinka (2006) for method of evaluating the appropriate estimate $\hat{h}_{opt}$.

Kernel estimate of the hazard function – example



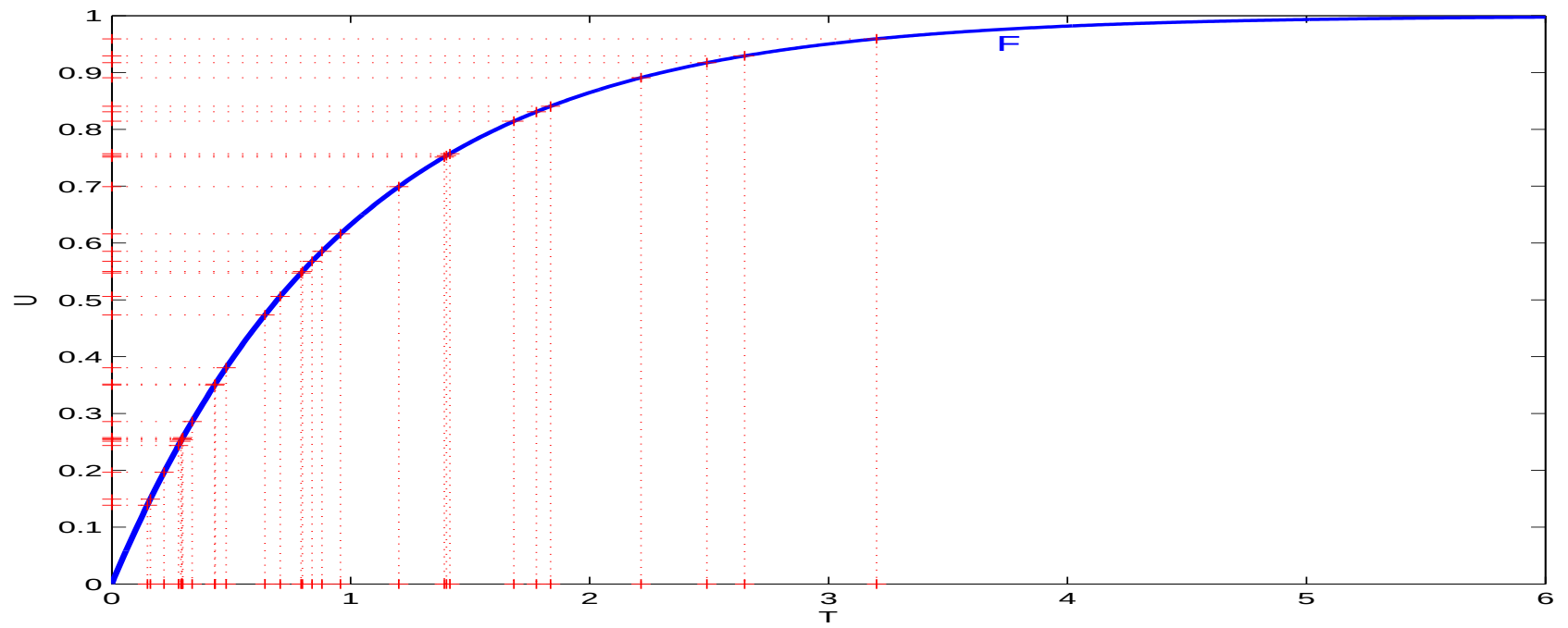


Simulation of lifetimes

For given hazard function λ we have (see (2))

$$F(x) = 1 - e^{-\int_0^x \lambda(t)dt}$$

The lifetimes $T_1,\ldots,T_n$ can be evaluated numerically by re-sampling random variables $U_1,\ldots,U_n$ uniformly distributed on interval $[0,1]$.



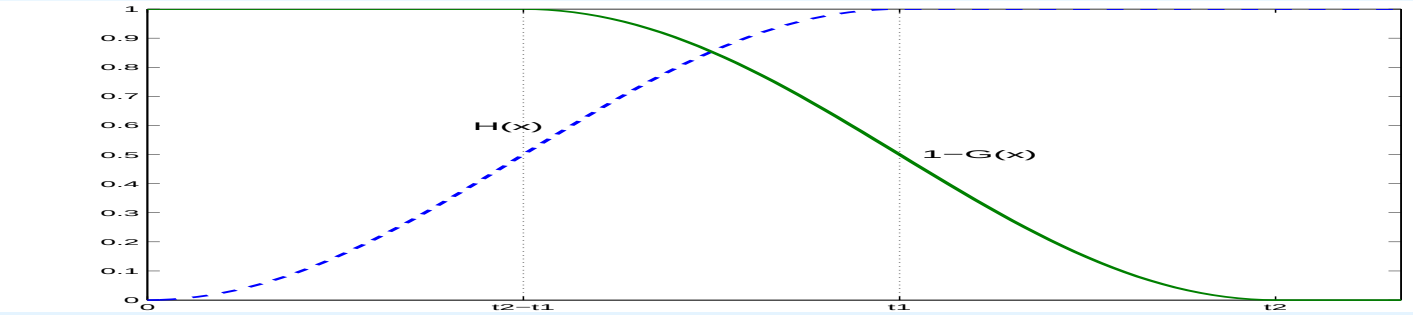
Simulation of censoring times

Real situation:

Let's have a clinical study dealing with some disease. The research begins in time t_0 (we can suppose $t_0 = 0$). Patients come to the study randomly in interval $[t_0, t_1]$, the begin of treatment is given by random variable B with cumulative distribution function H. The coming of patients is broken in time t_1, but the study may continue to some time $t_2 \geq t_1$ when it is finished.

The censorship time is $C = t_2 - B$ For the survival function $\bar{G}$ we have

$$\bar{G}(x) = H(t_2 - x)\,.$$



Cumulative distribution function for coming of patients (H)

and survival function of censoring times ($\bar{G} = 1 - G$).

Let us recall

$$h_{opt}^5 = \frac{V(K)\Lambda}{n\beta_2^2 D_2}$$

for

$$V(K) = \int_{-1}^{1} K^2(x)\mathrm{d}x, \quad \beta_2 = \int_{-1}^{1} x^2 K(x)\mathrm{d}x,$$

$$\Lambda = \int_0^\tau \frac{\lambda(x)}{\bar{F}(x)\bar{G}(x)}\mathrm{d}x, \quad D_2 = \int_0^\tau \left(\lambda^{(2)}(x)\right)^2 \mathrm{d}x.$$

Choice of τ: naturally $\tau = t_2$, $\Rightarrow$ problem with counting Λ as $\bar{G}(t_2) = 0$.
Solution: for $\bar{G}$ let us take such λ that

$$\bar{F}(t_2) > 0, \quad \frac{\lambda(x)}{\bar{G}(x)} = O(1), \text{ for } x \to t_2 \,.$$

As a result of this property we have $\lambda(t_2) = 0$ and for $\lambda \in C^2[0,T]$ also $\lambda'(t_2) = 0$ as λ is non-negative.

In all simulations let the begins of treatment B be uniformly distributed on $[0, t_1]$. Due to this fact the cumulative distribution function C is uniformly distributed on $[t_2 - t_1, t_2]$.

Simulation 1

λ: unimodal hazard function on $[0, t_2]$:

$$\lambda(x) = x(2-x)^2$$

$$F(x) = 1 - e^{\frac{x^2}{12}(3x^2 - 16x + 24)}.$$

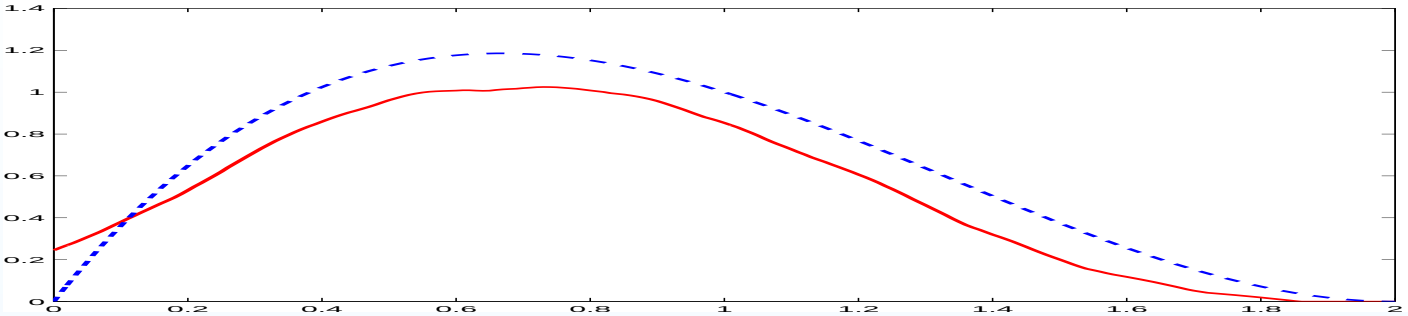
Let K be the Epanechnikov kernel and $n = 100$

Case A: $t_1 = 1$, $t_2 = 2$, $h_{opt} = 0.4437$

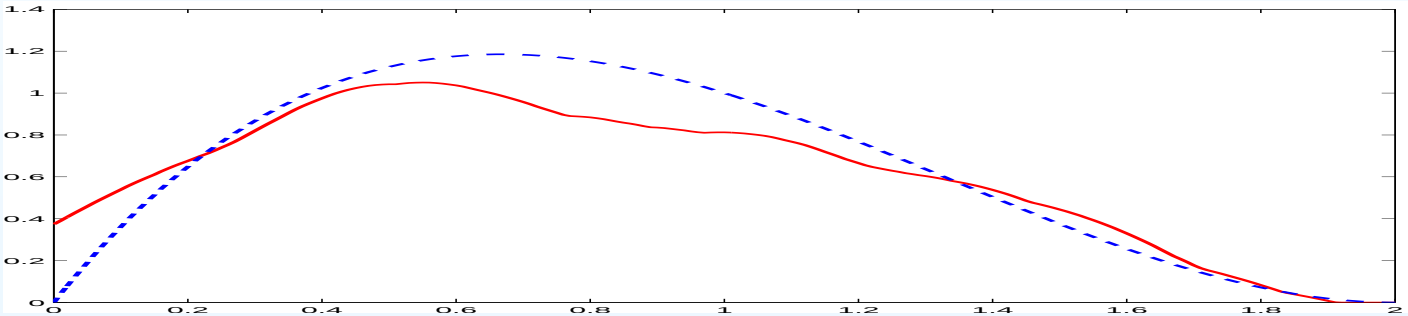
Case B: $t_1 = 1.5$, $t_2 = 2$, $h_{opt} = 0.4721$

Case C: $t_1 = 2$, $t_2 = 2$, $h_{opt} = 0.4993$

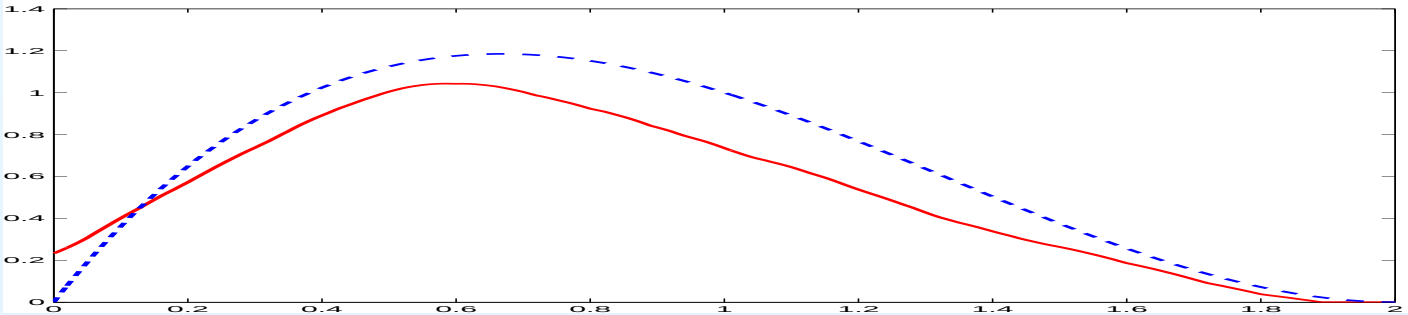
Estimate of λ for optimal bandwidth



Case A: λ – dashed line, estimate – solid line

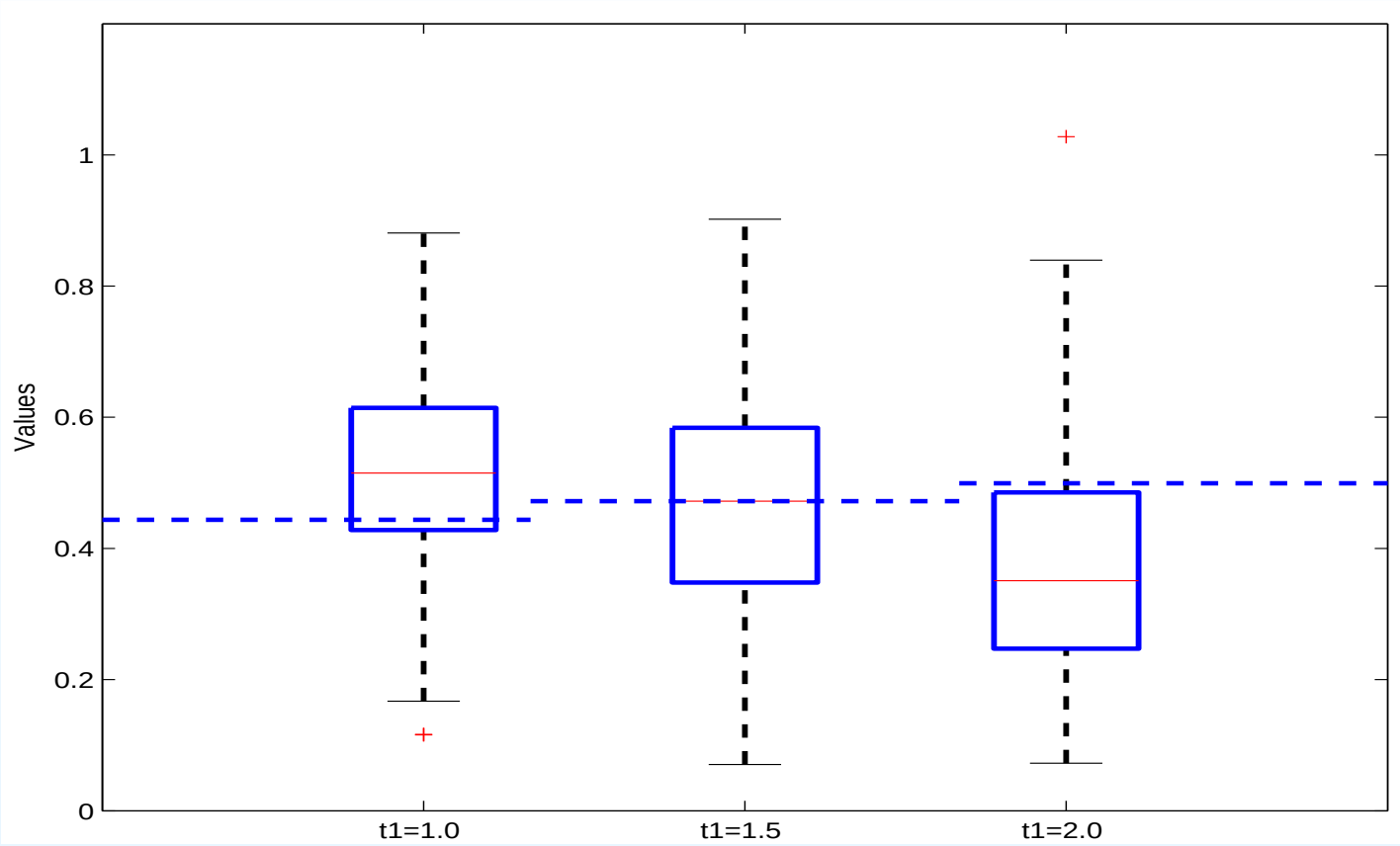


Case B: λ – dashed line, estimate – solid line



Case C: λ – dashed line, estimate – solid line

Estimate of $hopt$ for 200 repetitions



Dashed lines: optimal bandwidths

Simulation 2

λ: unimodal hazard function on $[0, t_2]$:

$$\lambda(x) = \frac{1}{100}\left(1 - \cos\frac{2 * \pi}{t_2}x\right)$$

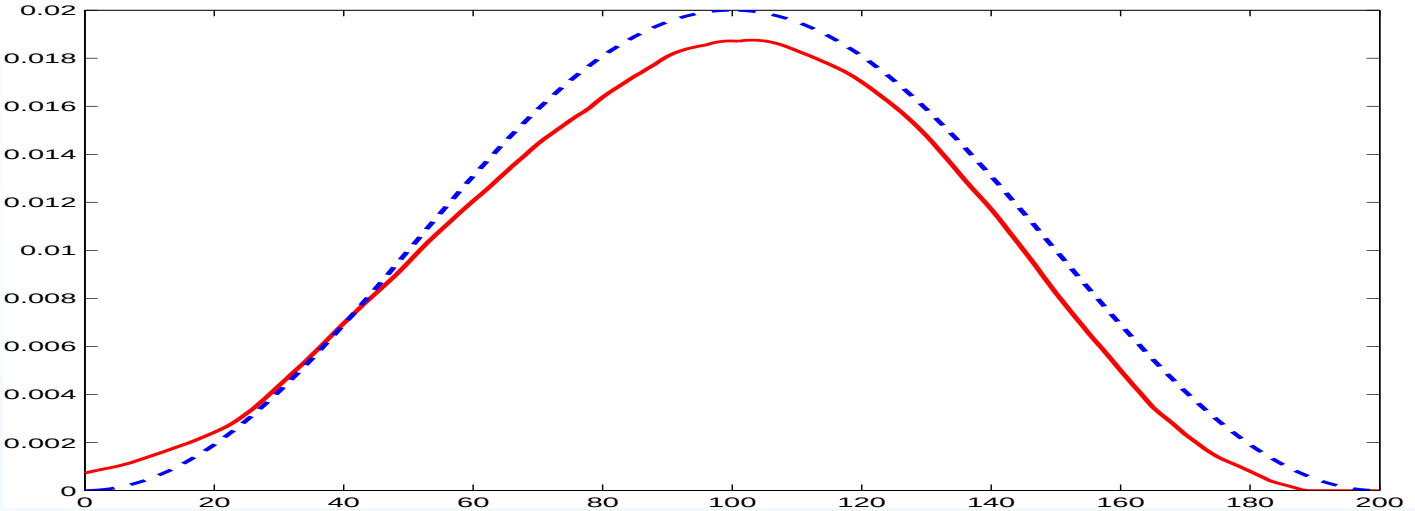
$$F(x) = 1 - e^{\frac{1}{100}(\frac{t_2}{2*pi}\sin\frac{2*pi}{t_2}x - x)}$$

Let K be the Epanechnikov kernel and $n = 100$

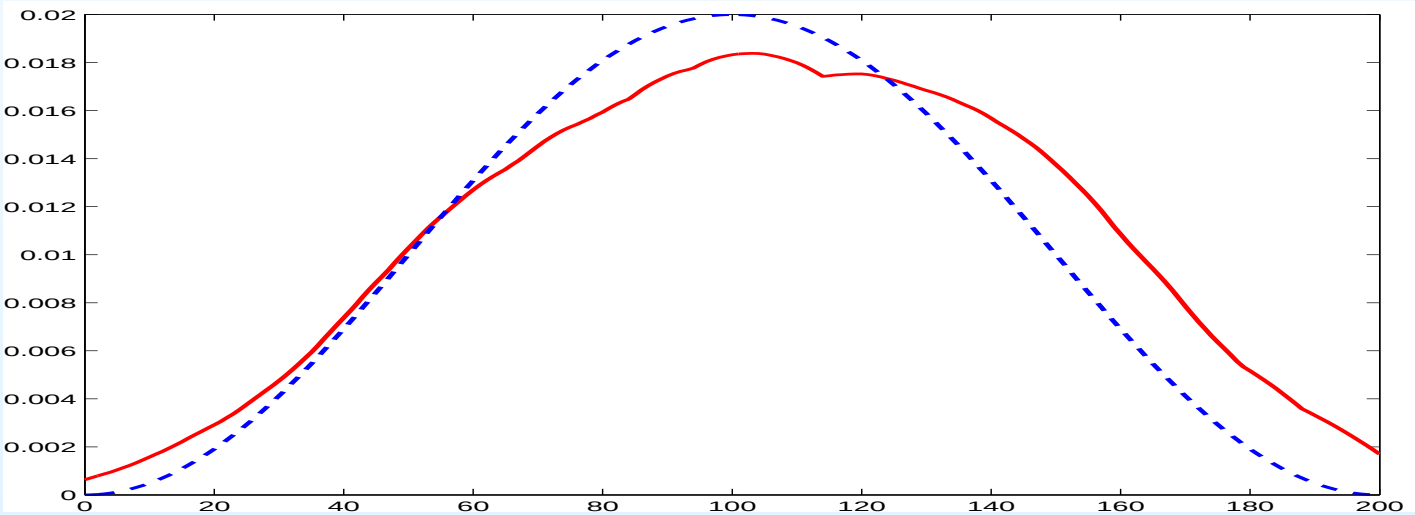
Case A: $t_1 = 100$, $t_2 = 200$, $h_{opt} = 43.703$

Case B: $t_1 = 150$, $t_2 = 200$, $h_{opt} = 47.122$

Estimate of λ for optimal bandwidth

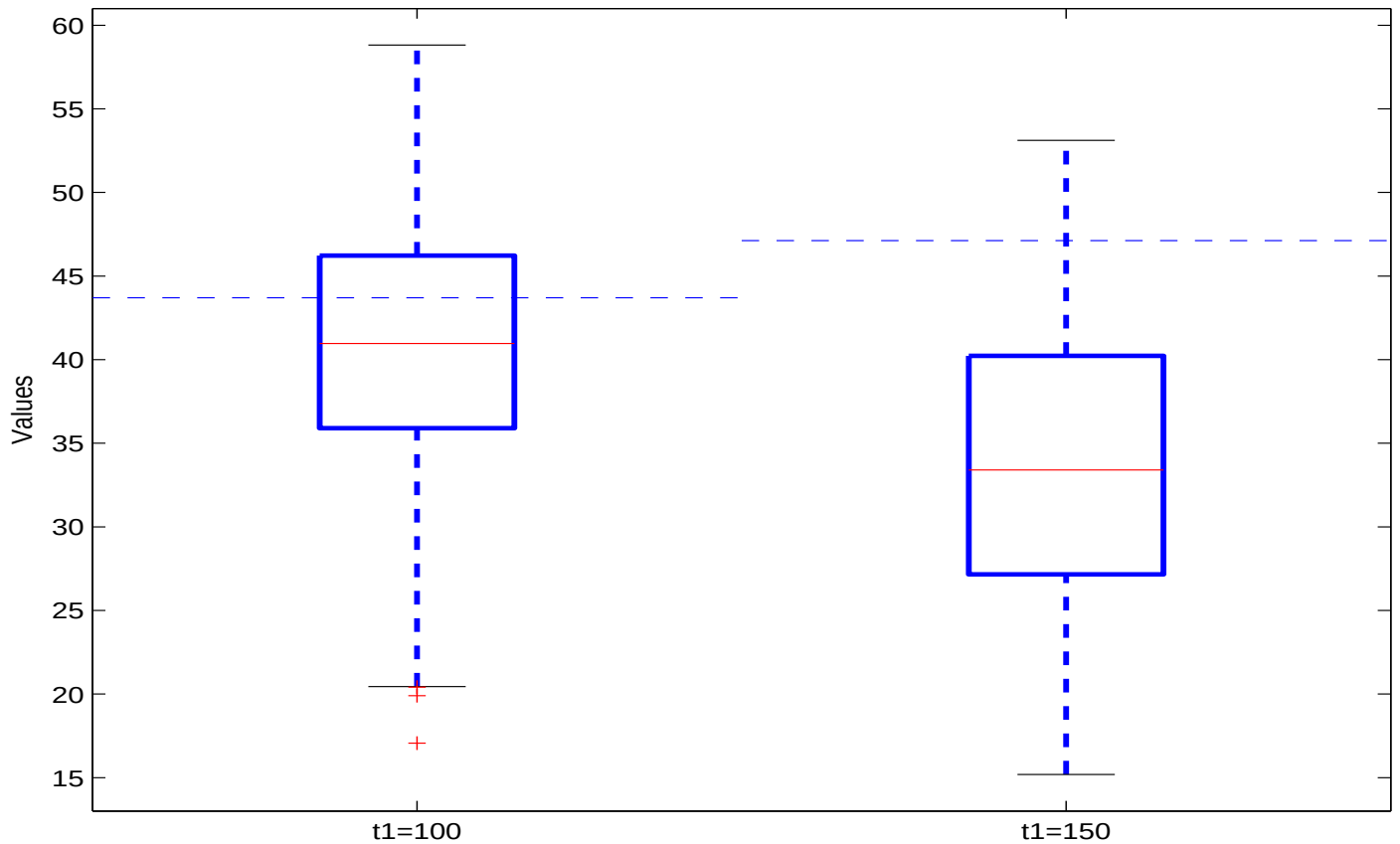


Case A: λ – dashed line, estimate – solid line



Case B: λ – dashed line, estimate – solid line

Estimate of $hopt$ for 200 repetitions



Dashed lines: optimal bandwidths

Simulation 3

λ: bimodal hazard function on $[0, t_2]$:

$$\lambda(x) = \frac{1}{100}\left(1 - \cos\frac{4 * \pi}{t_2}x\right)$$

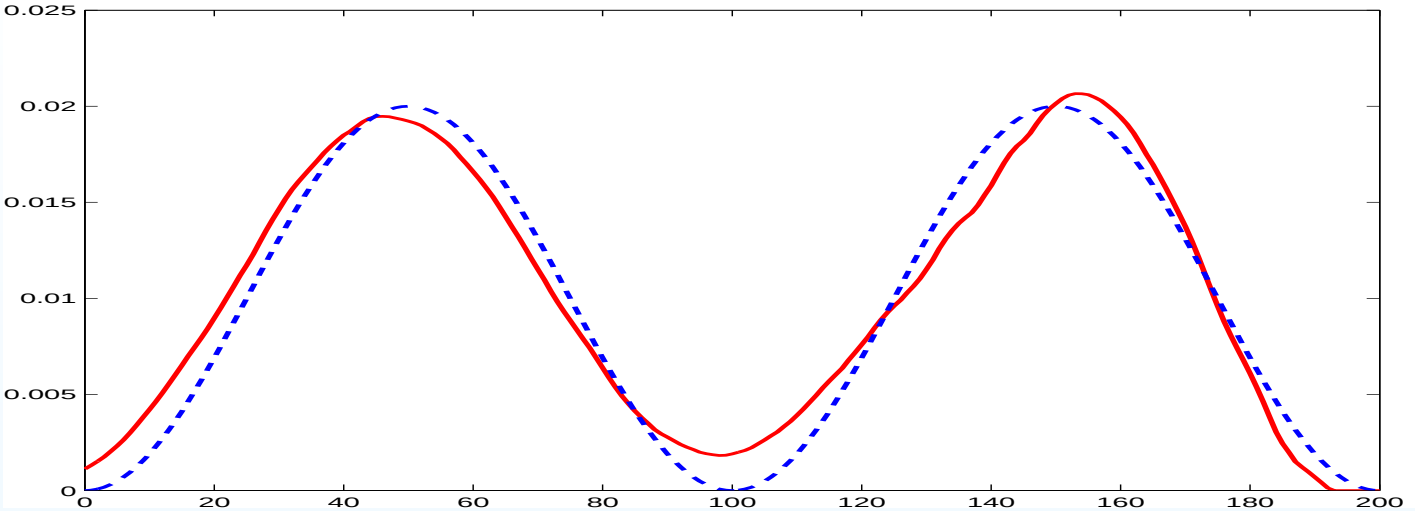
$$F(x) = 1 - e^{\frac{1}{100}(\frac{t_2}{4*pi}\sin\frac{4*pi}{t_2}x - x)}$$

Let K be the Epanechnikov kernel and $n = 200$

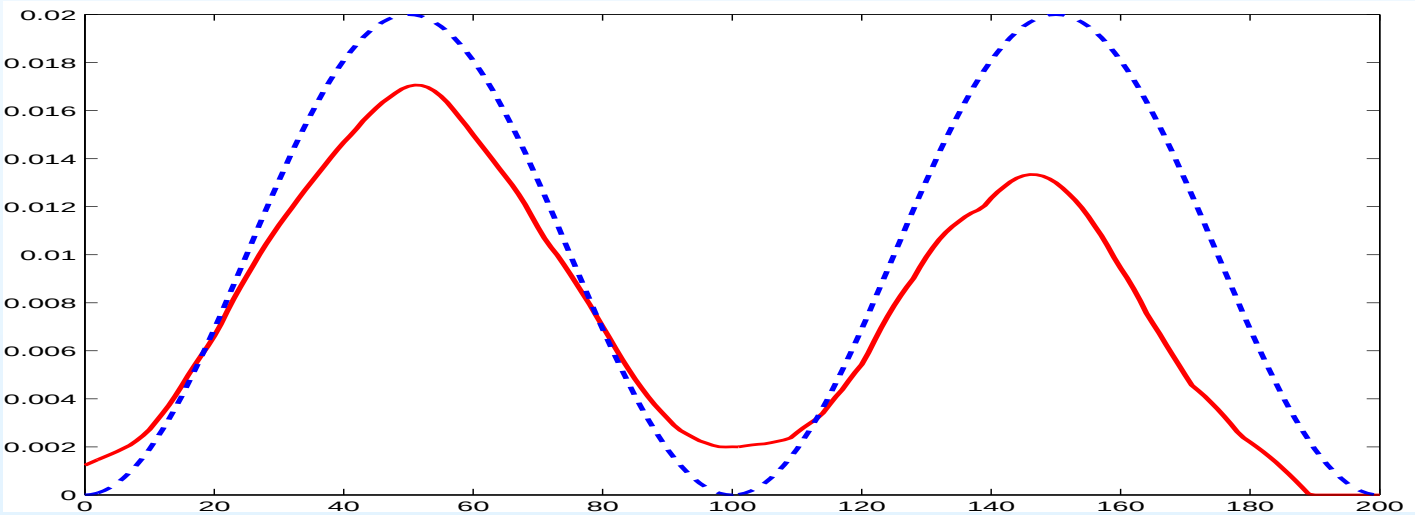
Case A: $t_1 = 100$, $t_2 = 200$, $h_{opt} = 23.443$

Case B: $t_1 = 150$, $t_2 = 200$, $h_{opt} = 25.255$

Estimate of λ for optimal bandwidth

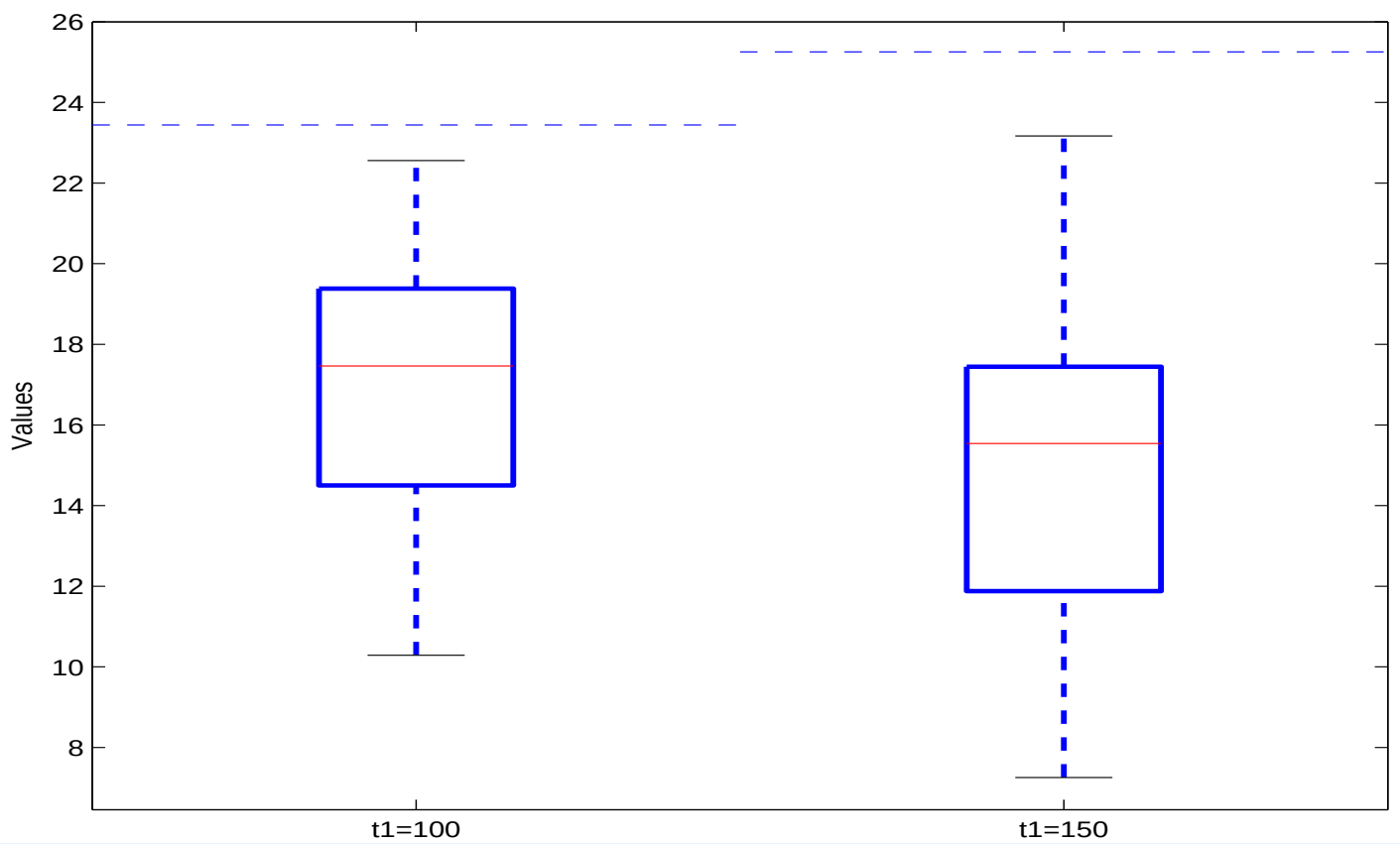


Case A: λ – dashed line, estimate – solid line



Case B: λ – dashed line, estimate – solid line

Estimate of $hopt$ for 200 repetitions



Dashed lines: optimal bandwidths

Conclusion

- The simulations indicate that the proposed method of generating random censored data for given cumulative distribution function C and hazard function λ can be well applied for testing the algorithms of survival analysis.
- At the same time the simulations show that the method of bandwidth choice proposed in Horová & Zelinka (2006) gives worse results for the greater frequency of censored data, but the estimates of optimal bandwidth are still well usable.

References

Collett D.: *Modelling Survival Data in Medical Research*. Chapman & Hall/CRC: Boca Raton-London-New York-Washington, D.C., 2003.

Horová I., Zelinka J., Budíková M.: Estimates of Hazard Functions for Carcinoma Data Sets. *Environmetrics*, **17**, 239–255, 2006.

Horová I., Zelinka J.: (2006) Kernel Estimates of Hazard Functions for Biomedical Data Sets. In *Applied Biostatistics: Case studies and Interdisciplinary Methods*, Springer, 2006.

Horová I., Pospíšil Z., Zelinka J.: Semiparametric Estimation of Hazard Function for Cancer Patients, *Sankhya*, **69**, 494–513, 2008.

Horová I., Pospíšil Z., Zelinka J.: Hazard function for cancer patients and cancer cell dynamics, *Journal of Theoretical Biology*, textbf258, 437–443, 2009.

References

Müller H.G., Wang J.L.: Nonparametric Analysis of Changes in Hazard Rates for Censored Survival Data: An alternative Change-Point Models. *Biometrika*, **77**(2), 305–314, 1990.

Ramlau-Hansen H.: Counting Processes Intensities by Means of Kernel Functions. *The Annals of Statistics*, **11**(2), 453–466, 1983.

Tanner M.A., Wong W.H.: The Estimation of the Hazard Function from Randomly Censored Data by the Kernel Method. *The Annals of Statistics*, **11**(3), 989–993, 1983.

Uzunogullari U., Wang J.L.: A comparision of Hazard Rate Estimators for Left Truncated and Right Censored Data. *Biometrika*, **79**(2), 297–310, 1992.

Wand, I.P., Jones, I.C.: *Kernel smoothing*. Chapman & Hall, London, 1995.