Supplementary material to "Non-parametric cure rate estimation under insufficient follow-up using extremes"

Mikael Escobar-Bach * Ingrid Van Keilegom *

KU Leuven *KU Leuven*

April 30, 2019

In this supplement, we provide additional simulation results for the generalized extreme value distribution with $\gamma = -0.7$ or 1, and with $p = 0.25, 0.5$ or 0.75. The results are obtained from $N = 200$ iterations and for samples of size $n = 1000$. Four aspects of the proposed estimation method are studied in this supplement : the selection of the tuning parameter y, the misspecification of the sign of the extreme value index γ, the case of sufficient follow-up, and the construction of confidence intervals for p.

1 Selection of the tuning parameter y

As mentioned in Sections 4 and 5, the estimator $\widehat{p}_y$ requires the choice of the tuning parameter y. In this section, we will further inspect the sensitivity of our estimator $\widehat{p}_y$ to the choice of y through a small simulation study, which underlines the usefulness of our selection method. We perform the computations of $\widehat{p}_y$ when y is fixed to either 0.5 or 0.6, and the results are shown in Figures 1 and 2, respectively. The plots represent the results

*ORSTAT, KU Leuven, Naamsestraat 69, B-3000 Leuven, Belgium. Email: mikael.escobarbach@kuleuven.be and ingrid.vankeilegom@kuleuven.be (corresponding author).
Financial support from the European Research Council (2016-2021, Horizon 2020 / ERC grant agreement No. 694409) is gratefully acknowledged.

for the mean (left) and MSE (right) of the KME $\widehat{p}_n$ and our estimator $\widehat{p}_y$ as a function of the ratio $\tau_c/\tau_{0.95}$. From these figures, we observe that :

- For $y = 0.5$ and $\gamma = 1$, our estimator approaches in a satisfactory way the value of p from a small value of $\tau_c/\tau_{0.95}$ on, and outperforms the KME. When $\gamma = -0.7$, $\widehat{p}_y$ has difficulties in estimating p with a discrepancy similar to that of the KME when $\tau_c/\tau_{0.95}$ is large enough.

- For $y = 0.6$ and $\gamma = 1$, our estimator is close to the KME and has no advantage over the latter. When $\gamma = -0.7$, $\widehat{p}_y$ is a good approximation of p for a wide range of values of $\tau_c/\tau_{0.95}$, with reduced bias compared to the KME.

- Overall, the MSE curves for $\widehat{p}_y$ are higher than those for the KME $\widehat{p}_n$, though both are decreasing as a function of $\tau_c/\tau_{0.95}$.

Those observations particularly emphasize the necessity to use an appropriate data-driven parameter selection. On one hand, this provides an appropriate y for each $\gamma \in \mathbb{R}$ with proper bias-correcting effect for $\widehat{p}_y$, while, on the other hand, it greatly reduces the variability of the estimator.

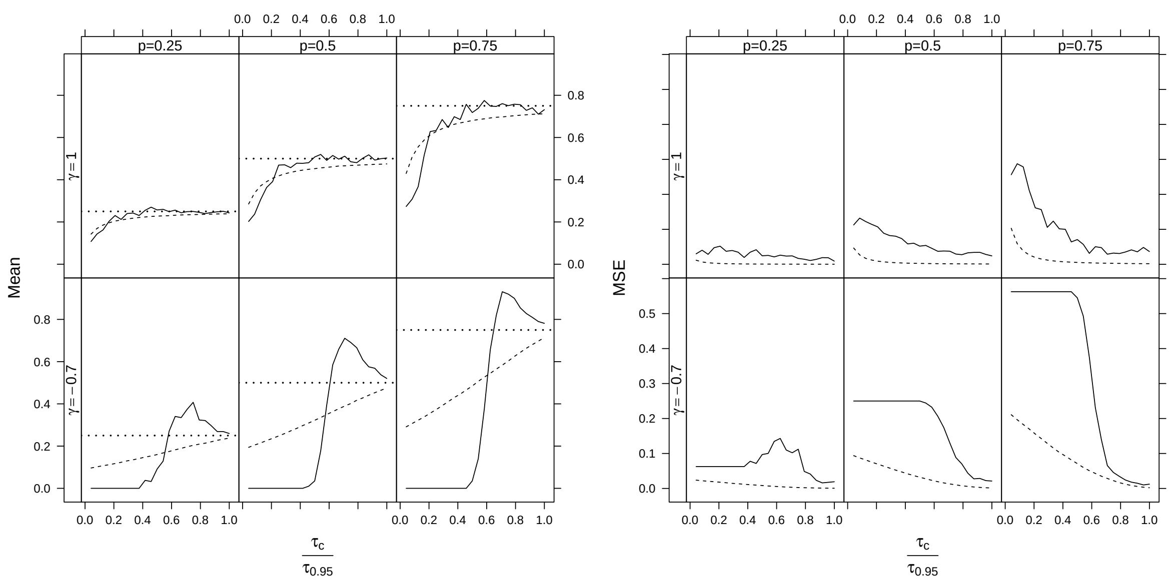

Figure 1: Mean (left) and MSE (right) of the KME $\widehat{p}_n$ (dashed) and our estimator $\widehat{p}_y$ (solid) when $y = 0.5$. The plots correspond to $p = 0.25, 0.5, 0.75$ and $\gamma = -0.7, 1$.

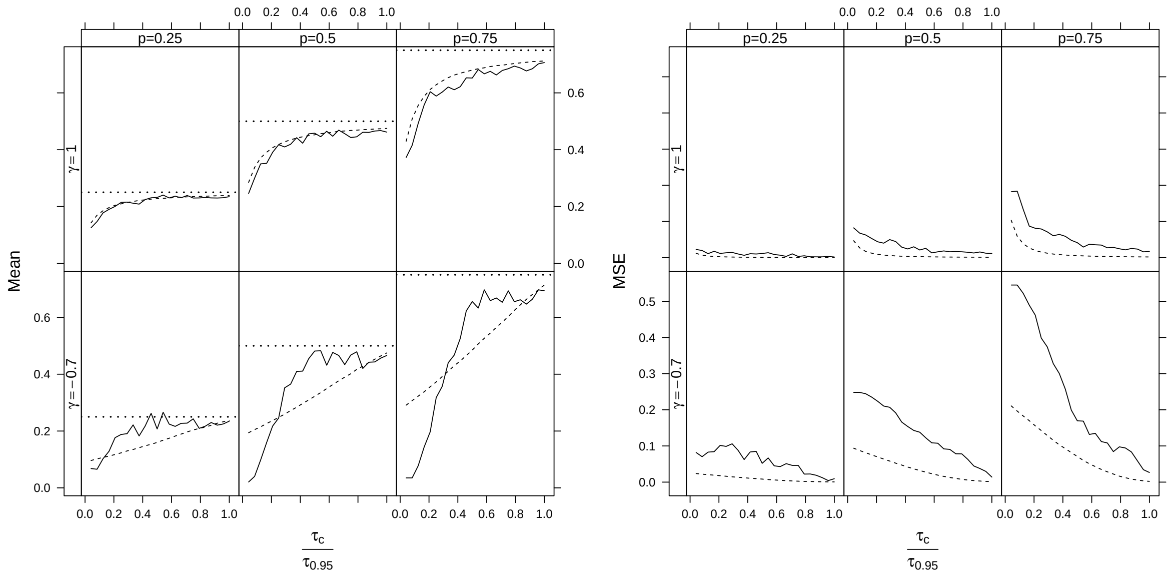

Figure 2: Mean (left) and MSE (right) of the KME $\widehat{p}_n$ (dashed) and our estimator $\widehat{p}_y$ (solid) when $y = 0.6$. The plots correspond to $p = 0.25, 0.5, 0.75$ and $\gamma = -0.7, 1$.

2 Misspecification of the sign of γ

In this section, we investigate the robustness of our method against misspecification of the sign of the extreme value index γ.

In order to apply our method, it is necessary to know the value of the right-end point τ_0 and so of the sign of γ. However, this information might be difficult to obtain in practice and one could seek some estimation approach for γ applying to cure models. As mentioned in the introduction, various techniques have been developed in survival analysis with right censored data, but they all apply to the case without cure fraction and mostly when $\tau_c = \tau_0$. The problem of estimating γ in our setting thus remains unsolved and presents an interesting but difficult challenge, which we will not address in this paper.

Nevertheless, we conducted a small simulation study when the sign of γ is missspecified in the procedure. In Figures 3 and 4, the left panel represents the mean (top) and MSE (bottom) of the KME $\widehat{p}_n$ and of our estimator $\widehat{p}_{y*}$ as a function of the ratio $\tau_c/\tau_{0.95}$. Similarly, the right panel shows the boxplots of the selected y-values along with $\tau_c/\tau_{0.95}$. In Figure 3, γ equals -0.7 but no initial transformation of the data has been performed, while in Figure 4, γ equals 1 and the function ψ has been initially applied to the data with $\tau_0 = \frac{5}{4}\widehat{\tau}_n$.

As expected, we see that our method does not provide the same satisfactory correction as when the sign of γ is correctly determined. In both configurations, $\widehat{p}_{y*}$ shows little difference with the KME and both hardly approach p. However, the MSE of $\widehat{p}_{y*}$ is similar or slightly better than that of the KME $\widehat{p}_n$.

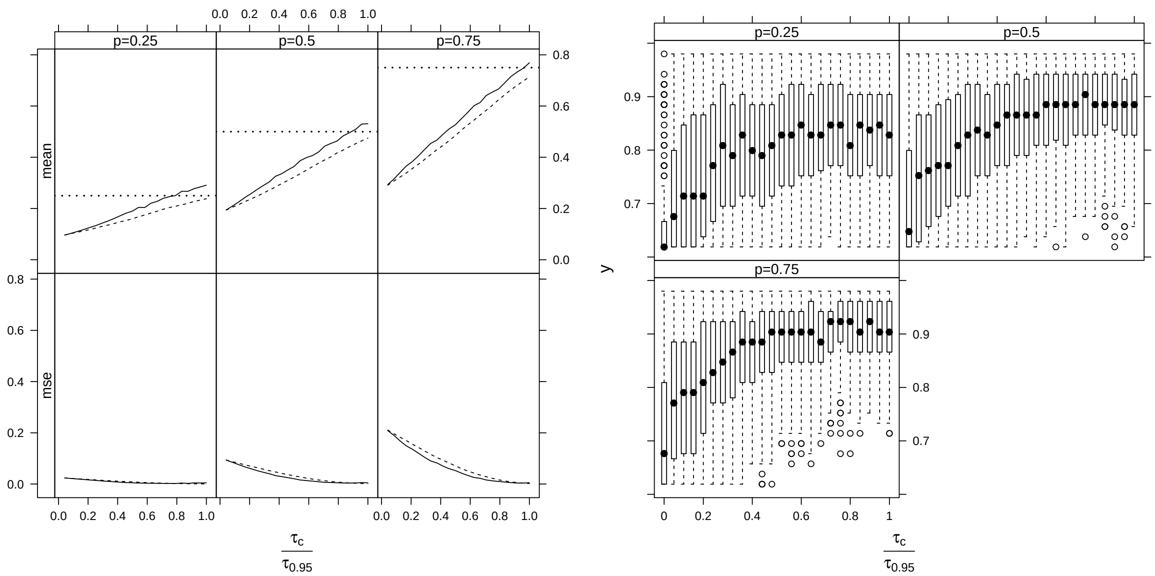

Figure 3: Left panel : Mean (top) and MSE (bottom) of the KME $\widehat{p}_n$ (dashed) and our estimator $\widehat{p}_{y*}$ (solid). Right panel : boxplots of the selected y-values. The plots correspond to $p = 0.25, 0.5, 0.75$ and $\gamma = -0.7$.

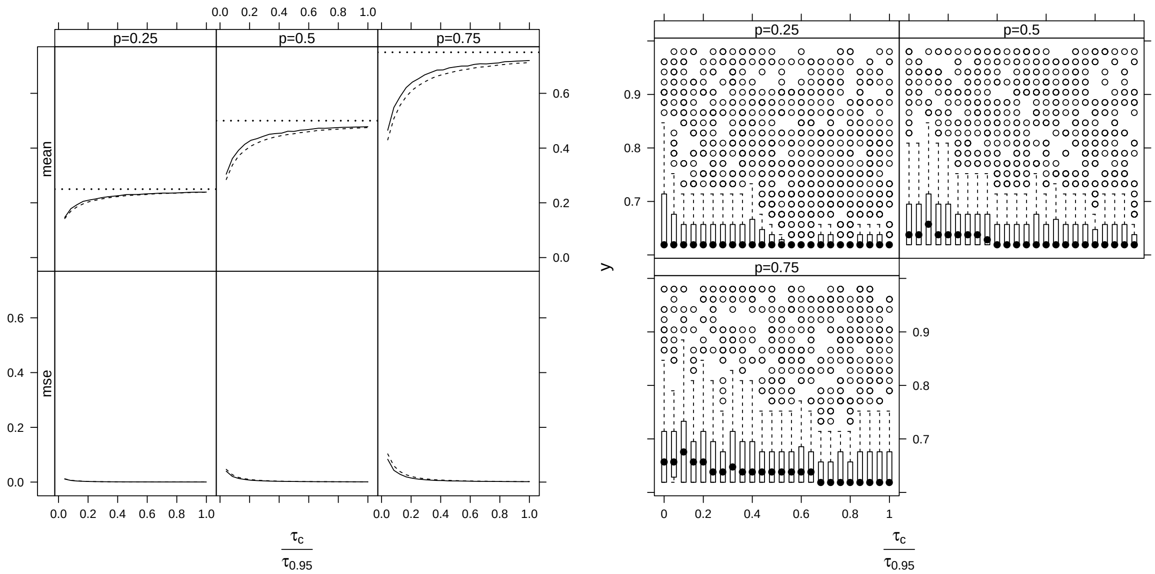

Figure 4: Left panel : Mean (top) and MSE (bottom) of the KME $\widehat{p}_n$ (dashed) and our estimator $\widehat{p}_{y*}$ (solid). Right panel : boxplots of the selected y-values. The plots correspond to $p = 0.25, 0.5, 0.75$ and $\gamma = 1$.

3 The case of sufficient follow-up

Another concern might arise regarding the practicability of our estimator $\widehat{p}_{y*}$ when the follow-up is sufficient. As it is often hard to verify in practice whether the follow-up is sufficient or not, it is important to study the robustness of our estimator against the case of sufficient follow-up. Despite the fact that our method is not designed for such a context, one might expect that our estimator performs well even when the follow-up is sufficient.

In Figure 5 we compare the KME and our estimator by means of boxplots under two different configurations with a sufficient follow-up.

- The boxplots on the first row correspond to $\gamma = 1$ and assume that the censoring variable follows an exponential distribution with parameter $\lambda = 1/\tau_{0.9}$, where $\tau_{0.9}$ is the 90th-percentile of the extreme value distribution G_γ. This particularly implies that $\tau_c = \tau_0$.
- The boxplots on the second row correspond to $\gamma = -0.7$ and assume that the

censoring variable follows an uniform distribution over the interval $[0, \tau_c]$, where $\tau_c = -2/\gamma - \tau_{0.95}$. This particularly implies that $\tau_0 < \tau_c$.

From these boxplots, we can see that apart from some outliers, our estimator $\widehat{p}_{y*}$ performs equally well as the KME $\widehat{p}_n$, and no significant differences allow us to differentiate the two estimators. This is expected in theory since $\widehat{p}_y$ might be seen as the KME, apart from a term that depends on the increments $\widehat{F}_n(\widehat{\tau}_n) - \widehat{F}_n(s\widehat{\tau}_n)$ with $s = y^2$ or y. In the context of sufficient follow-up, the latter terms tend to zero which reduces our estimator to the KME.

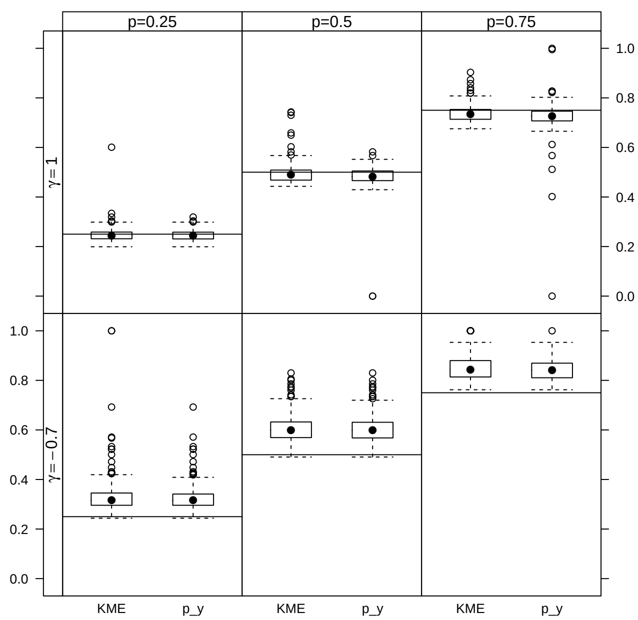

Figure 5: Boxplots of the KME $\widehat{p}_n$ (left) and of our estimator $\widehat{p}_{y*}$ (right). The plots correspond to $p = 0.25, 0.5, 0.75$ and $\gamma = -0.7, 1$. The horizontal line represents the true value of p.

4 Confidence intervals for p

In Theorems 3.1 and 3.2 the asymptotic normal distributions of both the KME $\widehat{p}_n$ and our estimator $\widehat{p}_y$ are provided. We use our simulated data to compute the empirical coverage probabilities of 95% confidence intervals for p using these asymptotic normality results.

To do so, we define the estimators

$$\widehat{a}_0 := (1-\widehat{b}_y)^2, \quad \widehat{a}_1 := -2\widehat{b}_y(1+\widehat{b}_y) \quad \text{and} \quad \widehat{a}_2 := \widehat{b}_y^2,$$

where

$$\widehat{b}_y := \frac{\widehat{F}_n(\widehat{\tau}_n) - \widehat{F}_n(y\widehat{\tau}_n)}{\widehat{F}_n(y^2\widehat{\tau}_n) - 2\widehat{F}_n(y\widehat{\tau}_n) + \widehat{F}_n(\widehat{\tau}_n)}.$$

Since F is assumed to be continuous, we also define the following estimator of the variance $v(t)$ in (3.1):

$$\widehat{v}(t) := \sum_{i=1}^{n} \frac{(\widehat{F}_n(Y_{(i)}) - \widehat{F}_n(Y_{(i-1)}))\mathbb{1}_{\{Y_{(i)}\leqslant t\}}}{(1-\widehat{F}_n(Y_{(i)}))^2(1-\widehat{F}_{n,c}(Y_{(i-1)}))},$$

where $\widehat{F}_{n,c}(Y_{(0)}) = 0$ and $\widehat{F}_{n,c}$ is the product-limit estimator of the censoring distribution given by

$$\widehat{F}_{n,c}(t) := \prod_{Y_{(i)}\leqslant t} \left(1 - \frac{1-\delta_{(i)}}{n-i+1}\right), \quad t \in \mathbb{R}.$$

It follows that the asymptotic variance of the KME $\widehat{p}_n$ and of our estimator $\widehat{p}_y$ can respectively be estimated by

$$\widehat{\sigma}^2_{KM} := \widehat{v}(\widehat{\tau}_n)(1-\widehat{F}_n(\widehat{\tau}_n))^2 \quad \text{and} \quad \widehat{\sigma}^2_y := \sum_{i,j=0}^{2} \widehat{a}_i\widehat{a}_j(1-\widehat{F}_n(y^i\widehat{\tau}_n))(1-\widehat{F}_n(y^j\widehat{\tau}_n))\widehat{v}(y^{i\vee j}\widehat{\tau}_n),$$

leading to the following confidence intervals:

$$\left[\widehat{p}_n - q\frac{\widehat{\sigma}_{KM}}{\sqrt{n}}, \widehat{p}_n + q\frac{\widehat{\sigma}_{KM}}{\sqrt{n}}\right] \quad \text{and} \quad \left[\widehat{p}_y - q\frac{\widehat{\sigma}_y}{\sqrt{n}}, \widehat{p}_y + q\frac{\widehat{\sigma}_y}{\sqrt{n}}\right],$$

where q is defined as the 95th-percentile of the standard normal distribution. In Table 1, we present the coverage probabilities of these intervals.

Table 1: Coverage probabilities of 95% confidence intervals for p, when $\tau_c/\tau_{0.95}$ equals $0.25, 0.5, 0.75$ or 1.

		$p=0.25$				$p=0.5$				$p=0.75$			
γ		0.25	0.5	0.75	1	0.25	0.5	0.75	1	0.25	0.5	0.75	1
-0.7	$\widehat{p}_n$	0.00	0.01	0.40	0.85	0.00	0.00	0.06	0.81	0.00	0.00	0.00	0.77
	$\widehat{p}_{y*}$	0.73	0.89	0.98	0.88	0.59	0.81	0.96	0.93	0.47	0.75	0.93	0.95
1	$\widehat{p}_n$	0.27	0.69	0.83	0.90	0.01	0.29	0.60	0.69	0.00	0.10	0.26	0.44
	$\widehat{p}_{y*}$	0.91	0.95	0.93	0.94	0.96	0.97	0.95	0.98	0.91	0.92	0.98	0.96

We can conclude from these tables that the empirical coverage probabilities increase with the value of the ratio $\tau_c/\tau_{0.95}$, which is expected since for increasing τ_c, both the estimators of $F(\tau_c)$ and $p_y(\tau_c)$ converge to p. Overall, the results are better for $\gamma = 1$ than for $\gamma = -0.7$, though we observe a huge discrepancy between the KME $\widehat{p}_n$ and our estimator $\widehat{p}_{y*}$. On one hand, the probabilities for the KME never manage to reach 0.95 with more difficulties when p is large. We can particularly observe that many results are close to zero when $\gamma = -0.7$, or exceed 0.80 only for $\tau_c = \tau_{0.95}$. On the other hand, the results of $\widehat{p}_{y*}$ tend to be better. For $\gamma = -0.7$, the probabilities are close to 0.95 as soon as $\tau_c/\tau_{0.95} \geqslant 0.75$, while for $\gamma = 1$, the coverage probabilities are close to the nominal value for a wide range of values of $\tau_c/\tau_{0.95}$.