Refining Kaplan-Meier Estimation with the Generalized Pareto Model for Survival Analysis

Yi He
University of Amsterdam

Liang Peng
Georgia State University

Dabao Zhang
University of California, Irvine

Zifeng Zhao
University of Notre Dame

October 24, 2024

Abstract

The Kaplan-Meier estimator is widely recognized as the leading nonparametric method for estimating survival functions from censored data. However, it faces challenges with tail estimation and cannot extrapolate beyond the maximum observed data point, particularly when the largest observation is censored. To address these limitations, we enhance the Kaplan-Meier estimator by fitting the upper tail of the survival function to a generalized Pareto model. This approach improves tail estimation and extends survival estimates beyond the observed maximum, regardless of whether the largest observation is censored. We derive the joint asymptotic behavior of the Kaplan-Meier estimator in both central and tail regions by analyzing exceedances over a high, finite threshold, leading to more accurate approximations. Furthermore, we establish that the confidence intervals from a random weighted bootstrap method are asymptotically correct and demonstrate its coverage performance through numerical analysis. We illustrate the estimation and inference advantages of our refined estimator in an application to the National Job Training Partnership Act study.

Keywords: Survival analysis; Generalized Pareto distribution; Random weighted bootstrap; Asymptotic normality; Duration data.

1 Introduction

Since introducing the product limit estimator of the survival function by Kaplan and Meier (1958), it has become one of the most widely used tools for analyzing lifetime data. According to PubMed[1], over the past ten years, more than 10,000 papers annually have cited the Kaplan-Meier estimator (see Figure 1). Due to its flexibility in handling censored data nonparametrically, the Kaplan-Meier estimator is also employed to estimate the mean and variance of survival times. Furthermore, it has served as a foundation for the development of many advanced statistical methods in survival analysis, such as test statistics for comparing treatment effects on survival times (Efron, 1967) and iterated least squares estimators for accelerated failure time models (Jin, Lin and Ying, 2006).

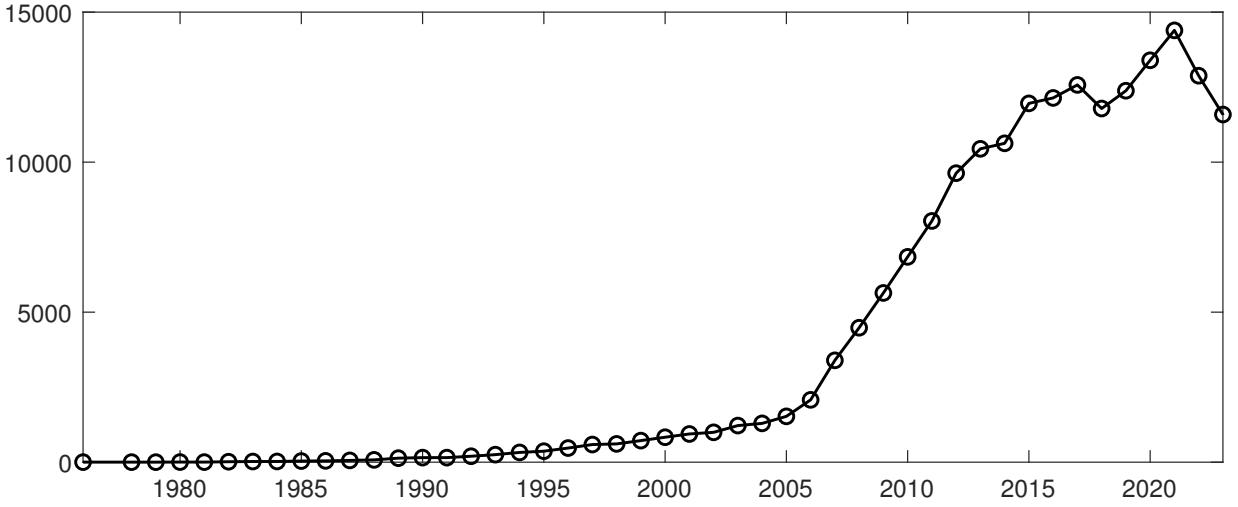

Figure 1: Annual numbers of papers citing the Kaplan-Meier estimator according to PubMed.

An often-used procedure to estimate the full-domain survival function is to treat the largest censored observation as uncensored (Efron, 1967). This naive setting may lead to serious bias in the survival function and subsequent procedures. Let the observed data

[1] `https://pubmed.ncbi.nlm.nih.gov/?term=Kaplan-Meier`

under right censoring be $X_i = \min\{T_i, C_i\}$ as i.i.d. copy of $X = \min\{T, C\}$, where T_i and C_i are random lifetime and censoring variable for individual i respectively. Specifically, we can calculate the sample mean using the Kaplan-Meier estimator of the survival function of T_i, denoted by $\widehat{S}_0(\cdot)$, given by

$$\widehat{\mu} = \int_0^\infty t d\widehat{S}_0(t) = \sum_{i=1}^{n} X_i \{\widehat{S}(X_i-) - \widehat{S}(X_i)\} \tag{1}$$

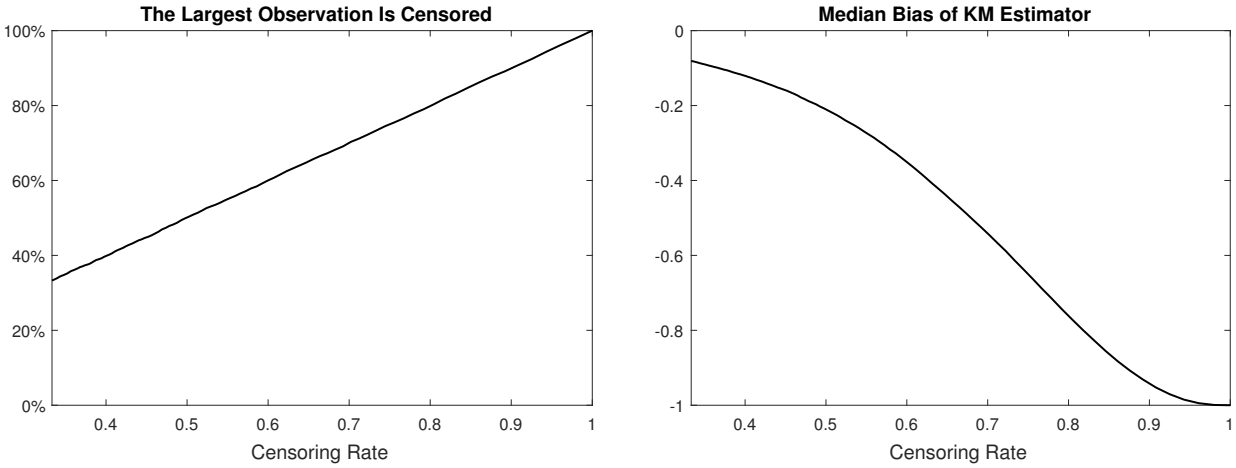

Figure 2: The probability of the largest observed value is censored (left) and biases in estimating mean lifetime from a size 100 sample using the Kaplan-Meier estimator (right). The survival time follows the standard exponential distribution, and the censoring time follows an exponential distribution with a mean ranging from 0 to 2.

The left panel of Figure 2 shows the probability that the largest observed survival time is censored in a sample of 100 individuals following a standard exponential distribution, with right censoring by an exponentially distributed variable whose mean $\mathbb{E}C$ ranges from 0 to 2. The x-axis represents the overall censoring rate, $1/(1 + \mathbb{E}C)$, which ranges from 1/3 to 1. The probability of the largest observation being censored aligns with the limit derived by Maller and Zhou (1993), equaling the overall censoring rate.

The right panel of Figure 2 illustrates that the Kaplan-Meier (KM) estimator tends to underestimate the mean lifetime in this scenario. The magnitude of the median bias, defined

as the difference between the median estimate and the true mean, increases significantly as the censoring rates rise. All results were obtained using Monte Carlo simulations with 100,000 replications.

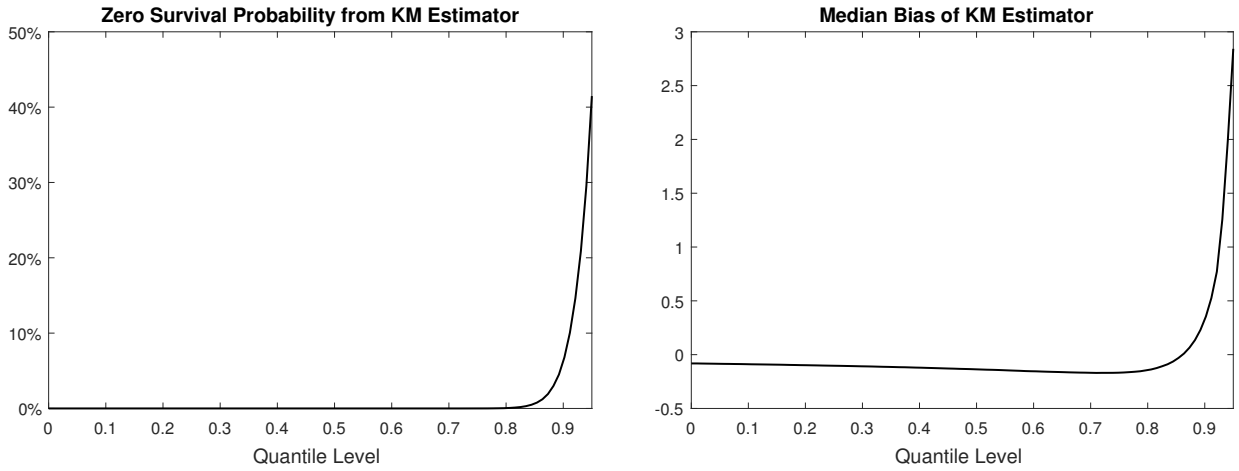

Figure 3: Probability of obtaining a zero estimate for the survival probability at various quantile levels (left) and median bias of the Kaplan-Meier estimator, defined as the difference between its median and the true value (right). Survival times follow a standard exponential distribution while censoring times follow an exponential distribution with a mean of 2.

Similarly, for any threshold t_0 within the support of lifetime distribution, the KM estimator of the mean residual life $e_T(t_0) \equiv \mathbb{E}[T_i - t_0 | T_i > t_0]$ is given by

$$\widehat{e}_T(t_0) = \frac{\int_{t_0}^{\infty} \widehat{S}_0(t)dt}{\widehat{S}_0(t_0)} - t_0 = \frac{\sum_{i:X_i \geq t_0} X_i\{\widehat{S}_0(X_i-) - \widehat{S}_0(X_i)\}}{\widehat{S}_0(t_0)} - t_0. \tag{2}$$

As the mean residual life increasingly depends on later survival times, its estimator, based on the Kaplan-Meier estimator, is even more severely affected by ignoring the largest censored values. This is demonstrated in Figure 3 for a standard exponential lifetime censored by an exponential censoring variable with a mean of 2. When the mean residual life is evaluated at a large time, the Kaplan-Meier estimator becomes unreliable. Eventually, it

cannot provide any estimate beyond the largest observed value. The left panel of Figure 3 shows that the probability of this issue becomes non-trivial and increases with t_0 at high quantile levels. The right panel shows the bias of the median of the Kaplan-Meier estimator for the mean residual life, which initially underestimates at lower quantile levels but dramatically overestimates at higher quantile levels. All these values were computed using the Monte Carlo method with 100,000 replications, each based on a sample size of 100.

On the other hand, the generalized Pareto distribution can approximate the residual life well over a long time. In particular, the extreme value theory states that when the lifetime distribution is in the domain of attraction of extreme value distribution, there exists a function $\beta(u) > 0$ such that

$$\lim_{u \to \tau_0} \sup_{0 \le x < \tau_0 - u} |S_0(x+u)/S_0(u) - G(x \mid \gamma, \beta(u))| = 0, \tag{3}$$

where S_0 denotes the population survival distribution of lifetime T and τ_0 is its right endpoint, i.e., $\tau_0 = \sup\{x : S_0(x) > 0\}$ and

$$G(x \mid \gamma, \beta) = (1 + \gamma x/\beta)^{-1/\gamma}, \quad 1 + \gamma x/\beta > 0,$$

is the survival function of the generalized Pareto distribution with the shape parameter γ and scale parameter β; see Balkema and de Haan (1974), and the overviews by Resnick (1987) and Embrechts, Klüppelberg, and Mikosch (1997).

Therefore, we propose modeling the residual life beyond a finite, possibly unknown, sufficiently high threshold using the generalized Pareto distribution (GPD). In line with He et al. (2022), this approach recognizes that practitioners typically work with a finite threshold and offers improved asymptotic theory, enabling us to derive non-degenerate joint limits for central and tail estimators. By fitting the GPD to the tail using a censored maximum likelihood method, we obtain a more efficient estimator of the tail survival distribution

while accommodating the possibility of the largest observation being censored. To avoid estimating the asymptotic variance in its complex form and to enhance finite-sample performance, we propose constructing confidence intervals using a randomly weighted bootstrap method. This approach is straightforward to implement, and we prove that it is asymptotically correct.

Our semiparametric model does not specify the central part of the distribution below the threshold. By combining the KM estimator for the regions below the threshold with our GPD estimator for the regions above, we develop a comprehensive asymptotic theory for a refined survival distribution estimator that extends across the full domain, even beyond the observed data range. This asymptotic theory applies to a much broader scope than existing literature, such as Einmahl, Fils-Villetard, and Guillou (2008) and Beirlant, Guillou, and Toulemonde (2010), which are only applicable to three specific cases where the endpoint of the lifetime variable must be smaller than that of the censoring variables. We have relaxed this assumption, as it is often violated in real-life applications, including the National Job Training Partnership Act study to be discussed here.

The rest of the paper is organized as follows. Section 2 develops a comprehensive asymptotic theory for the point estimation procedure and a random weighted bootstrap solution for interval estimation. Section 3 demonstrates the good performance of our refined estimation in various settings where the KM estimator fails. Section 4 provides an application to the national Job Training Partnership Act study, illustrating the advantages of our refined estimator in tail inference and extrapolation beyond the observed data range. We conclude the paper in Section 5. All the proofs are available in the supplementary document.

2 Asymptotic Theory

Consider a lifetime random variable $T > 0$ with a continuous survival function S_0 and a (generalized) quantile function Q_0. For a threshold u_0, possibly unknown, with exceedance probability $\alpha_0 = S_0(u_0) \in (0, 1)$, we make the following assumption regarding the exceedance $T - u_0$. Let $(x)_+ = \max\{x, 0\}$ denote the positive part of x.

Assumption 2.1 (Generalized Pareto Model). There exist a shape parameter $\gamma_0 \in \mathbb{R}$ and a scale parameter $\sigma_{\alpha_0} > 0$ with $\alpha_0 = S_0(u_0)$ such that, for $x \geq 0$

$$\mathbb{P}\left(T > u_0 + x | T > u_0\right) = G(x|\gamma_0, \sigma_{\alpha_0}) = \begin{cases} \left(1 + \frac{\gamma_0 x}{\sigma_{\alpha_0}}\right)_+^{-1/\gamma_0}, & \gamma_0 \neq 0, \\ \exp\left(-\frac{x}{\sigma_{\alpha_0}}\right), & \gamma_0 = 0. \end{cases}$$

The shape parameter γ_0 is called the extreme value index for the lifetime distribution. When $\gamma_0 < 0$, there is a finite right endpoint $\tau_0 \equiv u_0 - \frac{\sigma_{\alpha_0}}{\gamma_0}$ in the support of the distribution of T, i.e., $S_0(t) = 0$ for all $t \geq \tau_0$. When $\gamma_0 = 0$, $T - u_0 \mid T > u_0$ has an exponential distribution with mean σ_{α_0}. When $\gamma_0 > 0$, T has a heavy tail with up to $1/\gamma_0$-th finite moments. Note that σ_{α_0} is also a function of u_0 through $\alpha_0 = S_0(u_0)$.

Observe that, for any higher threshold $u > u_0$, the exceedance $T - u \mid T > u$ follows the generalized Pareto distribution with the same shape parameter γ_0 but a different scale parameter $\sigma_\alpha = (\alpha_0/\alpha)^{\gamma_0}\sigma_{\alpha_0}$, where $\alpha = S_0(u)$ is the probability of exceeding u. Specifically,

$$\mathbb{P}\left(T > u + x \mid T > u\right) = G(x \mid \gamma_0, \sigma_\alpha),$$

where $G(x \mid \gamma_0, \sigma_\alpha)$ denotes the generalized Pareto distribution function with shape parameter γ_0 and scale parameter σ_α.

Let $T_1, \ldots, T_n$ be independent lifetime random variables with a common survival function S_0 satisfying Assumption 2.1. Let $C_1, \ldots, C_n$ be independent censoring random variables, independent of the T_i's, with a possibly non-continuous and possibly defective com-

mon survival function S_C (that is, we may have $\lim_{x\to\infty} S_C(x) > 0$). We observe the censored data $(X_1, \delta_1), \ldots, (X_n, \delta_n)$, where

$$X_i = \min\{T_i, C_i\} = \delta_i T_i + (1 - \delta_i) C_i, \quad \delta_i = \mathbb{1}[T_i \leq C_i].$$

Let $S = S_0 \cdot S_C$ denote the survival distribution of X_i.

Let $X_{1:n} \leq \ldots \leq X_{n:n}$ denote the order statistics of $X_1, \ldots, X_n$. Denote by $\delta_{i,n}$ the induced order statistics of $\delta_1, \cdots, \delta_n$ associated with $X_{i:n}$, such that $(X_{i:n}, \delta_{i,n}) \in \{(X_i, \delta_i) : 1 \leq i \leq n\}$. The Kaplan and Meier (1958) estimator of the lifetime survival function S_0 is given by

$$\widehat{S}_0(t) = \prod_{X_{i:n} \leq t} \left(\frac{n-i}{n-i+1} \right)^{\delta_{i,n}}. \tag{4}$$

Note that $\widehat{S}_0(X_{n:n}) = 0$ if $X_{n:n}$ is not censored, but $\widehat{S}_0(X_{n:n}) > 0$ otherwise. In the latter case, one may redefine $\widehat{S}_0(x) = 0$ for $x > X_{n:n}$ if necessary.

Under Assumption 2.1, we can refine this estimator using the generalized Pareto model. Let us choose a sufficiently large threshold, denoted by $u_n < X_{n:n}$, with an estimated exceeding probability

$$\widehat{\alpha}_n = \widehat{S}_0(u_n). \tag{5}$$

Then, we can approximate $S_0(t)$ for $t > u_n$ by

$$\widehat{S}_0(t \mid \gamma, \log\sigma) = \widehat{\alpha}_n G\left(t - u_n \mid \gamma, \log\sigma\right), \tag{6}$$

where G denotes the generalized Pareto survival function defined in Assumption 2.1 with appropriate parameters $(\gamma, \log\sigma)$ such that $1 + \frac{\gamma(X_{n:n} - u_n)}{\sigma} > 0$. One advantage of the generalized Pareto model is its ability to extrapolate beyond the data range, even when the data maximum is below the endpoint of the true distribution support due to the censoring mechanism.

To use (6), we estimate the parameters by the maximum likelihood method for censored data. Given an exceedance $X - u_n = x > 0$ and the associated censoring indicator δ, the censored log-likelihood function (for $1 + \gamma x/\sigma > 0$) is given by

$$\begin{aligned}\ell(\gamma, \log\sigma | x, \delta) = &\begin{cases} \log\left(-G'(x|\gamma,\sigma)\right) & \delta = 1 \\ \log G(x|\gamma,\sigma) & \delta = 0 \end{cases} \\ =&\delta \log(-G'(x|\gamma,\sigma)) + (1-\delta)\log G(x|\gamma,\sigma) \\ =&\delta \log\lambda(x|\gamma,\sigma) + \log G(x|\gamma,\sigma),\end{aligned}$$

where

$$\lambda(x|\gamma,\sigma) = -\frac{G'(x|\gamma,\sigma)}{G(x|\gamma,\sigma)} = \left(1+\frac{\gamma}{\sigma}x\right)^{-1}\frac{1}{\sigma}$$

is the hazard rate for the generalized Pareto model. Now, under Assumption 2.1, we can compute the log-likelihood function explicitly given by

$$\ell(\gamma, \log\sigma | x, \delta) = -\delta\left\{\log\left(1+\frac{\gamma}{\sigma}x\right) + \log\sigma\right\} - \frac{1}{\gamma}\log\left(1+\frac{\gamma}{\sigma}x\right),$$

which is well-defined at $\gamma = 0$ by

$$\ell(0, \log\sigma | x, \delta) := \lim_{\gamma\to 0}\ell(\gamma, \log\sigma | x, \delta) = -\delta\log\sigma - \frac{x}{\sigma}.$$

The total log-likelihood for the sample exceedances is therefore given by

$$\sum_{i=1}^{n}\ell(\gamma, \log\sigma | X_i - u_n, \delta_i)\mathbb{1}[X_i > u_n],$$

which is finite if and only if

$$1 + \gamma(X_{n:n} - u_n)/\sigma > 0.$$

When there is no censoring ($\delta_i = 1$ for all i), the likelihood reduces to the standard one for fitting GPD to the data $X_i = T_i$ and one finds relevant asymptotic theory in He et al. (2022).

We assume that our threshold statistic u_n satisfies the following conditions.

Assumption 2.2. The following conditions hold.

(a) The threshold $u_n = u_n(X_1, \ldots, X_n)$ is an arbitrary measurable statistic such that $S_0(u_n) \xrightarrow{p} \bar{\alpha}$ for some $\bar{\alpha} \in (0, \alpha_0)$, where S_0 is the lifetime survival function and '$\xrightarrow{p}$' denotes convergence in probability;

(b) The survival function of C, namely S_C, is Lipschitz continuous and positive in a neighborhood of the limiting threshold $\bar{u} := Q_0(1-\bar{\alpha})$ and has a bounded variation on (u_0, τ), where $\tau = \sup\{x : S(x) > 0\}$ denotes the right endpoint of survival distribution S of censored observation $X_i = \min\{T_i, C_i\}$.

The first condition allows a general threshold statistic, such as any appropriate quantile statistic or any appropriate fixed value. The convergence rate towards the limiting threshold can be arbitrary. The second condition on S_C requires only continuity in a neighborhood around the (limiting) threshold to avoid some irregular thresholding effects. The bounded variation condition is only needed for using the integration by parts in our proofs, and it allows for many discontinuous functions that are discontinuous at a countable set of points.

Using Gertsbakh (1995)'s formula, the unconditional Fisher information matrix

$$\mathcal{I}(\bar{\alpha}) = -\frac{1}{\bar{\alpha}}\mathbb{E}[\nabla^2_{(\gamma,\log\sigma)}\ell(\gamma, \log\sigma \mid X_i - \bar{u}, \delta_i)\mathbb{1}[X > \bar{u}]] \tag{7}$$

associated with the limiting threshold $\bar{u}$ has the integral form

$$\mathcal{I}(\bar{\alpha}) = \mathbb{E}[\mathcal{I}(\gamma_0, \log\sigma_{\bar{\alpha}} \mid Z)\mathbb{1}[Z > 0]], \quad Z = C - \bar{u},$$

where $\mathcal{I}(\gamma, \log\sigma \mid z)$ is the conditional Fisher information matrix given by

$$\begin{aligned}\mathcal{I}(\gamma, \log\sigma \mid z) = \int_0^z & -G'(x \mid \gamma, \sigma)s(\gamma, \log\sigma \mid x)s^\top(\gamma, \log\sigma \mid x)\,dx \\ & + G(z \mid \gamma, \sigma)w(\gamma, \log\sigma \mid z)w^\top(\gamma, \log\sigma \mid z),\end{aligned}$$

where $s(\gamma, \log\sigma \mid x) = \nabla_{(\gamma,\log\sigma)} \log(-G'(x \mid \gamma, \sigma))$ denotes the lifetime score functions for generalized Pareto distributions with respect to $(\gamma, \log\sigma)^\top$, $w(\gamma, \log\sigma \mid x) = \nabla_{(\gamma,\log\sigma)} \log G(x \mid \gamma, \sigma)$ represents the censoring score function, and $A^\top$ denotes the transpose of vector or matrix A. Note that $\mathcal{I}(\gamma, \log\sigma \mid z)$ is positive definite for every $z > 0$, and $\mathcal{I}(\bar{\alpha})$ is also positive definite by part (b) of Assumption 2.2.

The following theorem establishes the existence of the maximum likelihood estimators (MLEs) for the generalized Pareto parameters and their joint asymptotic normality with the Kaplan-Meier (KM) estimator. Let $\widetilde{S}(x) = \mathbb{P}(X_i > x, \delta_i = 1)$ denote the (improper) survival distribution when the observation is not censored. Denote weak convergence by '$\xrightarrow{w}$', and let $D(I)$ represent the (generalized) Skorokhod space of functions on the interval I that may have jump discontinuities; see, e.g., Billingsley (1999), Section 12.

Theorem 2.1. *Suppose that Assumptions 2.1 and 2.2 hold with a true parameter $\gamma_0 > -\frac{1}{2}$.*

(1) With probability tending to 1, there exists a maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_n = (\widehat{\gamma}, \log\widehat{\sigma})^\top$ in the local parameter space

$$\Theta_n^\varepsilon = \left\{ \boldsymbol{\theta} \in \mathbb{R}^2 : \left\| \boldsymbol{\theta} - \boldsymbol{\theta}_0^{(n)} \right\| < n^{-1/2+\varepsilon} \right\}, \tag{8}$$

for any $\varepsilon \in (0, \min\{\gamma_0 + 1/2, 1/2\})$, where $\boldsymbol{\theta}_0^{(n)} = (\gamma_0, \log\sigma_{\alpha_n})^\top$ denotes the true values adaptive to the threshold statistic u_n.

(2) Any maximum likelihood estimator sequence from part (i) is asymptotically normal jointly with the product-limit process $\{\widehat{S}_0(t) : t \le \tau\}$ for any point τ with $S(\tau) > 0$ in such a way that, in the product space $D([0, Q(1-\tau)]) \times \mathbb{R}^2$

$$\left(\sqrt{n} \left(\frac{\widehat{S}_0(\cdot)}{S_0(\cdot)} - 1 \right), \sqrt{n\bar{\alpha}} \left(\widehat{\gamma} - \gamma_0 \right), \sqrt{n\bar{\alpha}} \left(\frac{\widehat{\sigma}}{\sigma_{\alpha_n}} - 1 \right) \right) \xrightarrow{w} (Z(S(\cdot)), \Gamma, \Lambda)$$

where the process Z, Γ, and Λ are jointly Gaussian and related through Brownian

bridges B_1 and B_2 that are jointly Gaussian with cross-covariance structure

$$cov(B_1(s), B_2(t)) = \min\{\widetilde{\rho}(s), t\} - st,\ \widetilde{\rho}(s) = \mathbb{P}(S(X_i) < s, \delta_i = 1),$$

in the following way:

(i) $Z(s) = \int_s^1 B_1(x)x^{-2}d\widetilde{\rho}(x) - \int_s^1 x^{-1}dB_2(\widetilde{\rho}(x))$

(ii) $(\Gamma, \Lambda)^\top = [\mathcal{I}(\bar{\alpha})]^{-1}\Upsilon$, *where the Fisher information matrix* $\mathcal{I}(\bar{\alpha})$ *defined in* (7) *is positive definite, and the random vector* Υ *is given by*

$$\Upsilon = -\begin{bmatrix} 0 \\ B_2(\widetilde{S}(Q_0(1-\bar{\alpha}))) \end{bmatrix} + \int_0^1 B_2(\widetilde{S}(Q_0(1-\bar{\alpha}t)))\upsilon_{1,\gamma_0}(t)dt$$
$$- \int_0^1 B_1(S(Q_0(1-\bar{\alpha}t)))\upsilon_{2,\gamma_0}(t)dt$$

with the vector functions

$$\upsilon_{1,\gamma_0}(t) = \begin{bmatrix} -t^{\gamma_0-1} \\ \gamma_0 t^{\gamma_0-1} \end{bmatrix}, \quad \textit{and} \quad \upsilon_{2,\gamma_0}(t) = \begin{bmatrix} t^{-1}\left(\frac{t^{\gamma_0}-1}{\gamma_0}\right) \\ -t^{\gamma_0-1} \end{bmatrix}.$$

When $\gamma_0 = 0$, we interpret the first entry of $\upsilon_{2,\gamma_0}(t)$ as its continuous extension given by $t^{-1}\log t$.

Recall our threshold statistic u_n and take any upper bound $\tau > \bar{u}$ such that $\mathbb{P}(u_n < \tau) \to 1$. With the Kaplan-Meier estimator $\widehat{S}_0(t), t \in [0, u_n]$, and the maximum likelihood estimator $\widehat{\gamma}$, and $\widehat{\sigma}$, we propose the following refined estimator of the lifetime survival function on the whole positive line,

$$\widehat{S}_0(t; u_n) = \begin{cases} \widehat{S}_0(t) & t \le u_n \\ \widehat{S}_0(u_n)G(t-u_n|\widehat{\gamma}, \widehat{\sigma}) & t > u_n \end{cases},$$

where $G(\cdot|\gamma,\sigma)$ is the generalized Pareto survival function defined in Assumption 2.1. Accordingly, we can estimate the mean of T, for $\widehat{\gamma}<1$,

$$\widehat{\mu}_T = \int_0^{\infty} \widehat{S}_0(t;u_n)dt \tag{9}$$

$$= \int_0^{u_n} \widehat{S}_0(t)dt + \widehat{S}_0(u_n) \times \frac{\widehat{\sigma}}{1-\widehat{\gamma}}. \tag{10}$$

Observe that, when $u_n > u_0$, we can also expand the true survival function beyond u_n and the mean and variance of T adaptively as follows,

$$S_0(t) = S_0(u_n)G(t-u_n|\gamma_0,\sigma_{\alpha_n}),$$
$$\mu_T = \int_0^{\infty} S_0(t)dt = \int_0^{u_n} S_0(t)dt + S_0(u_n)\frac{\sigma_{\alpha_n}}{1-\gamma_0}.$$

The corollary below follows directly from the delta method.

Corollary 2.1. *Under the conditions of Theorem 2.1,*

(1) $\sqrt{n}\left(\widehat{S}_0(\cdot;u_n) - S_0(\cdot)\right) \xrightarrow{w} \mathbb{S}(\cdot)$ *in* $D([0,\infty))$*, where*

$$\mathbb{S}(t) = S_0(t)Z(S(\min\{t,\bar{u}\})) + \bar{\alpha}(\Gamma,\Lambda)^{\top}\nabla_{(\gamma,\log\sigma)}G((t-\bar{u})_+|\gamma_0,\sigma_{\bar{\alpha}}),$$

where

$$\nabla_{(\gamma,\log\sigma)}G(x|\gamma_0,\sigma_{\bar{\alpha}}) = G(x|\gamma_0,\sigma_{\bar{\alpha}}) \begin{bmatrix} \frac{1}{\gamma_0^2}\left\{\log\left(1+\frac{\gamma_0 x}{\sigma_{\bar{\alpha}}}\right) - \frac{\gamma_0 x/\sigma_{\bar{\alpha}}}{1+\gamma_0 x/\sigma_{\bar{\alpha}}}\right\} \\ \frac{x/\sigma_{\bar{\alpha}}}{1+\gamma_0 x/\sigma_{\bar{\alpha}}}, \end{bmatrix}$$

and, when $\gamma_0 = 0$*, first entry on the right-hand-side should be interpreted as it continuous extension given by* $G(x|0,\bar{\sigma})\cdot\frac{1}{2}\left(\frac{x}{\sigma}\right)^2$.

(2) If provided that $\gamma_0 < 1$*,*

$$\sqrt{n}\left(\widehat{\mu}_T - \mu_T\right) \xrightarrow{d} \mathbb{M}$$

where '$\xrightarrow{d}$*' denotes convergence in distribution, and*

$$\mathbb{M} = \int_0^{\bar{u}} Z(S(t))dt + \bar{\alpha}\frac{\sigma_{\bar{\alpha}}}{1-\gamma_0}\left\{Z(S(\bar{u})) + \frac{1}{1-\gamma_0}\Gamma + \Lambda\right\}.$$

Using Vervaat (1972) Lemma (see also Appendix A in de Haan and Ferreira, 2006), under the Skorokhod construction, the weak convergence of the refined estimator of the survival function also implies the weak convergence of the refined quantile estimator under mild differentiability conditions. Consider the refined estimator of the lifetime quantile function

$$\widehat{Q}_0(1-p;u_n) = \begin{cases} \widehat{S}_0^{\leftarrow}(1-p) & p \geq \widehat{\alpha}_n, \\ \\ u_n + Q_G(1-p/\widehat{\alpha}_n;\widehat{\gamma},\widehat{\sigma}) & p < \widehat{\alpha}_n, \end{cases}$$

where $\widehat{\alpha}_n = \widehat{S}_0(u_n)$, '$\leftarrow$' denotes the left-continuous inverse, and $Q_G(\cdot;\gamma,\sigma)$ is the quantile function of $G(\cdot \mid \gamma,\sigma)$ given by

$$Q_G(1-p;\gamma,\sigma) = \begin{cases} \frac{\sigma}{\gamma}(p^{-\gamma}-1) & \gamma \neq 0, \\ \\ \sigma \log(1/p) & \gamma = 0. \end{cases}$$

In particular, when $\gamma_0 < 0$, we can estimate the end-point τ_0 by

$$\widehat{\tau}_0 = Q_G(1;\widehat{\gamma},\widehat{\sigma}) = u_n - \frac{\widehat{\sigma}}{\widehat{\gamma}},$$

where the second equation holds with probability tending to one due to the consistency of $\widehat{\gamma}$.

The next corollary gives the weak convergence of our refined (high) quantile estimators.

Corollary 2.2. *Under the conditions of Theorem 2.1, for any compact interval $I \subset (0,1)$ on which S_0 is strictly decreasing and continuously differentiable,*

$$\sqrt{n}\left(\widehat{Q}_0(\cdot;u_n) - Q_0(\cdot)\right) \xrightarrow{w} \mathbb{Q}(\cdot),$$

in $D(I)$ where $\mathbb{Q}(1-p) = \frac{p}{S_0'(Q_0(1-p))}\mathbb{S}(Q_0(1-p))$. In particular, for every high quantile at survival probability level $p \in (0,\alpha_0)$,

$$\frac{\sqrt{n\bar{\alpha}}}{\sigma_p}\left(\widehat{Q}_0(1-p;u_n) - Q_0(1-p)\right) \xrightarrow{d} q\left(\frac{\bar{\alpha}}{p}\right)^{\top}(\Gamma,\Lambda) + Z(S(\bar{u})),$$

where

$$q(t) = \left(\int_1^t \left(\frac{s}{t}\right)^{\gamma_0} \frac{\log s}{s}\, ds,\ \frac{1 - t^{-\gamma_0}}{\gamma_0} \right)^\top, \quad t > 0,$$

and it should be interpreted by continuity as $\left(\frac{1}{2}(\log t)^2, \log t\right)^\top$ *when* $\gamma_0 = 0$. *When* $\gamma_0 < 0$, *the same holds for* $p = 0$ *in the sense that* $\sqrt{n\bar{\alpha}}(\widehat{\tau}_0 - \tau_0) \xrightarrow{d} -\frac{\sigma_{\bar{\alpha}}}{\gamma_0}\left(-\frac{1}{\gamma_0}\Gamma + \Lambda\right)$.

Remark 1. Although we study different statistics separately in Theorem 2.1 and Corollaries 2.1–2.2, their weak convergence holds jointly, with the limiting random elements defined on the same probability space.

Remark 2. To ensure that our estimator of the endpoint is no smaller than the data maximum, one can use $\widehat{\tau}_0 = \max\{u_n - \widehat{\sigma}/\widehat{\gamma}, X_{n:n}\}$ and the same asymptotic holds when $\gamma_0 > -1/2$.

In general, the asymptotic variance above takes a complex form that depends on the unknown functions S_0, $\widetilde{S}$, and S. To facilitate convenient inference and improve finite-sample coverage, we propose constructing confidence intervals using the random weighted bootstrap procedure described below. Consider the true parameter θ_0 from one of the following or its log transformation: μ_T, $S_0(t)$ for some $t \in (0, \tau_0)$, or $Q_0(1-p)$ for some $p \in (0,1)$, as well as the endpoint $\tau_0 = Q_0(1)$ if it exists. Denote the refined estimator by $\widehat{\theta}$. We propose the following interval inference procedure for θ_0:

(Step 1) Draw a random sample of size n from a subexponential distribution with mean one and variance one, such as the standard exponential distribution. Denote the sample as $\xi_1^{(b)}, \ldots, \xi_n^{(b)} > 0$. Let $\xi_{i,n}^{(b)}$ be the induced order statistics of $\{\xi_i^{(b)} : i = 1, \ldots, n\}$, associated with $X_{i:n}$.

(Step 2) Choose a threshold statistic $u_n^{(b)}$ satisfying $u_n^{(b)} = u_n + o_p(1)$, which may depend on $\xi^{(b)} = (\xi_1^{(b)}, \ldots, \xi_n^{(b)})^{\top T}$, or simply as u_n. Solve for the weighted maximum likelihood

estimators $\widehat{\gamma}^{(b)}$ and $\widehat{\sigma}^{(b)}$ by maximizing the following:

$$\sum_{i=1}^{n} \xi_i^{(b)} \ell_i(\gamma, \log\sigma \mid X_i - u_n^{(b)}, \delta_i) \mathbb{1}[X_i > u_n^{(b)}].$$

Additionally, calculate the random weighted Kaplan-Meier (KM) estimator as follows:

$$\widehat{S}_0(t; \xi^{(b)}) = \prod_{X_{i:n} \le t} \left(1 - \frac{\xi_{i,n}^{(b)}}{\sum_{j \ge i} \xi_{j,n}^{(b)}}\right)^{\delta_{i,n}}.$$

Replace $(\widehat{\gamma}, \widehat{\sigma}, \widehat{S}_0(\cdot))$ with $(\widehat{\gamma}^{(b)}, \widehat{\sigma}^{(b)}, \widehat{S}_0(\cdot; \xi^{(b)}))$ to obtain the random weighted estimator $\widehat{\theta}^{(b)}$.

(Step 3) Repeat the above steps B times to obtain the set of estimators $\{\widehat{\theta}^{(b)}\}_{b=1}^{B}$. For a sufficiently large B, this process yields the conditional distribution of $\widehat{\theta}^{(b)}$ given the observations. Based on this, we can construct confidence intervals according to the following theorem.

Theorem 2.2. *Suppose the conditions of Theorem 2.1 hold. Consider the true parameter θ_0 from one of the following, or its log transformation, provided it exists: the lifetime mean μ_T, a survival probability $S_0(t)$ for some $t \in (0, \tau_0)$, a quantile $Q_0(1-p)$ for some $p \in (0,1)$, as well as the endpoint $\tau_0 = Q_0(1)$ if it exists, or the extreme value index γ_0. Then for any bootstrap threshold statistic $u_n^{(b)} = u_n + o_p(1)$ with $\alpha_n^{(b)} := 1 - F(u_n^{(b)})$, the bootstrap procedure is asymptotically valid in the following sense.*

(1) With probability tending to 1, there exists a maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_n^{(b)} = (\widehat{\gamma}, \log\widehat{\sigma})^\top$ in the local parameter space

$$\Theta_{n,b}^{\varepsilon} = \left\{ \boldsymbol{\theta} \in \mathbb{R}^2 : \left\| \boldsymbol{\theta} - \boldsymbol{\theta}_0^{(n,b)} \right\| < n^{-1/2+\varepsilon} \right\}, \tag{11}$$

for any $\varepsilon \in (0, \min\{\gamma_0 + 1/2, 1/2\})$, where $\boldsymbol{\theta}_0^{(n,b)} = \left(\gamma_0, \log\sigma_{\alpha_n^{(b)}}\right)^\top$ denotes the true values adaptive to the bootstrap threshold.

(2) For a one-sided confidence interval at any confidence level $a \in (0, 0.5)$*,*

$$\mathbb{P}\left(\theta_0 \leq \widehat{\theta} + \widetilde{c}_n(a)\right) \to 1 - a, \text{ and } \mathbb{P}\left(\theta_0 \geq \widehat{\theta} - \widetilde{c}_n(a)\right) \to 1 - a,$$

where

$$\widetilde{c}_n(a) = \max\{-q_n(a), q_n(1-a)\},$$
$$q_n(p) \equiv \inf\left\{x : \mathbb{P}\left(\widehat{\theta}^{(b)} - \widehat{\theta} \leq x | (X_1, \delta_1), \ldots, (X_n, \delta_n)\right) \geq p\right\}.$$

(3) For a two-sided confidence interval at any confidence level $a \in (0, 1)$*,*

$$\mathbb{P}\left(\left|\widehat{\theta} - \theta_0\right| \leq c_n(a)\right) \to 1 - a,$$

where

$$c_n(a) = \inf\left\{x : \mathbb{P}\left(\left|\widehat{\theta}^{(b)} - \widehat{\theta}\right| \leq x | (X_1, \delta_1), \ldots, (X_n, \delta_n)\right) \geq 1 - a\right\}. \tag{12}$$

Furthermore, the results extend to $\theta_0 = \sigma_{\alpha_n}$*, as well as its log-transformation, if we maintain the same threshold* $u_n^{(b)} \equiv u_n$ *in the bootstrap samples.*

Remark 3. For the two-sided confidence interval, one can also use a slightly more conservative option with higher coverage, $c_n(a) = \widetilde{c}_n(1 - a/2)$, which is also asymptotically correct.

Observe that the theorem implies that one can construct a consistent estimator of the asymptotic variance of $\widehat{\theta}$ via $c_n(a)$, leading to the following corollary for two-sample tests when combined with the asymptotic normality of $\widehat{\theta}$ from Theorem 2.1 and Corollaries 2.1–2.2.

Corollary 2.3. *Consider two independent samples indexed by* $s \in \{1, 2\}$*, with raw estimators* $\widehat{\theta}_s$ *and randomly weighted estimators* $\widehat{\theta}_s^{(b)}$ *of the true parameter* $\theta_{0,s}$*, respectively, both*

satisfying the conditions in Theorem 2.2. Let $c_{n,s}(a)$ *denote the bootstrap critical value* (12) *for each sample* $s \in \{1, 2\}$ *for some specific* $a \in (0, 1)$*. Suppose we estimate the difference* $\theta_0 \equiv \theta_{0,2} - \theta_{0,1}$ *by* $\widehat{\theta} \equiv \widehat{\theta}_2 - \widehat{\theta}_1$*. Then,*

$$[\widehat{s}_a(\widehat{\theta})]^{-1} \left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, 1),$$

where

$$\widehat{s}_a(\widehat{\theta}) = \sqrt{(c_{n,1}(a))^2 + (c_{n,2}(a))^2}/\Phi^{-1}(1 - a/2), \tag{13}$$

and Φ^{-1} *denotes the inverse of the standard normal distribution function.*

Remark 4. As a byproduct, we obtain that the bragging (bootstrap robust aggregating) estimator

$$\widehat{\theta}_B = \text{Median}\left(\widehat{\theta}^{(b)} \mid (X_1, \delta_1), \ldots, (X_n, \delta_n)\right) \tag{14}$$

is asymptotically indistinguishable from $\widehat{\theta}$, and therefore the results above apply to both. Since it is well known that maximum likelihood estimation for the generalized Pareto distribution is not a globally concave problem and can suffer from local minima, we recommend using this bragging estimator, especially in small samples.

3 Simulation Study

We consider three settings and compare the results for the Kaplan-Meier (KM) estimator and our refined estimator based on the generalized Pareto model for estimating three different sets of parameters for the lifetime distribution: the mean, survival probabilities, and quantiles. The sample size is fixed at $n = 2000$, and the results are reported over 10,000 Monte Carlo replications. For each replication, we generate $B = 500$ bootstrap datasets to construct two-sided confidence intervals using the bragging estimator from Remark 4.

3.1 Unbounded but short lifetime

We revisit the example from the introduction. The lifetime variables, T_i, and censoring variables, C_i, are both exponentially distributed with light tails and an extreme value index of $\gamma_0 = 0$. We are interested in estimating the lifetime mean, $\mu_T = 1$. While the mean of T_i is fixed, we vary the mean of the censoring variable C_i, decreasing from 2, 1.8, ..., 1, corresponding to an increase in the censoring rate from 1/3 to 1/2. We use the 90% sample quantile as our threshold u_n and the confidence intervals of μ_T. We report the results for the confidence intervals constructed based on the log transformation $\theta_0 = \log \mu_T$ to ensure positive estimates. The results are very similar to those obtained using $\theta_0 = \mu_T$ directly.

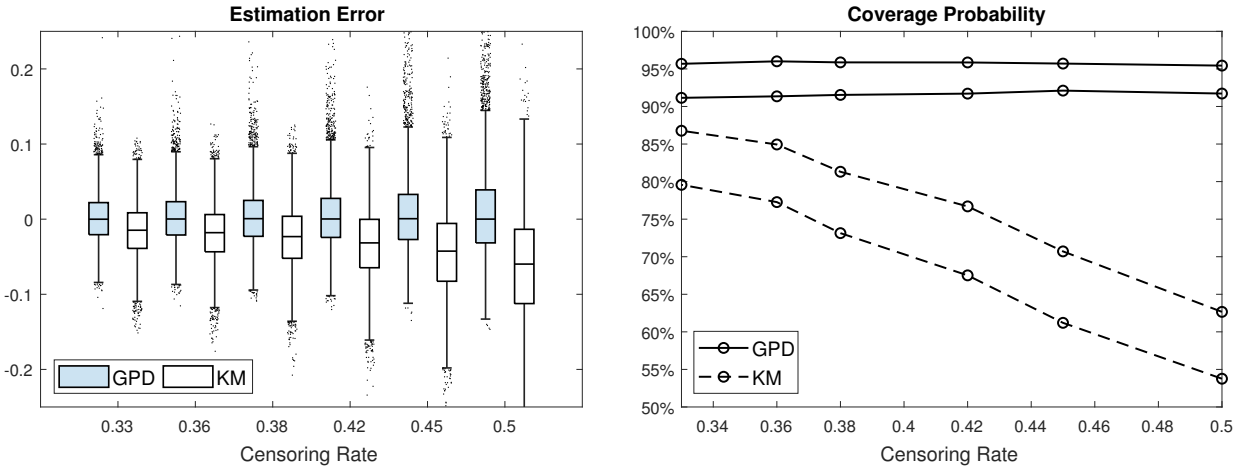

Figure 4: Box plot of estimation errors (left) and the coverage probabilities of the 95% and 90% confidence intervals (right) for the lifetime mean using our refined estimator (GPD) and the Kaplan-Meier (KM) estimator across different censoring rates. Both lifetime and censoring variables follow an exponential distribution.

The box plots of the estimation error in Figure 4 show that the KM estimator suffers from estimation bias. This bias causes the bootstrap confidence intervals to significantly undercover the true lifetime mean. As the censoring rate increases, the likelihood of the largest observations being censored grows (see Figure 2), resulting in a greater bias in the

KM estimator and an almost linear decline in coverage probability, as shown on the right side of Figure 4.

In contrast, our refinement using the generalized Pareto distribution (GPD) nearly eliminates the bias, and the bootstrap confidence intervals maintain relatively stable coverage across different censoring rates, as depicted in Figure 4. The intervals achieve nearly correct coverage at both the 95% and 90% confidence levels.

3.2 Bounded lifetime

We calibrate the lifetime variables T_i from the generalized Pareto distribution with $\gamma_0 = -0.4$, which is well fitted to the Australian AIDS data in Venables and Ripley (2002) by using the sample median as our threshold u_n. We restrict the lifetime distribution to have a bounded support of $(0, 8)$. We also calibrate the censoring variables C_i from a half Cauchy distribution with a median of 0.9 and support on $(0, \infty)$. The censoring rate is approximately 64%. Figure 5 shows that our refined estimator outperforms the KM estimator beyond the threshold, exhibiting a smaller median absolute estimation error and shorter confidence intervals on average while maintaining nearly correct coverage levels of 95% and 90% at each quantile level when applying the random weighted bootstrap method. Our generalized Pareto estimator loses slight coverage probability in very high quantiles close to the endpoint, primarily due to a small probability in finite samples for the extreme value index estimator $\widehat{\gamma}$ being close to 0. This issue is well-documented in extreme value theory even without censoring(see, e.g., Li and Peng, 2012) and typically diminishes as the sample size increases, owing to the consistency of $\widehat{\gamma}$. However, addressing this finite-sample problem is beyond the scope of this paper.

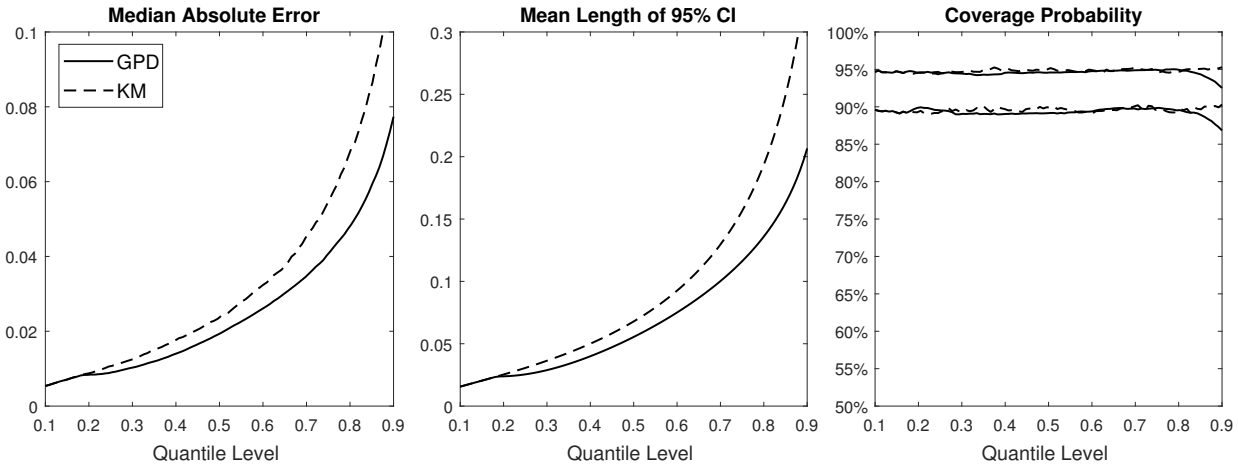

Figure 5: Comparison of the performance between the Kaplan-Meier (KM) estimator and our refined estimator (GPD) for survival probabilities on a log scale, i.e., $\log S_0(t)$, across different quantiles t. From left to right are the median absolute error, average length of the 95% confidence intervals, and coverage probabilities of the 95% and 90% bootstrap confidence intervals. The lifetime is generalized Pareto with endpoint $\tau_0 = 8$, and the censoring variable is half Cauchy.

3.3 Unbounded long lifetime censored by bounded variable

We calibrate the quantiles of the lifetime variable T_i from our empirical application to unemployment duration data in the next section, using the generalized Pareto distribution with $\gamma_0 = 0.6$, unit scale, and full support on $(0, \infty)$. This distribution has a finite mean of approximately 2.5 but an infinite variance. The censoring variables are defined as $C_i = 3B_i$, where B_i follows a Beta$(4, 1)$ distribution, yielding a censoring rate of approximately 24%. We fix the threshold to be $u_n = 0.5$.

In this setting, unlike the previous examples, the outcome $X_i = \min\{T_i, C_i\}$ is bounded by $\tau = 3$ due to censoring, while the lifetime variable is unbounded. To make the KM estimator feasible for quantiles beyond the endpoint $\tau = 3$, we extrapolate linearly beyond the data range. The generalized Pareto estimator, on the other hand, naturally extrapolates

and requires no special adjustment.

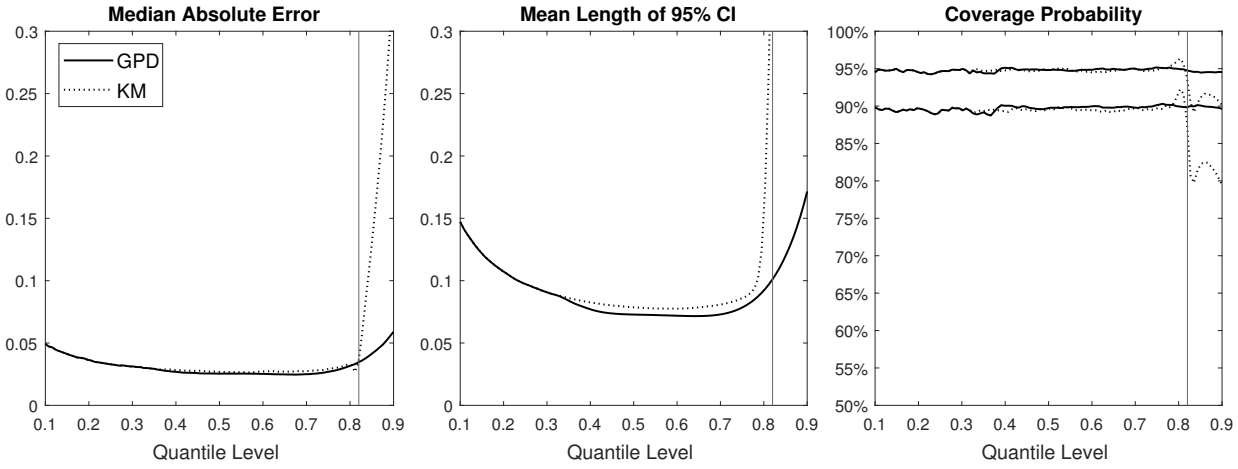

Figure 6: Comparison of the performance between the Kaplan-Meier (KM) estimator and our refined estimator (GPD) for quantiles on a log scale, i.e., $\log Q_0(p)$, across different quantile levels p from 0.1 to 0.9. From left to right are the median absolute error, average length of the 95% confidence intervals, and coverage probabilities of the 95% and 90% bootstrap confidence intervals. The vertical line indicates the quantile level of endpoint $\tau = 3$ with respect to the lifetime distribution.

Figure 6 shows that our refined estimator performs similarly to the KM estimator in the central region but significantly outperforms it around and beyond the endpoint, showing smaller median estimation error and shorter confidence intervals on average. The KM estimator performs well until it approaches the outcome endpoint, approximately the 0.82 quantile of the lifetime distribution, after which its estimation error increases dramatically. In contrast, the extrapolation based on the fitted generalized Pareto distribution provides a smooth transition beyond the outcome endpoint, resulting in a substantially smaller median absolute error and shorter confidence intervals on average. Moreover, the bootstrap confidence intervals based on our refined estimator maintain almost correct coverage, while those based on the KM estimator tend to undercover, even when they are much wider,

beyond the data range.

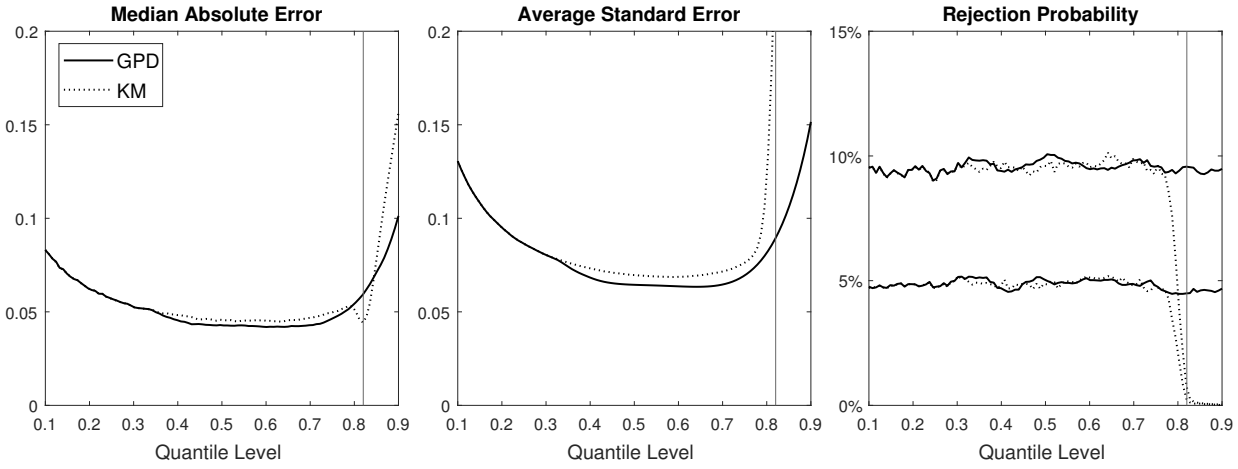

Figure 7: Comparison of the performance between the Kaplan-Meier (KM) estimator and our refined estimator (GPD) for the difference in quantiles, on a log scale, over two independent samples. From left to right are the median absolute estimation error, average of the estimated standard error, and rejection probabilities of the two-sided t-test for equal quantiles at the 5% and 10% significance levels. The vertical line indicates the quantile level of the endpoint $\tau = 3$ with respect to the lifetime distribution.

Next, we show that our bootstrap procedure can be used for a two-sample test according to Corollary 2.3. We generate another independent sample from the same data-generating process but with half the sample size, that is, a sample size of 1000. Then, we estimate the difference in their quantiles on a log scale and evaluate the rejection probability of the two-sided t-test of quantile equality at significance levels of 5% and 10%, using the bootstrapped standard error (13) with $a = 0.1$ (the results for $a = 0.05$ are similar and hence omitted).

Again, our refined estimator outperforms the KM estimator almost everywhere above the threshold, especially beyond the endpoint, as shown in Figure 7. The KM estimator suffers from large estimation errors in the tail, resulting in nearly zero rejection rates around

and beyond the endpoint. In contrast, our refined estimator maintains almost correct size across most quantile levels.

4 Empirical Illustration

We illustrate the advantage of our refined estimator by analyzing a US dataset from the national job training study initially analyzed by Bloom et al. (1997). This study is a large-scale randomized experiment evaluating the programs funded by the Job Training Partnership Act (JTPA) of 1982. Each individual was randomly assigned to either the control or treatment group. About 2/3 of the individuals were assigned to the treatment group, so they were allowed to enroll in a JTPA-funded training program. The rest, 1/3 of the individuals assigned to the control group, were generally excluded from receiving JTPA services for 18 months, though they may still participate in another training program. Following Abadie, Angrist, and Imbens (2002), we split the analysis for females and males. Furthermore, we split the sample into two subsamples according to the random assignment.

We are interested in the duration T_i between the treatment assignment and finding employment as in Frandsen (2015) and Ba et al. (2017). However, this duration is censored by the duration C_i between the random assignment and the follow-up interview. Since the follow-up surveys were scheduled relative to the treatment assignment date rather than a fixed calendar date, we consider these censoring events to be independent, as argued by Frandsen (2015).

We obtained the cleaned dataset directly from the online supplement of Beyhum, Tedesco, and Van Keilegom (2024), which consists of 7,876 individuals with no missing values in their analysis. The censoring rate is similar across the control and treatment groups, with approximately 15% for males and 24% for females. To convert the duration

data from days to years, we divide the duration in days by 365.

We then fit censored generalized Pareto models to the long-term unemployment duration data by taking the conventional threshold as 27 weeks (approximately half a year) to all groups divided by treatment assignment and sex. Figure 8 demonstrates the good fit of the generalized Pareto model to the quantile function obtained from the KM estimator across all groups. The quantile function for the treatment group is almost everywhere lower than the control group for the females, while the quantile function for the male treatment group is similar to the male control group at the lower quantile levels. This aligns with Abadie, Angrist, and Imbens (2002)'s observation that quantile treatment effects for males are only significant at higher quantiles.

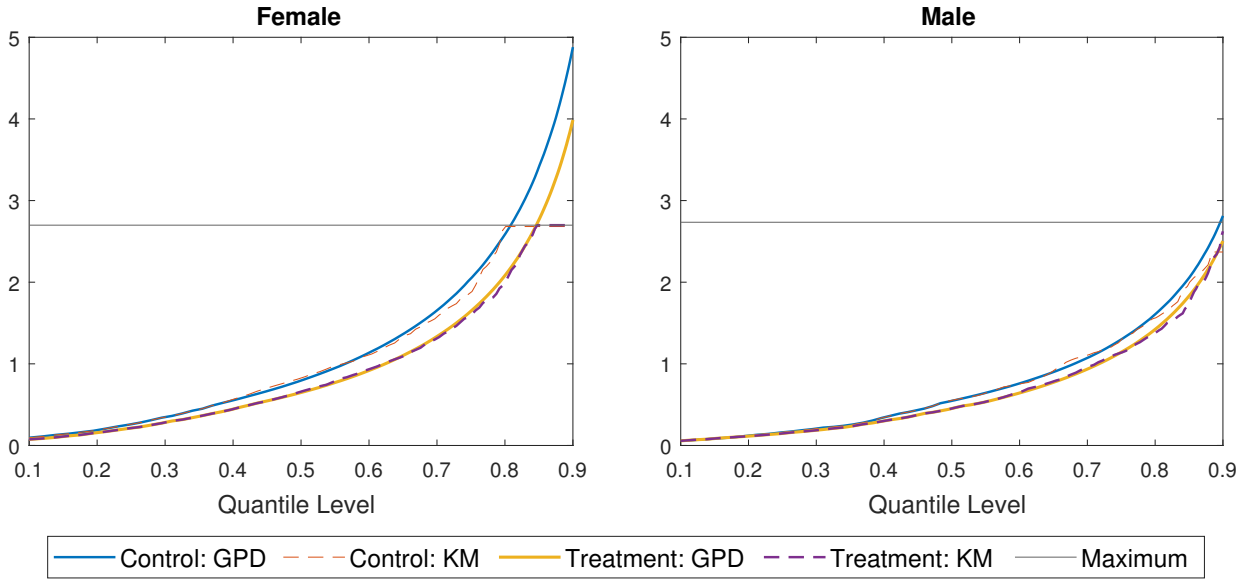

Figure 8: Quantile function based on the KM estimator and our refined estimator (GPD). The dotted line shows the threshold, and the dashed line shows the maximum of the uncensored duration over both control and treatment groups.

The KM estimator of the quantile function is truncated at the maximum uncensored duration within each group, as the largest observations are censored. This truncation

results in a substantial underestimation of the duration mean compared to our refined estimator, as shown in Table 1. We also report the extreme value index $\widehat{\gamma}$ and the scale parameter $\widehat{\sigma}$ for the fitted generalized Pareto model. Our refined estimator, benefiting from extrapolation, yields values more than four times higher than the KM estimator for females and more than twice as high for males. This large difference arises because the fitted tail of the lifetime distribution is too heavy to maintain a finite variance, with an extreme value index $\widehat{\gamma} > 1/2$ for all groups except the male treatment group (for which we could not reject this null hypothesis either).

The fitted generalized Pareto parameters are similar between the treatment and control groups, suggesting that the hazard rate beyond the threshold is comparable. In fact, in an unreported analysis, we cannot reject the equality of these parameters using the bootstrap confidence interval derived from Corollary 2.3. This suggests that the conditional training effect becomes insignificant given that a female or male individual has already been unemployed for more than 27 weeks. This finding is consistent with Ba et al. (2017).

Lifetime Mean $\widehat{\mu}_T$				
	Female, Control	Female, Treatment	Male, Control	Male, Treatment
KM	0.59	0.59	0.55	0.53
GPD	2.66	2.39	1.31	1.12
Generalized Pareto Parameters				
	Female, Control	Female, Treatment	Male, Control	Male, Treatment
$\widehat{\gamma}$	0.69	0.71	0.51	0.44
$\widehat{\sigma}$	1.20	0.98	0.94	0.95

Table 1: Comparison of the KM estimator and GPD estimator of the lifetime mean.

On the other hand, we find unconditional effects at various quantile levels. To test for

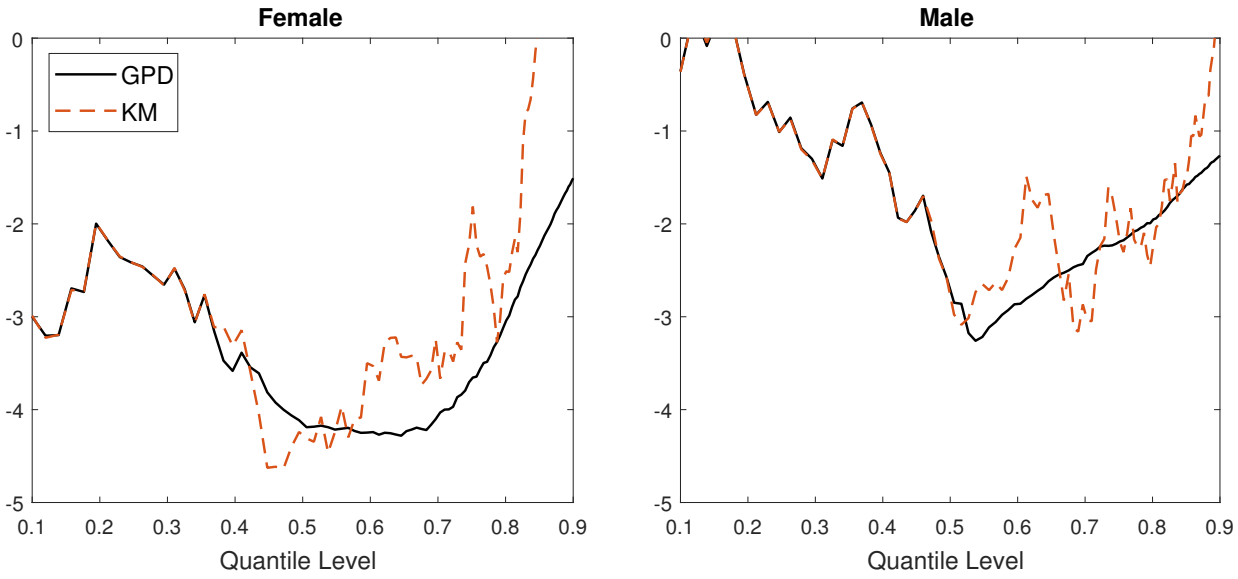

Figure 9: The pointwise t-statistic for testing the equality of quantiles between the control and treatment groups across various quantile levels.

the equality of (unconditional) quantiles between the control and treatment groups, we present in Figure 9 the pointwise t-statistic, constructed using the bootstrapped standard error from Corollary 2.3 with $a = 0.1$ (the results for $a = 0.05$ are nearly identical and are omitted). Like in our simulation study, the KM estimator suffers from size issues and yields only small t-statistics at high quantile levels. In contrast, our refined estimator consistently produces large negative t-statistics for females across every quantile level, while for males, the t-statistics remain small at the lower quantiles.

5 Concluding Remarks

The Kaplan-Meier (KM) estimator is widely used in survival analysis to estimate the distribution function under censoring, but it is reliable only within the observed data range. When the largest observation(s) are censored, the KM estimator becomes improper, leading to significant bias in estimating tail information beyond the range, as well as in estimating

the mean of the lifetime distribution if the tail is not negligible. Motivated by extreme value theory, we propose a refined method by fitting a generalized Pareto distribution (GPD) beyond a finite but sufficiently high threshold. While we find the GPD model highly effective in practice, developing a distribution-free goodness-of-fit test remains an important avenue for future research.

Our approach combines the non-parametric robustness of the KM estimator in the central region with the parametric efficiency of the GPD in the tail. This semiparametric method allows for reliable extrapolation beyond the data range. We have developed a comprehensive theory establishing the asymptotic normality of our refined estimator, along with the asymptotic validity of an easy-to-implement random weighted bootstrap method for inference. This bootstrap procedure performs well in finite-sample settings, as demonstrated by both simulation studies and empirical applications.

Additionally, we provided a straightforward extension of our approach for a two-sample test, which can be useful for causal analysis, as illustrated in our empirical application to a large-scale job training study. Extending this approach to a regression framework could yield more insightful analysis and is an interesting direction for future work.

SUPPLEMENTARY MATERIAL

The supplementary material includes the proofs for all the theorems from Section 2.

References

Abadie, A., Angrist, J., and Imbens, G. (2002), "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70(1), 91–117.

Ba, B. A., Ham, J. C., LaLonde, R. J., and Li, X. (2017), "Estimating (Easily Interpreted) Dynamic Training Effects from Experimental Data," *Journal of Labor Economics*, 35(S1), S149–S200.

Balkema, A., and de Haan, L. (1974), "Residual Life Time at Great Age," *The Annals of Probability*, 2, 792–804.

Beirlant, J., Guillou, A., and Toulemonde, G. (2010), "Peaks-Over-Threshold Modeling under Random Censoring," *Communications in Statistics—Theory and Methods*, 39(7), 1158–1179.

Beyhum, J., Tedesco, L., and Van Keilegom, I. (2024), "Instrumental Variable Quantile Regression under Random Right Censoring," *The Econometrics Journal*, 27(1), 21–36.

Billingsley, P. (1999), *Convergence of Probability Measures*, 2nd ed., New York: Wiley.

Bloom, H. S., Orr, L. L., Bell, S. H., Cave, G., Doolittle, F., Lin, W., and Bos, J. M. (1997), "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study," *Journal of Human Resources*, 32(3), 549–576.

de Haan, L., and Ferreira, A. (2006), *Extreme Value Theory: An Introduction*, New York: Springer.

Efron, B. (1967), "The Two Sample Problem with Censored Data," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 4, 831–853.

Einmahl, J. H. J., Fils-Villetard, A., and Guillou, A. (2008), "Statistics of Extremes under Random Censoring," *Bernoulli*, 14(1), 207–227.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997), *Modelling Extremal Events for Insurance and Finance*, Berlin: Springer.

Frandsen, B. R. (2015), "Treatment Effects with Censoring and Endogeneity," *Journal of the American Statistical Association*, 110(512), 1745–1752.

Gertsbakh, I. (1995), "On the Fisher Information in Type-I Censored and Quantal Response Data," *Statistics & Probability Letters*, 23, 297–306.

He, Y., Peng, L., Zhang, D., and Zhao, Z. (2022), "Risk Analysis via Generalized Pareto Distributions," *Journal of Business & Economic Statistics*, 40(2), 852–867.

Jin, Z., Lin, D. Y., and Ying, Z. (2006), "On Least-Squares Regression with Censored Data," *Biometrika*, 93(1), 147–161.

Kaplan, E. L., and Meier, P. (1958), "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, 53(282), 457–481.

Li, Z., and Peng, L. (2012), "Bootstrapping Endpoint," *Sankhyā A*, 74, 126–140.

Maller, R. A., and Zhou, S. (1993), "The Probability that the Largest Observation is Censored," *Journal of Applied Probability*, 30(3), 602–615.

Resnick, S. I. (1987), *Extreme Values, Regular Variation and Point Processes*, New York: Springer.

Venables, W. N., and Ripley, B. D. (2002), *Modern Applied Statistics with S*, 4th ed., New York: Springer.

Vervaat, W. (1972), "Functional Central Limit Theorems for Processes with Positive Drift and Their Inverses," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 23(4), 245–253.