International Symposium on Recent Advances in Statistics, in Honor of Ingrid Van Keilegom

An EM algorithm based estimator for the latency in mixture cure models

Mª Amalia Jácome[1], Ana López-Cheda[1] and Yingwey P. Peng[2]
[1] MODES group, University of A Coruña (Spain)
[2] Cancer Research Institute, Queen's University (Canada)

June 15, 2022

UNIVERSIDADE DA CORUÑA

gmodes DPTO. DE MATEMÁTICAS

Ana López Cheda
ana.lopez.cheda@udc.es

Y. P. Peng
pengp@queensu.ca

María Amalia Jácome
majacome@udc.es

First contact to Survival Analysis (LLN, 2004)

SEIO - FBBVA award (2021)

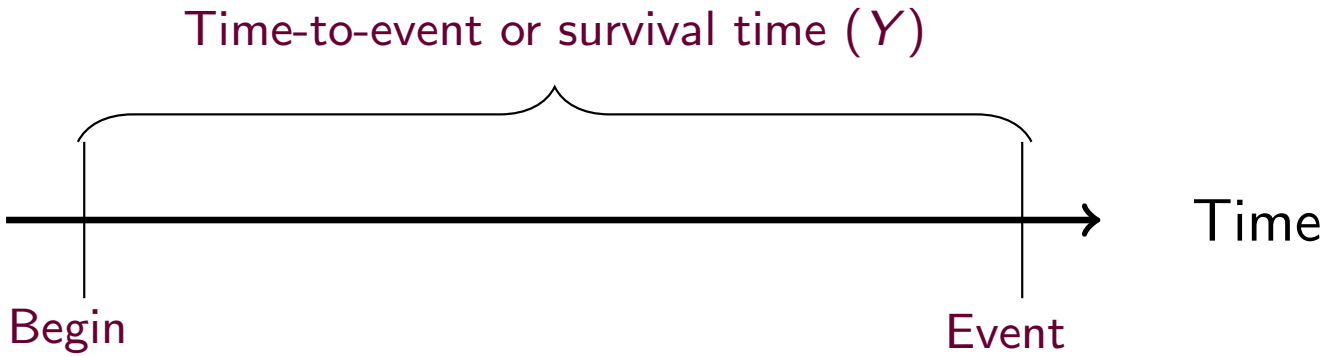
Best methodological contribution in the field of Statistics

[1] López-Cheda, A., Cao, R., Jácome, M. A. and Van Keilegom, I. (2017a). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Computational Statistics and Data Analysis*, 105, 144–165.

Outline

Introduction

Survival analysis



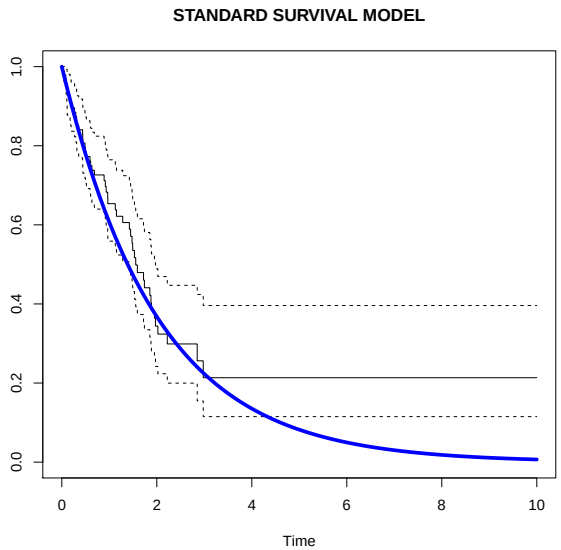
- Interest on survival function $\boxed{S(t) = 1 - F(t) = P(Y > t)}$
- Common challenge in practice is that an event is not always observed (censored observations).

Only a lower bound of the survival time is known (right censoring)

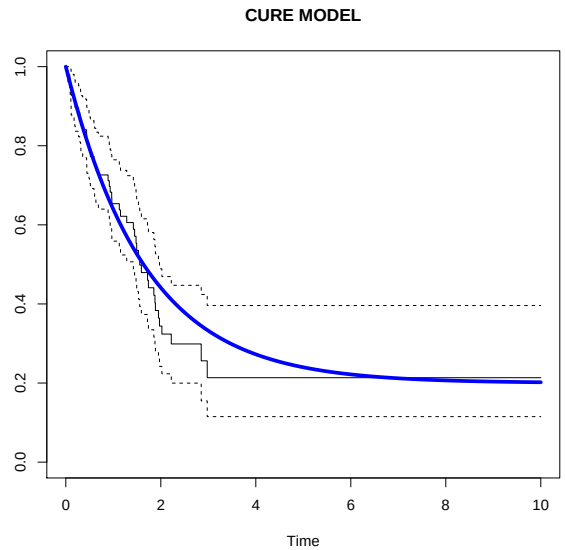
- Observed time $T = \min(Y, C)$.
- Event indicator $\delta = \mathbf{1}(Y \leq C)$.

Standard survival models vs. cure models

- **Standard survival models** suppose that all subjects are assumed to eventually experience the event ($Y < \infty$).
- **Cure models** assume that there is a proportion of subjects who will never experience the event and thus the survival curve reaches a plateau $P(Y = \infty) > 0$.

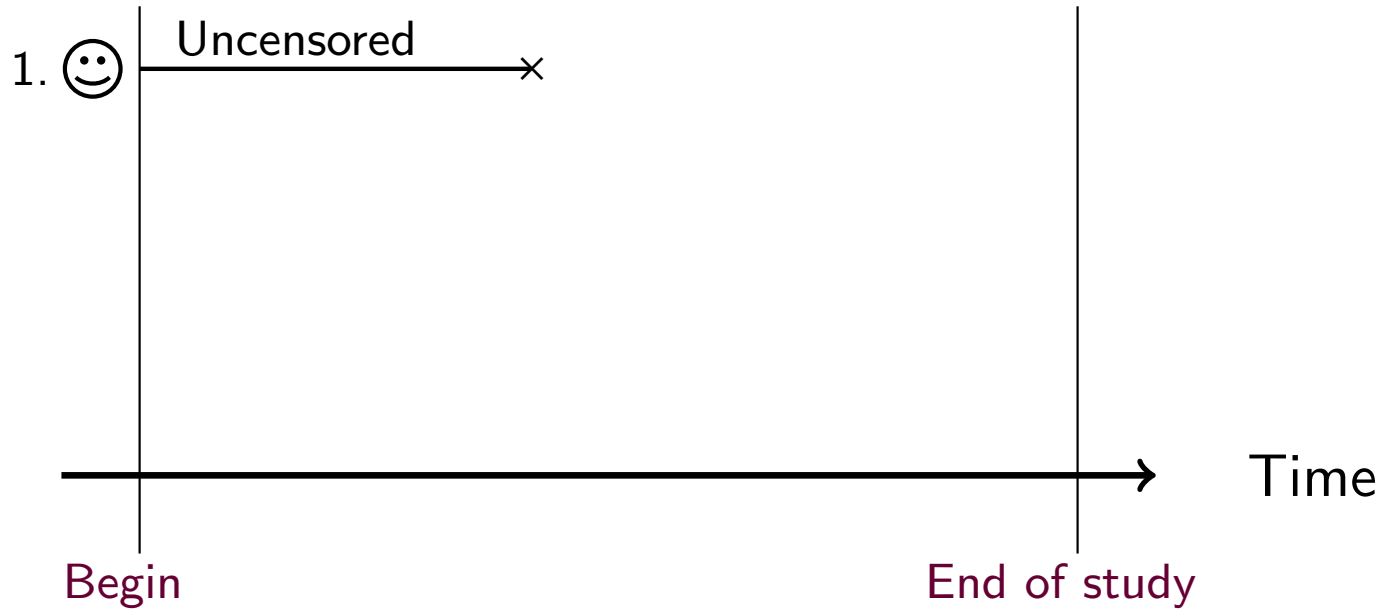


$$P(Y = \infty) = \lim_{t \to \infty} S(t) = 0.$$

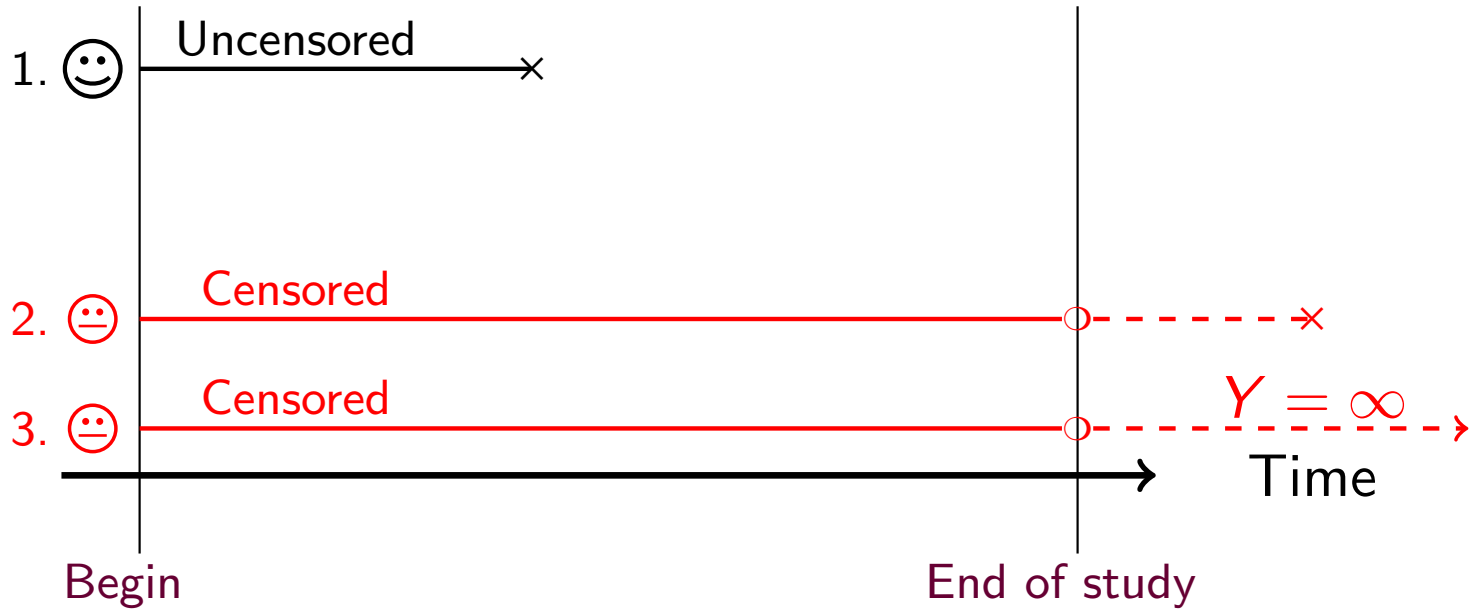


$$P(Y = \infty) = \lim_{t \to \infty} S(t) > 0.$$

Standard cure models



Standard cure models



Notation

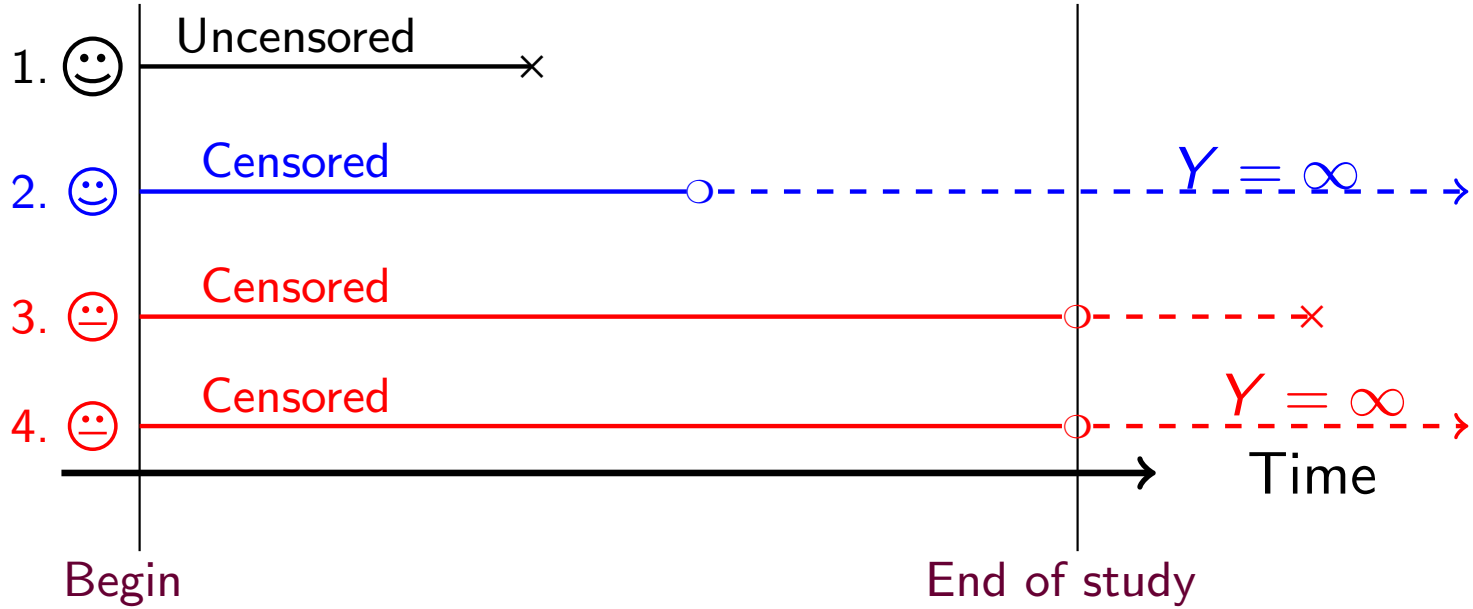
- Let Y be the time to event of interest, possibly censored at the censoring time C. With a cure fraction, the observations are $(\mathbf{X}, T, \delta)$:
  - $\mathbf{X}$ is a covariate vector.
  - $T = \min(Y, C)$ is the observed time.
  - $\delta = \mathbf{1}(Y \leq C)$ is the event indicator.
- Let $U = \mathbf{1}(Y < \infty)$ be the uncure indicator (latent variable).

Uncensored	$(\mathbf{X_i}, T_i = Y_i, \delta_i = 1)$	➡	$U_i = 1$
Censored	$(\mathbf{X_i}, T_i = C_i, \delta_i = 0)$	➡	$U_i =??$

Censoring hinders from classifying the censored observations to be classified as cured or uncured (U is not available for censored observations!).

(Cure status is partially known)

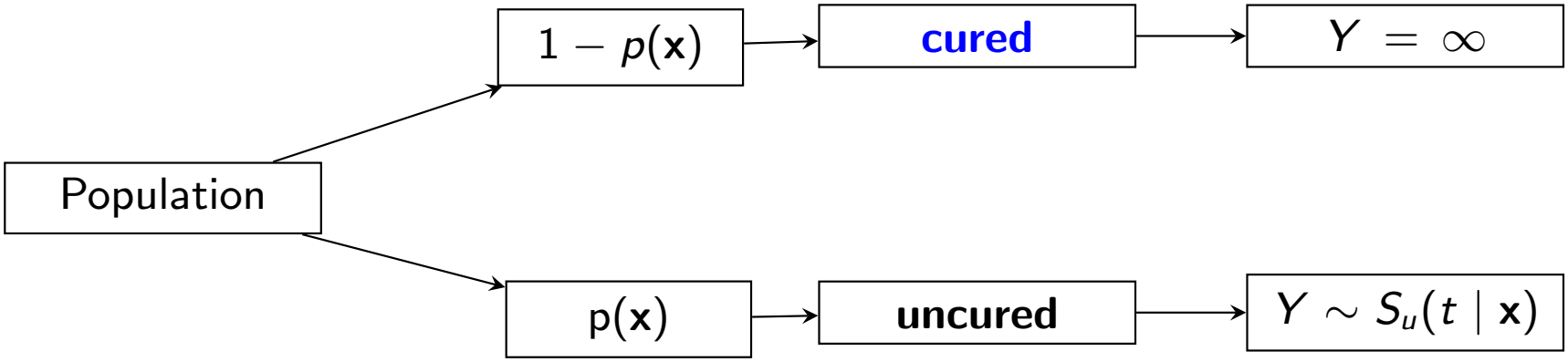
There might be situations where some censored observations are identified to be cured.



Cure threshold: Laska and Meisner (1992), Tan (2006), Nieto-Barajas and Yin (2008), Bernhardt (2016)
Based on a diagnostic test: Wu (2010)
Randomly: Betensky and Schoenfeld (2001), Safari et al (2021)

Mixture Cure Models (MCM)

- Boag (1949) introduced MCM where he defined that the survival function of the population of individuals was presented as a mixture:



$$S(t|\mathbf{x}) = \underbrace{1 - p(\mathbf{x})}_{\text{Cure probability}} + p(\mathbf{x}) \underbrace{S_u(t|\mathbf{x})}_{\text{Latency}}$$

- The **probability of cure**: $1 - p(\mathbf{x}) = P(U = 0 \mid \mathbf{X} = \mathbf{x})$.
- The **latency**: $S_u(t|\mathbf{x}) = P(Y > t \mid \mathbf{X} = \mathbf{x}, U = 1)$.

Nonparametric estimation of the latency survival function

Index

Nonparametric estimation in usual MCM

Nonparametric estimators for the **cure rate** (Xu and Peng, 2014; López-Cheda et al, 2017a; 2020) and for the **latency** (López-Cheda et al, 2017b) are based on the relations

$$\boxed{1 - p(x) = \lim_{t\to\infty} S(t|x) \quad \text{and} \quad S_u(t|x) = \frac{S(t|x) - (1 - p(x))}{p(x)}}$$

estimating $S(t|x)$ with the generalized product-limit estimator (Beran, 1981):

$$\widehat{S}_h(t|x) = \prod_{i=1}^{n} \left(1 - \frac{\delta_{[i]} B_{[i],h}(x) \mathbf{1}(T_{(i)} \leq t)}{\sum_{j=i}^{n} B_{[j],h}(x)}\right),$$

where $B_{[i],h}(x)$ are the Nadaraya-Watson weights

$$B_{[i],h}(x) = \frac{K_h\left(x - X_{[i]}\right)}{\sum_{j=1}^{n} K_h\left(x - X_j\right)}$$

and $K_h(\cdot) = K(\cdot/h)/h$ is a kernel function and h a smoothing parameter.

Usual nonparametric estimation in MCM

In the usual MCM, the NP estimator of the latency function (López-Cheda et al, 2017b) is

$$\widetilde{S}_{u,h}(t|x) = \frac{\widehat{S}_h(t|x) - (1 - \widehat{p}_h(x))}{\widehat{p}_h(x)}$$

It will be referred as **NPSXX estimator**.

Usual nonparametric estimation in MCM

In the usual MCM, the NP estimator of the latency function (López-Cheda et al, 2017b) is

$$\widetilde{S}_{u,h}(t|x) = \frac{\widehat{S}_h(t|x) - (1 - \widehat{p}_h(x))}{\widehat{p}_h(x)}$$

It will be referred as **NPSXX estimator**.

What if the covariates in the cure rate and the latency are different?

$$S(t|\mathbf{x}, \mathbf{z}) = 1 - p(\mathbf{z}) + p(\mathbf{z})S_u(t|\mathbf{x}).$$

The equivalent idea for a latency estimator

$$\frac{\widehat{S}_h(t|x, z) - (1 - \widehat{p}_h(z))}{\widehat{p}_h(z)}$$

is a function depending on both x and z $\Rightarrow$ It can not be considered an estimator of $S_u(t|x)$ in this model

Proposed estimator (NPSXZ)

We propose the following nonparametric product-limit estimator:

$$\hat{S}_{u,h}(t|x) = \prod_{t_i \leq t} \left(1 - \frac{\delta_i B_{i,h}(x)}{\sum_{j=1}^{n} w_j B_{j,h}(x)} \right),$$

where $\{w_1, \ldots, w_n\}$ is a vector of weights representing the conditional probability of being uncured $E(u_i|t_i, \delta_i, x_i, z_i)$, and u_i is the value of U for subject i. To estimate the weights, $\{w_1, \ldots, w_n\}$, we consider the EM algorithm.

EM algorithm

- **E-step**: We compute the conditional expectation of $l_n(p(\cdot), S_0(\cdot))$ with respect to u_i, which is equivalent to computing the conditional expectation of u_i:

$$\begin{aligned} & E[u_i|t_i, \delta_i, x_i, z_i, p^{(r-1)}(\cdot), S_u^{(r-1)}(\cdot)] \\ = \quad & \delta_i + (1-\delta_i)\frac{\hat{p}^{(r-1)}(z_i)\hat{S}_u^{(r-1)}(t_i|x_i)}{1-\hat{p}^{(r-1)}(z_i)+\hat{p}^{(r-1)}(z_i)\hat{S}_u^{(r-1)}(t_i|x_i)}, \end{aligned}$$

  where $\hat{p}^{(r-1)}(\cdot)$, $\hat{S}_u^{(r-1)}(\cdot)$ are the estimates of $p(\cdot)$, $S_u(\cdot)$ respectively in the last iteration of the EM algorithm.

- **M-step**: We maximize $l_{1n}(p(\cdot))$ and $l_{2n}(S_u(\cdot))$ after u_i is replaced with $E[u_i|t_i, \delta_i, x_i, z_i, p^{(r-1)}(\cdot), S_u^{(r-1)}(\cdot)]$ to update $p(\cdot)$ and $S_u(\cdot)$.

Implementation

1. Compute $\hat{p}_{h_{1,i}}(z_i)$ and $\widetilde{S}_{u,h_{2,i}}(t_i|x_i)$, the NP estimators in the MCM with one single covariate, using the R package `npcure` (López-de-Ullibarri et al, 2020).
2. Set $r = 1$ and denote $\hat{S}^{(0)}_{u,h_{2,i}}(t_i|x_i) = \widetilde{S}_{u,h_{2,i}}(t_i|x_i)$
3. **E-step**: for $r > 1$ compute

$$w_i^{(r)} = \delta_i + (1-\delta_i)\frac{\hat{p}_{h_{1,i}}(z_i)\hat{S}^{(r-1)}_{u,h_{2,i}}(t_i|x_i)}{1-\hat{p}_{h_{1,i}}(z_i)+\hat{p}_{h_{1,i}}(z_i)\hat{S}^{(r-1)}_{u,h_{2,i}}(t_i|x_i)}, \quad i = 1,\ldots,n.$$

4. **M-step**: update $\hat{S}^{(r)}_u(\cdot)$ from:

$$\hat{S}^{(r)}_{u,h_{2,i}}(t_i|x_i) = \prod_{t_k \le t_i}\left(1 - \frac{\delta_k B_{k,h_{2,i}}(x_i)}{\sum_{j=i}^{n} w_j^{(r)} B_{j,h_{2,i}}(x_i)}\right).$$

5. Repeat Step 3 and 4 until convergence.

Implementation

1. Compute $\hat{p}_{h_{1,i}}(z_i)$ and $\widetilde{S}_{u,h_{2,i}}(t_i|x_i)$, the NP estimators in the MCM with one single covariate, using the R package `npcure` (López-de-Ullibarri et al, 2020).
2. Set $r = 1$ and denote $\hat{S}^{(0)}_{u,h_{2,i}}(t_i|x_i) = \widetilde{S}_{u,h_{2,i}}(t_i|x_i)$
3. **E-step**: for $r > 1$ compute

$$w_i^{(r)} = \delta_i + (1-\delta_i)\frac{\hat{p}_{h_{1,i}}(z_i)\,\hat{S}^{(r-1)}_{u,h_{2,i}}(t_i|x_i)}{1-\hat{p}_{h_{1,i}}(z_i)+\hat{p}_{h_{1,i}}(z_i)\,\hat{S}^{(r-1)}_{u,h_{2,i}}(t_i|x_i)}, \quad i = 1,\ldots,n.$$

4. **M-step**: update $\hat{S}^{(r)}_u(\cdot)$ from:

$$\hat{S}^{(r)}_{u,h_{2,i}}(t_i|x_i) = \prod_{t_k \le t_i}\left(1 - \frac{\delta_k B_{k,h_{2,i}}(x_i)}{\sum_{j=i}^{n} w_j^{(r)} B_{j,h_{2,i}}(x_i)}\right).$$

5. Repeat Step 3 and 4 until convergence.

Alternative estimator (NPSXZ2)

The estimation of $S_u(t|x)$ can have a simplified version:

$$\widetilde{S}_{u,h}(t|x) = \prod_{t_i \leq t} \left(1 - \frac{\delta_i B_{i,h}(x)}{\sum_{j=1}^{n} \widetilde{w}_j B_{j,h}(x)} \right),$$

where $\widetilde{w}_j$ are estimated with the EM algorithm considering:

$$1 - \hat{p}_n = \hat{S}_n(t_n),$$

where $\hat{S}_n(t_n)$ is the Kaplan-Meier estimator of the survival function $S(t)$ evaluated at the largest observed time t_n.

- **Advantage**: No bandwidth needed when computing $\hat{p}_n$.
- **Disadvantage**: less efficient than the NPSXZ method since the NPSXZ estimator takes into account the dependency of the cure rate $1 - p(z)$ on the covariate Z.

Bandwidth selection

Bootstrap bandwidth selection method

For a given x, the optimal local MISE bandwidth $h(x)$ for the estimator $\hat{S}_{u,h}(t|x)$ is approximated by the minimizer of the bootstrap version of the MISE:

$$\text{MISE}^*_x(h) \simeq \frac{1}{B}\sum_{b=1}^{B}\int \left(\hat{S}^{*(b)}_{u,h(x)}(t|x) - \hat{S}_{u,g(x)}(t|x)\right)^2 w(t,x)dt,$$

- $\hat{S}^{*(b)}_{u,h(x)}(t|x)$ is the proposed estimator of $S_u(t|x)$ computed with bandwidth $h(x)$ and based on the b-th bootstrap sample
- $\hat{S}_{u,g(x)}(t|x)$ is the proposed estimator computed using a pilot bandwidth $g(x)$ and based on the original sample
- $w(t,x)$ is an appropriate weight function intended to give lower weight at the right tail of the distribution
- B is the number of bootstrap samples.

Bootstrapping in the MCM $S(t|x) = 1 - p(x) + S_u(t|x)$

Obvious bootstrap

For each i, generate X_i^* iid from the empirical distribution of $(X_1, \ldots, X_n)$, generate Y_i^* from the GPL estimator of the survival funtion $\hat{S}_g(t|X_i^*)$, generate C_i^* from the GPL estimator of the censoring distribution $\hat{G}_g(t|X_i^*)$, and compute:

$$T_i^* = \min(Y_i^*, C_i^*) \quad \text{and} \quad \delta_i^* = \mathbf{1}(Y_i^* \leq C_i^*)$$

Simple weighted bootstrap

For each i, generate X_i^* iid from the empirical distribution of $(X_i, \ldots, X_n)$, and generate (T_i^*, δ_i^*) from the weighted empirical distribution

$$\hat{F}_g(t, d|X_i^*) = \sum_{j=1}^{n} B_{g,j}(X_i^*)\mathbf{1}(T_j \leq t, \delta_j \leq d)$$

.

Without ties in the observed times, both methods are equivalent (Li and Datta, 2001; Safari et al, 2022)

Bootstrapping in the MCM $S(t|x,z) = 1 - p(z) + S_u(t|x)$

Obvious bootstrap

For each i, generate (X_i^*, Z_i^*) iid from the empirical distribution of $\{(X_1, Z_1), \ldots, (X_n, Z_n)\}$, generate Y_i^* from an estimator of the survival funtion $\hat{S}_g(t|X_i^*, Z_i^*)$, generate C_i^* from an estimator of the censoring distribution $\hat{G}_g(t|X_i^*, Z_i^*)$, and compute:

$$T_i^* = \min(Y_i^*, C_i^*) \quad \text{and} \quad \delta_i^* = \mathbf{1}(Y_i^* \leq C_i^*)$$

Simple weighted bootstrap

For each i, generate X_i^* iid from the empirical distribution of $(X_i, \ldots, X_n)$, and generate (T_i^*, δ_i^*) from the distribution

$$\hat{S}_g(t|X_i^*, Z_i^*)(1 - \hat{G}_g(t|X_i^*, Z_i^*))$$

.

Bootstrapping under this MCM model with the simple weighted resampling method becomes more complicated, given the dependency of $1 - p(z)$ and the latency $S_u(t|x)$ on different covariates. As a consequence, the obvious bootstrap is considered instead.

Simulation study

Index

Simulation study

Aim: To assess the finite sample performance of the proposed NPSXZ and NPSXZ2 estimators when the cure rate and the latency depend on different covariates.

Two existing methods are considered for reference:

- The **semiparametric method** by Peng and Dear (2000):
  - Implemented in the `smcure` package (Cai et al, 2012). It fits the cure rate $1 - p(z)$, with a logistic link function, and the latency $S_u(t|x)$, with a PH model.
- The **NPSXX estimator** by López-Cheda et al (2017a,b):
  - Implemented in the `npcure` package (López-Cheda et al, 2021).

Data generation

- The covariates X and Z are generated independently from a $U(-10, 20)$
- The censoring times, C, are generated from an $Exp(0.3)$
- The failure times, Y, are generated under three settings.

1. **Setting 1** Censoring rate 60.99%; overall cure rate 29.07%:

$$p(z) = \frac{\exp(\beta_0 + \beta_1 z)}{1 + \exp(\beta_0 + \beta_1 z)}$$

$$S_u(t|x) = \begin{cases} \dfrac{\exp(-\lambda(x)t) - \exp(-\lambda(x)\tau_0)}{1 - \exp(-\lambda(x)\tau_0)} & \text{if } t \leq \tau_0 \\ 0 & \text{if } t > \tau_0 \end{cases},$$

with $\beta_0 = 0.476$, $\beta_1 = 0.358$, $\tau_0 = 4.605$, and $\lambda(x) = \exp((x + 20)/40)$.

2. **Setting 2** Censoring rate 47.30%; overall cure rate 41.07%:

$$p(z) = \frac{\exp\left(\beta_0 + \beta_1 z + \beta_2 z^2 + \beta_3 z^3\right)}{1 + \exp\left(\beta_0 + \beta_1 z + \beta_2 z^2 + \beta_3 z^3\right)}$$

$$S_u(t|x) = \frac{1}{2}\left\{\exp[-\alpha(x)t^5] + \exp(-100t^5)\right\},$$

where $\beta_0 = 0.0476$, $\beta_1 = -0.2558$, $\beta_2 = -0.0027$, $\beta_3 = 0.004$, and $\alpha(x) = \exp(-0.01x^2)$.

3. **Setting 3**: same forms as in Setting 1 except that both $p(x)$ and $S_u(t|x)$ depend on the same covariate X.

- $n = 50$, 100, and 200
- $N = 1000$ samples
- The MISEs of each estimator of the latency function with Monte Carlo are approximated as follows:

$$\text{MISE}(x) \equiv \frac{1}{N} \sum_{j=1}^{N} \int \left(\hat{S}_u^{(j)}(t|x) - S_u(t|x) \right)^2 w(t,x) dt,$$

  - $\hat{S}_u^{(j)}(t|x)$ is the estimated latency survival function from sample j computed with each of the aforementioned method
  - $w(t,x) = 1(a_x \leq t \leq b_x)$ is the weight function, where $a_x = 0$ and b_x is the 90th pertentile of $S_u(t|x)$.

MISE in Setting 1 $S(t|x,z) = 1 - p(z) + S_u(t|x)$

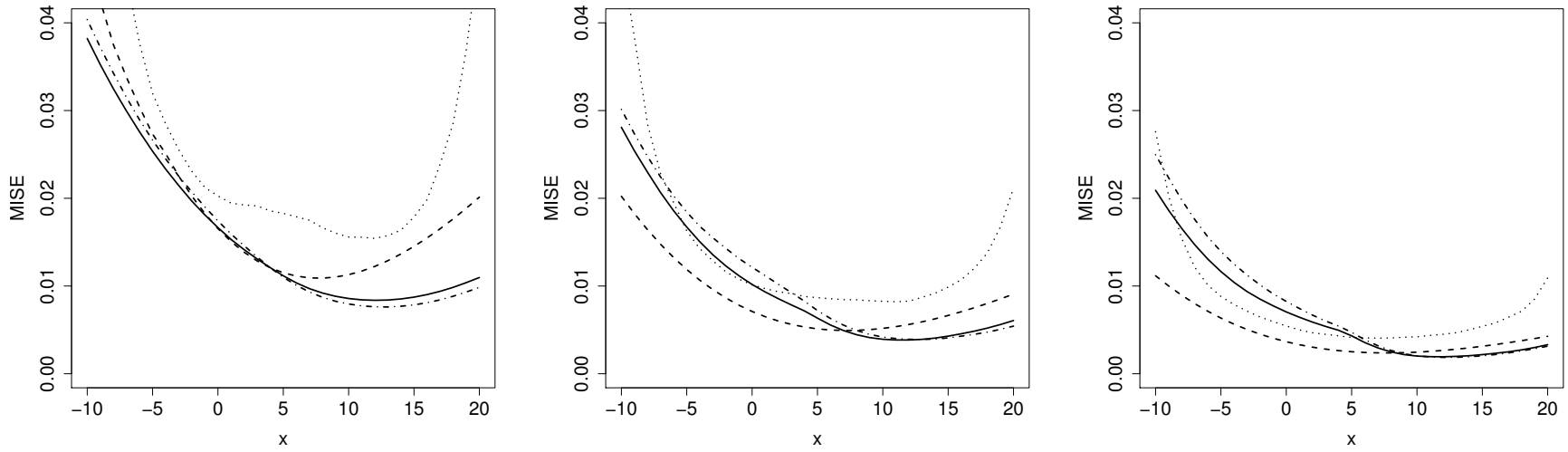


Figure 1: MISEs of the NPSXZ estimator (solid line), the NPSXZ2 estimator (dot-dashed line), the NPSXX estimator (dotted line), and the semiparametric estimator (dashed line) with sample sizes $n = 50$ (left), $n = 100$ (center) and $n = 200$ (right), for **Setting 1**.

MISE in Setting 2 $S(t|x,z) = 1 - p(z) + S_u(t|x)$

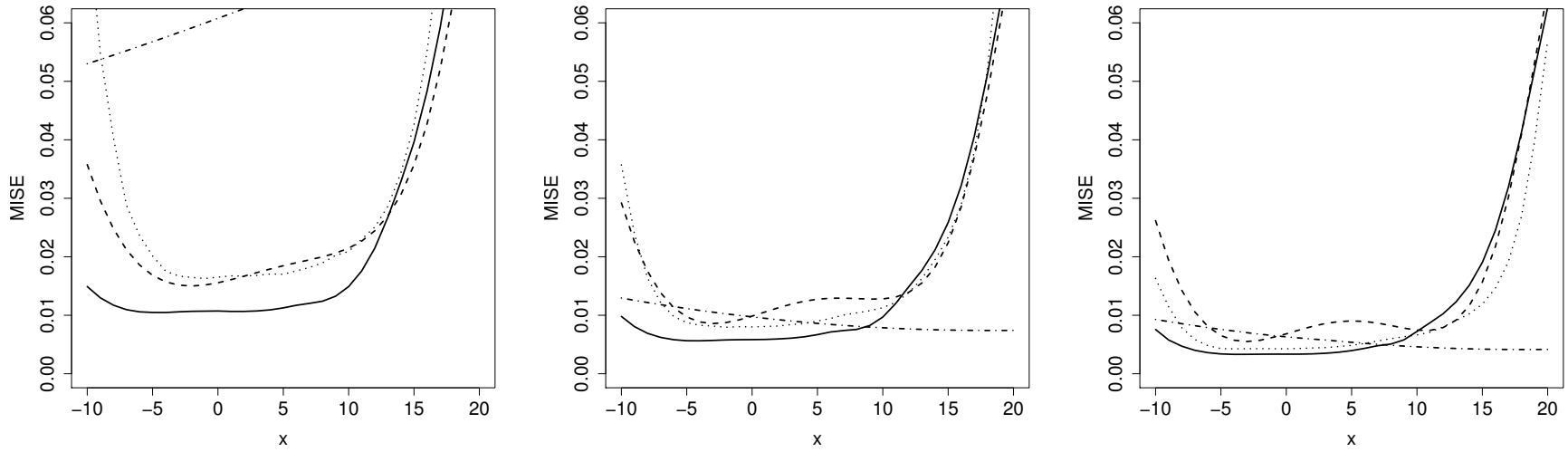


Figure 2: MISEs of the NPSXZ estimator (solid line), the NPSXZ2 estimator (dot-dashed line), the NPSXX estimator (dotted line), and the semiparametric estimator (dashed line) with sample sizes $n = 50$ (left), $n = 100$ (center) and $n = 200$ (right), for **Setting 2**.

MISE in Setting 3 $S(t|x) = 1 - p(x) + S_u(t|x)$

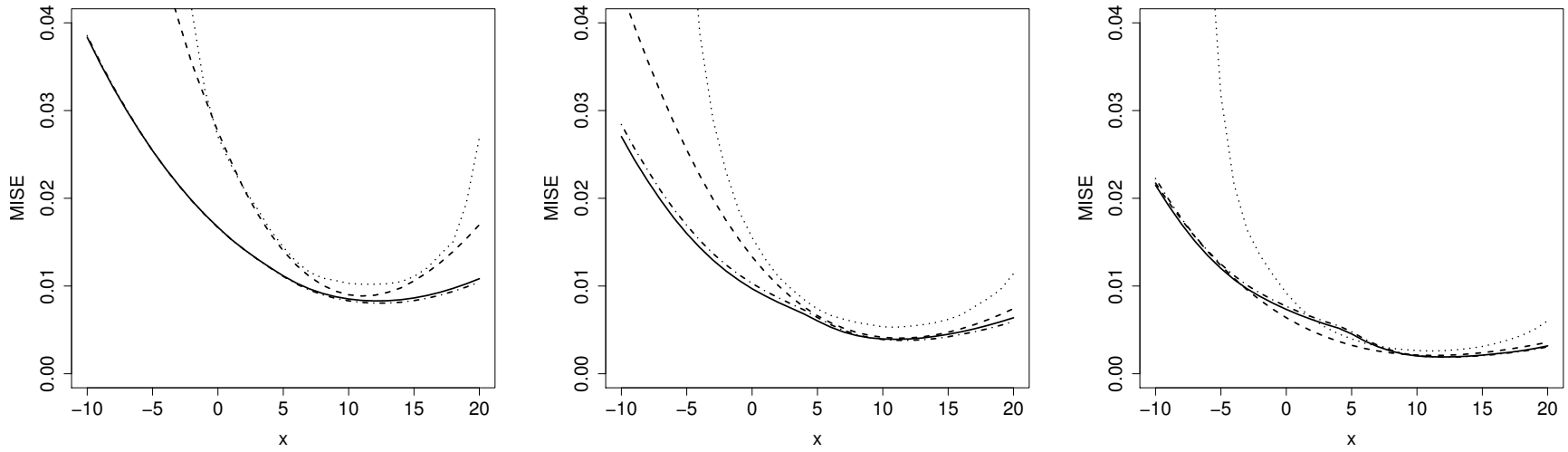


Figure 3: MISEs of the NPSXZ estimator (solid line), the NPSXZ2 estimator (dot-dashed line), the NPSXX estimator (dotted line), and the semiparametric estimator (dashed line) with sample sizes $n = 50$ (left), $n = 100$ (center) and $n = 200$ (right), for **Setting 3**.

Robustness of the NPSXZ estimator to the choice of the bandwidth

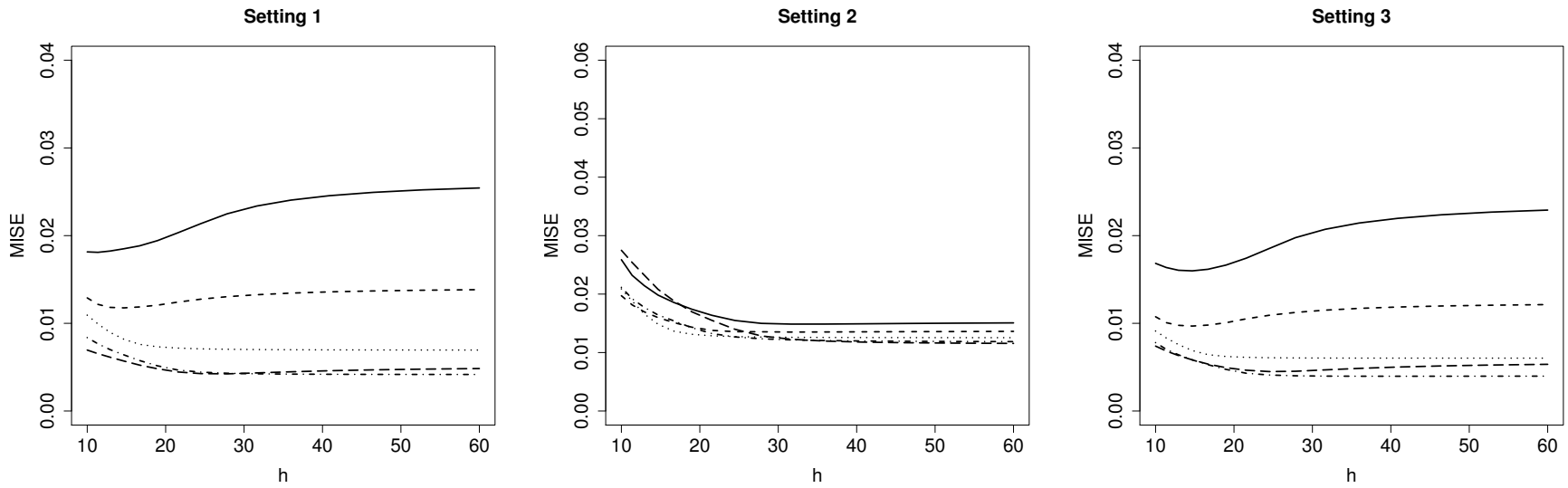


Figure 4: MISEs vs h for the NPSXZ estimator, with sample size $n = 100$, for Setting 1 (left) and Setting 2 (center) and Setting 3 (right), for covariate values $x = -5$ (solid line), $x = 0$ (dashed line), $x = 5$ (dotted line), $x = 10$ (dot-dashed line) and $x = 15$ (long dashed line).

An application to time to bankruptcy

Index

Real data set: banks (time to bankrupcy)

- 500 commercial banks insured by the Federal Deposit Insurance Corporation (FDIC), studied by Beretta and Heuchenne (2019)
- Event of interest: bankruptcy or bank's closure by the FDIC.
  - 5.6% banks experienced the event of interest (bankruptcy)
- Follow-up time: period 2006 - 2017

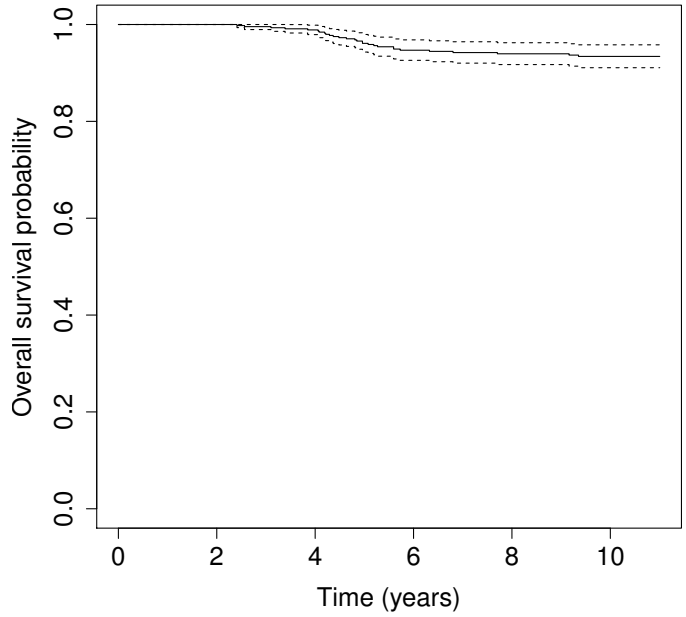


Figure 5: KM estimation of the survival time until bankrupcy of the commercial banks dataset.

Objectives

- To estimate the probability of not becoming bankrupt **(cure rate)**
- To estimate the distribution of the time to bankruptcy for the banks that will be bankrupt **(latency)**

We consider a set of bank-specific explanatory variables related to some of the 5 components of the well-known CAMEL rating system: capital adequacy, asset quality, management efficiency, earnings, and liquidity.

Covariates

- **COREDEP (Z)**: Retail deposits, the most stable source of funding for lending activities.

  Prob. bankruptcy Decreases significantly for large values of COREDEP ($p_{BH} < 0.001$, $p_{KS} < 0.001$, $p_{CvM} < 0.001$)

  Latency Not affected significantly ($p_{BH} = 0.2638$).

- **LOANS (X)**: Total loans, measures the asset quality and it is usually the least liquid and most risky asset.

  Prob. bankruptcy Not affected significantly ($p_{BH} = 0.8294$, $p_{KS} = 0.1771$ $p_{CvM} = 0.1006$)

  Latency High values of LOANS are associated to longer times to bankruptcy ($p_{BH} = 0.0063$).

- **ROA (X, Z)**: Return on assets, the capability of a bank to generate earnings. Large values of ROA are associated with stronger and safer banks.

  Prob. bankruptcy Affected significantly ($p_{BH} < 0.001$, $p_{KS} < 0.001$, $p_{CvM} < 0.001$)

  Latency Affected significantly ($p_{BH} = 0.0154$).

Probability of cure $1 - p(x)$ (not bankrupt)

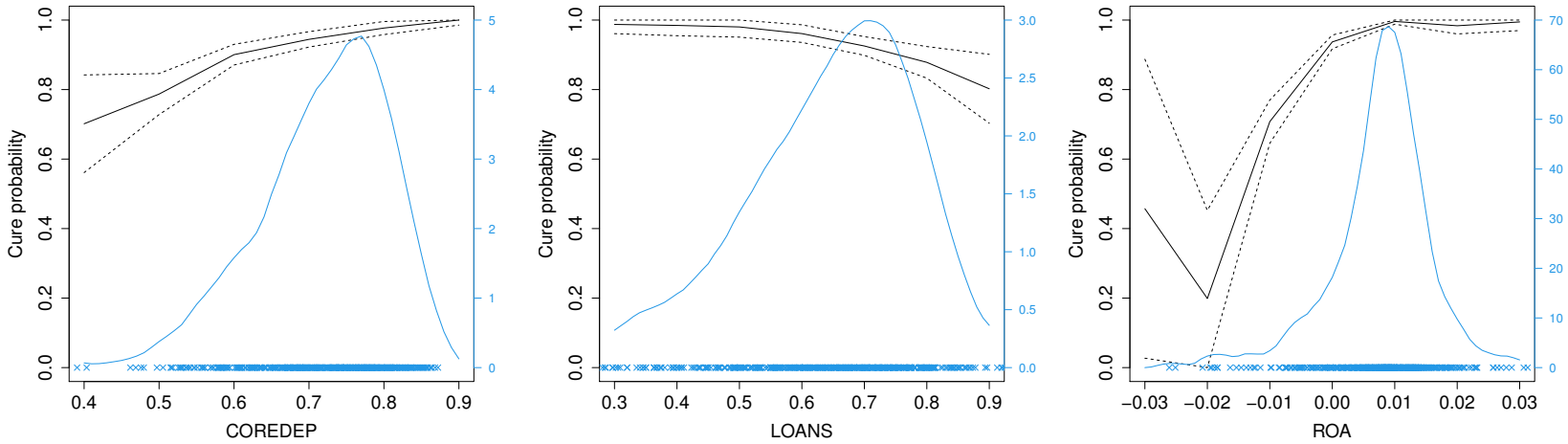


Figure 6: NP estimation of the probability of immune to bankruptcy (solid black line) as a function of COREDEP (left), LOANS (center) and ROA (right). The 95% confidence intervals (dashed black lines) are computed using the percentile bootstrap method. The blue line represents the Parzen–Rosenblatt density estimations of the covariates, using Sheather and Jones' plug-in bandwidth.

The probability of being immune to bankruptcy:

- increases as COREDEP or ROA increases
- does not appear to depend on LOANS

Estimation of $S_u(t|x)$

with X = LOANS and Z = COREDEP for $1 - p(z)$

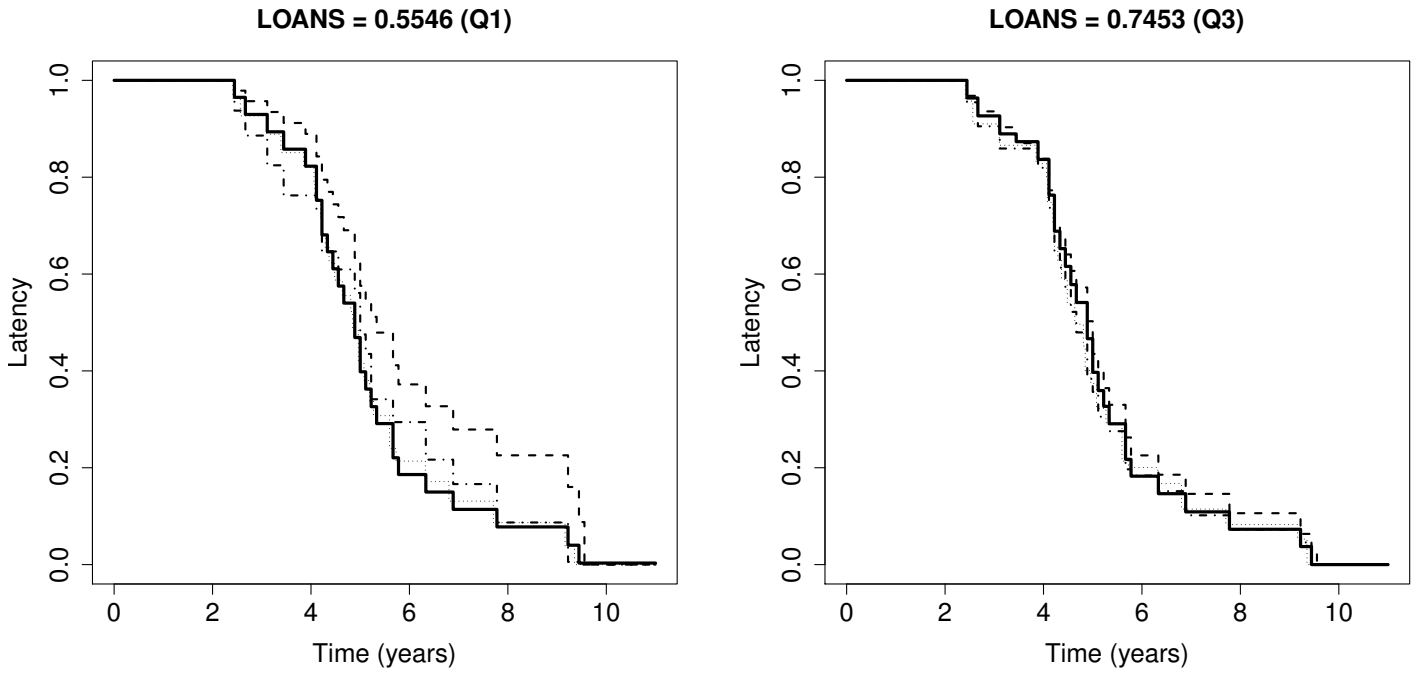


Figure 7: Estimation of the latency with the NPSXZ estimator (solid line), the NPSXZ2 estimator (dot-dashed line), the semiparametric estimator (dashed line), and the NPSXX estimator (dotted line) when LOANS = 0.5546 (left) and LOANS = 0.7453 (right).

Estimation of $S_u(t|x)$

with X = ROA and Z = ROA for $1 - p(z)$

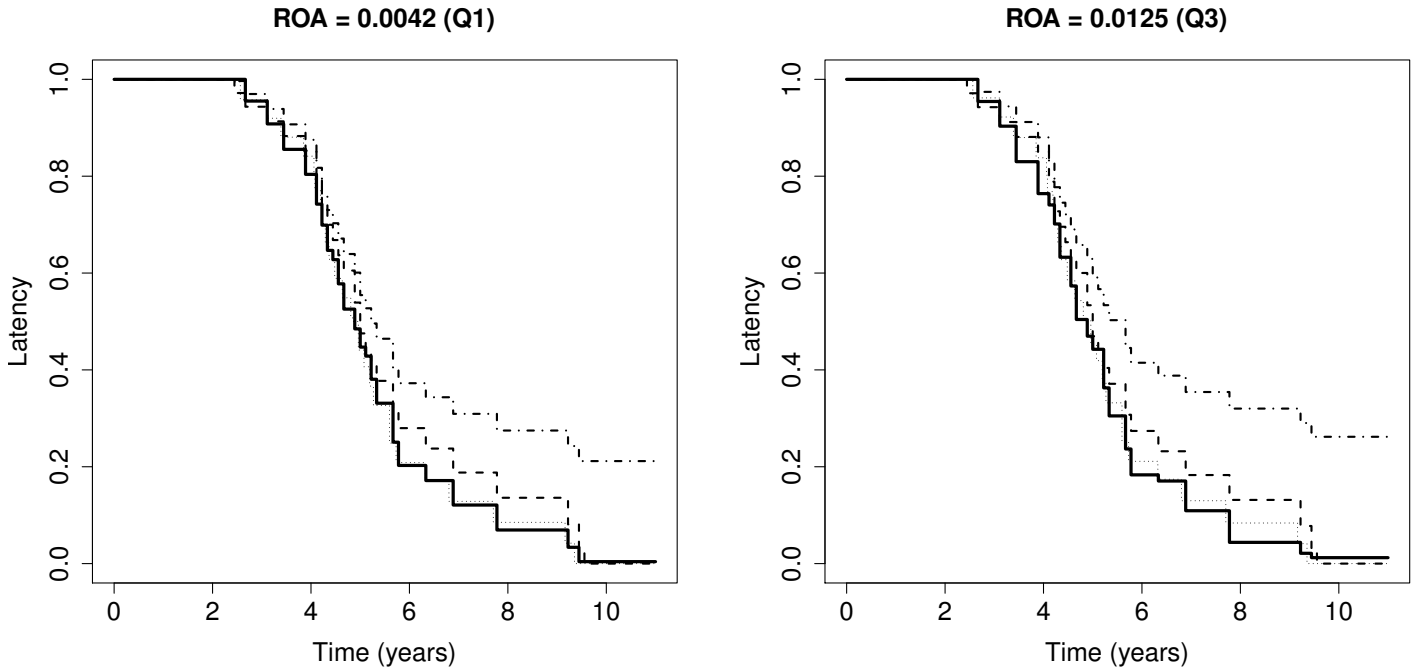


Figure 8: Estimation of the latency with the NPSXZ estimator (solid line), the NPSXZ2 estimator (dot-dashed line), the semiparametric estimator (dashed line), and the NPSXX estimator (dotted line) when ROA = 0.0042 (left) and 0.0125 (right).

Conclusions

Index

Conclusions

- Two nonparametric estimators for the latency distribution, NPSXZ and NPSXZ2, were introduced in the MCM:
  - They do not require the covariates in the cure rate and latency parts to be the same
  - They do not involve any parametric assumptions
  - They can be applied to discrete and categorical covariates
  - They hinge on a suitable choice of the smoothing parameter or bandwidth $\rightarrow$ bootstrap bandwidth selector
- **Simulation study**: the proposed estimators perform better than the existing estimators when the assumptions for the existing estimators are not fulfilled, while they still maintain their strong performance when the assumptions for the existing estimators are fulfilled
- **Real data**: differences in the latency survival estimates from the proposed estimators and the existing ones $\rightarrow$ the assumptions of the existing methods may not hold

Future work

- Asymptotic properties for the proposed estimators.
- Multi-dimensional covariates:
  - Using multivariate weight functions $\mathbf{B}_{i,\mathbf{h}}(\mathbf{x})$ in $\mathbb{R}^p$, for example, the NW weights with a multivariate product kernel $\mathbf{K}(\mathbf{u}) = \prod_{l=1}^{p} K_l(u_l)$
  - Using single index-models.
- Generalization to left censoring, interval censoring, truncation, time-dependent covariates and dependent censoring

References

[1] Beretta, A. and Heuchenne, C. (2019). Variable selection in proportional hazards cure model with time-varying covariates, application to US bank failures. *Journal of Applied Statistics*, 46(9), 1529-1549.

[2] Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society, Series B*, 11, 15–44.

[3] Cai, C., Zou, Y., Peng, Y. and Zhang, J. (2012). smcure: Fit Semiparametric Mixture Cure Models. R package version 2.0. `https://CRAN.R-project.org/package=smcure`

[4] López-Cheda, A., Cao, R., Jácome, M. A. and Van Keilegom, I. (2017a). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models.*Computational Statistics and Data Analysis*, 105, 144–165.

[5] López-Cheda, A., Jácome, M. A. and Cao, R. (2017b). Nonparametric latency estimation for mixture cure models. *TEST*, 26, 353—376.

[6] López-Cheda, A., Jácome, M. A. and López-de-Ullibarri, I. (2021). npcure: An R Package for Nonparametric Inference in Mixture Cure Models *The R Journal*, 13, 21–41.

[7] Peng, Y. and Dear, K. B. G. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, 56, 237-243.

[8] Xu, J. and Peng, Y. (2014). Nonparametric cure rate estimation with covariates. *The Canadian Journal of Statistics*, 42(1), 1–17.