Nonparametric Threshold Detection for Cost Distributions

Benedikt Funke* Masayuki Hirukawa†
TH Köln Ryukoku University

4th April 2024

Abstract

The tail of a distribution beyond some threshold is of importance and interest in academics and industries. In this paper, we develop a nonparametric threshold detection method designed specially for cost distributions. Our approach originates from tailoring the existing techniques for change point or jump location detection in statistics to stylized facts of cost distributions. Considering that the threshold is located in the right-tail region, we propose to employ the asymmetric gamma kernel when constructing the diagnostic function. It is demonstrated that our jump location estimator is consistent with a faster convergence rate than the parametric one and asymptotically normal when suitably implemented. Since this estimator tends to underestimate the jump location in finite samples, we advocate correcting its bias. Monte Carlo simulations and real data examples illustrate attractive properties and practical relevance of our proposal in several different use cases.

Keywords: distribution of costs; gamma kernel; heavy-tailedness; incomplete gamma function; cross validation; threshold detection.
JEL Classification Codes: C13; C14; G22.
MSC 2010 Codes: 62G07; 62G20; 62P05; 91B30.

* *Corresponding Author*; Institute for Insurance Studies, TH Köln - University of Applied Sciences, Gustav-Heinemann-Ufer 54, 50968 Köln, Germany; e-mail: benedikt.funke@th-koeln.de.

† Faculty of Economics, Ryukoku University, 67 Tsukamoto-cho, Fukakusa, Fushimi-ku, Kyoto 612-8577, Japan; e-mail: hirukawa@econ.ryukoku.ac.jp.

1 Introduction

The *tail* of a distribution beyond some threshold gathers special interest in academics and in several industrial sectors. It is a source of large market fluctuations or even volatility clustering in finance. The region in an income distribution corresponds to the 'elite income' tier. In the field of non-life insurance, a handful of severe losses within a collection of policies beyond some threshold can significantly contribute to the total claim amount. In geophysics and hydrology, tail analyses are conducted from the viewpoint of catastrophic disasters.

At the same time, because distributions in these examples behave quite differently below and above a threshold, it is also widely recognized that a single model cannot capture characteristics over the while ranges of such distributions. In fact, the tail part of an income distribution above a threshold is modelled and studied separately from the part that is below the threshold and contains most of observations (called the *bulk* part hereinafter); see Cowell, Ferreira and Litchfield (1998) and Jenkins (2017), for instance. Moreover, modelling the whole range of a loss distribution is of particular importance and interest in actuarial science, particularly in non-life insurance. Then, splicing (e.g., Klugman et al., 2019) or composite modelling (e.g., Cooray and Ananda, 2005; Scollnik and Sun, 2012) -- the practice of modelling the parts below and above a threshold differently -- is frequently employed.

Several research questions arise in this context. For example, where does the tail part begin in an income distribution? Or where should we switch to the tail model in a loss distribution? The aim of this paper is to propose a nonparametric method of finding a threshold in these distributions. Incomes and actuarial losses, as well as wages and consumption expenditures, are examples of cost variables. Our estimation strategy is grounded on prior knowledge (or stylized facts) about shapes of distributions of such variables. Distributions of cost variables have support on $\mathbb{R}_+$ with a natural boundary at the origin and are highly right-skewed with a concentration of observations in the vicinity of the origin and a long tail with sparse data. As revealed shortly, we employ a specific kernel to capture these characteristics while avoiding possible misspecification by relying on a particular parametric model.

In addition, we translate the problem of threshold estimation into that of change point (or jump location) detection, which has been actively studied in statistics. Our translation or interpretation of the problem is motivated by the original idea of splicing. In this idea, two different probability density functions (pdfs) are fitted for the regions below and above a threshold or a splicing point, and continuity of the distribution at this point is not required.[1] Accordingly, the words "threshold" and "jump" are used exchangeably hereinafter, whenever no confusion may arise.

Following the literature on change point detection (e.g., Chu and Cheng, 1996; Couallier, 1999; Huh, 2002), we take the absolute difference of two kernel density estimates as the diagnostic function and define its maximizer as the jump location estimator.[2] Nonetheless, our procedure remarkably differs from previous ones. While they are designed to work in the middle part (i.e., near the peak) of a distribution, a threshold that divides the bulk and tail parts is supposed to be located on the right tail and far away from the peak of the distribution. To overcome such difficulty while exploiting prior information on shapes of cost distributions, we adopt the asymmetric gamma kernel by Chen (2000); our contribution is the first work in which an asymmetric kernel is applied for the problem of jump detection in density or regression curves, to the best of our knowledge. The kernel is defined as the pdf of the gamma distribution $G(x/b+1, b)$ for a design point $x\,(\geq 0)$ and a smoothing parameter $b\,(>0)$, and it is in the form of

$$K_{G(x,b)}(u) = \frac{u^{x/b}\exp(-u/b)}{b^{x/b+1}\Gamma(x/b+1)}\mathbf{1}\{u \geq 0\}, \tag{1}$$

where $\Gamma(\cdot)$ is the gamma function, and $\mathbf{1}\{\cdot\}$ is an indicator function. This kernel has an adaptive smoothing property via changing shapes automatically across design points, and a single smoothing parameter suffices to generate a variety of shapes. Thanks to this property, the gamma kernel can calibrate right-skewed densities with support on $\mathbb{R}_+$ (e.g., those of cost distributions) well.

Our jump location estimator is super-consistent (i.e., its convergence rate exceeds $\sqrt{n}$, where n is the sample size) and asymptotically normal when suitably implemented. While the estimator shares its asymptotic properties with the existing literature (e.g., Chu and Cheng, 1996; Couallier, 1999; Huh, 2002), our proof strategy is

[1] Many authors (e.g., Cooray and Ananda, 2005; Scollnik and Sun, 2012) additionally impose differentiability of the pdf at the splicing point to make the entire density smooth and reduce the number of parameters. This practice has computational advantage now that the splicing point can be expressed as a function of other model parameters. However, such a restriction on parameters results in less flexibility. From this viewpoint, Reynkens et al. (2017) first estimate the splicing point and then compute estimates of remaining model parameters.

[2] An alternative approach is proposed by Desmet et al. (2010), who transform kernel density estimation into kernel regression estimation via prebinning and applying an existing method of discontinuity detection for nonparametric regression curves. This approach has the disadvantage that prebinning substantially reduces the sample size which can be used for regression estimation.

totally different. The gamma kernel does not admit the location-scale transformation $K_h(u-x) = K\{(u-x)/h\}/h$ with a bandwidth $h\,(>0)$ or exchangeability between u and x, unlike standard symmetric kernels. Instead, approximations to the incomplete gamma, digamma and polygamma functions studied by Funke and Hirukawa (2019, 2023) are tailored for the technical proofs. This proof strategy is novel and of independent interest; see the Appendix and/or the Supplemental Material for more details. It is also demonstrated that a uniform approximation to our diagnostic function has a unique maximum. So far existence of the maximum in the diagnostic function has been simply suggested, not formally proven, in the literature, although it is a key ingredient for consistency of the jump location estimator. A concern is that the jump location estimator tends to generate negative biases in finite samples. However, an elementary bias correction can instantly improve bias properties of the estimator without inflating its variance, and thus we advocate putting the bias-corrected version to practical use.

Acknowledging that threshold estimation is a notoriously difficult problem, many authors have proposed various threshold detection methods, mainly in the field of non-life insurance. The methods are divided roughly into three categories, namely, (i) heuristic approaches, (ii) graphical diagnostics and (iii) automated procedures. For (i), a threshold is defined as a fixed quantile (DuMouchel, 1983) or determined by some formula depending on the sample size (Loretan and Phillips, 1994); see Scarrott and MacDonald (2012) for more details. Despite no theoretical justification, these methods are used by actuaries in practical applications. Examples of (ii) include the Hill plot and its variants (e.g., Resnick 1997), the mean excess plot (Davison and Smith, 1990), and the peaks-over-threshold method (e.g., Cunnane, 1973). These are easy to grasp and thus used regularly, whereas there is room for practitioners' discretion at the stage of identifying a threshold. Recent research has shifted toward (iii) such as the minimum Kolmogorov-Smirnov (KS) distance procedure (Clauset et al., 2009; Drees et al., 2020), sequential goodness-of-fit testing (e.g., Bader et al., 2018), and the minimum quantile discrepancy and automated Eye-Balling methods (Danielsson et al., 2019). These procedures circumvent arbitrariness but rely on a certain parametric model of the tail part including the generalized Pareto distribution (GPD). Our threshold detection procedure will be able to serve as a more objective and flexible alternative to these methods.

Furthermore, there will be potentially many applications of our proposal. The definition of a 'cost' is not strict, actually. Our threshold estimation procedure is expected to work equally for non-cost variables (e.g., quantities demanded, transaction volumes, etc.), as long as shapes of their distributions have similarities to those of costs. It can be also employed for threshold detection prerequisite for tracing out the evidence of illegal trading (James et al., 2023) and implementing extreme changes in changes (CIC) estimation (Sasaki and Wang, 2023), for instance.

The remainder of this paper is organized as follows. Section 2 overviews the jump location estimation using the gamma kernel and then recommends its practical implementation. In Section 3, convergence properties of the jump location estimator, namely, strong consistency and asymptotic normality, are explored. Section 4 conducts Monte Carlo simulations to compare finite-sample behaviors of our jump location estimator with those of several existing threshold or jump detection methods. Our aim is to show numerically the advantage of our proposal over these alternatives. In Section 5, the jump location estimation is applied to a couple of real world datasets. Section 6 concludes. Proofs of theorems and propositions are provided in the Appendix. Proofs of lemmata are deferred to the Supplementary Material, which is available on the second author's webpage.

This paper adopts the following notational conventions: '$a_n \sim b_n$' means that a_n/b_n converges to 1; '$a_n = o(b_n)$' signifies that a_n/b_n converges to 0; '$a_n = O(b_n)$' means that a_n/b_n is bounded; and we say that '$a_n \asymp b_n$' if there exist constants $0 < c_1 < c_2 < \infty$ so that $c_1 a_n \leq b_n \leq c_2 a_n$. For a function $h(x)$ and a point c, $h(c^-) = \lim_{x \uparrow c} h(x)$, $h(c^+) = \lim_{x \downarrow c} h(x)$ and $h^{(m)}(x) = d^m h(x)/dx^m$ denote the left and right limits, and the mth-order derivative, respectively. The abbreviation '$a.s.$' stands for "almost surely". Finally, the expression '$X \stackrel{d}{=} Y$' reads "A random variable X obeys the distribution Y."

2 Our Proposal: An Informal Overview

2.1 Estimation of the Jump Location

It is suspected that $f(x)$, the pdf of a 'cost' variable $X \in \mathbb{R}_+$, is discontinuous at t_0 on a prespecified closed interval $I_0 := \left[\underline{t}, \overline{t}\right]$ with $0 < \underline{t} < \overline{t} < \infty$. Throughout it is assumed that the interval I_0 is located in the right tail part of the underlying cost distribution. Prior knowledge on the interval is not at all unrealistic, because quite often practitioners have a rough idea about the location of the threshold through, for example, preliminary threshold estimates, historical experiences and/or empirical quantiles. Against this background, our method can complement existing practices and objectify the graphical analysis procedures mentioned in Section 1.

The model below basically follows those of Chu and Cheng (1996) and Couallier

(1999). A similar local structure can be also found in jump detection problems for nonparametric regression (e.g., Wu and Chu, 1993a,b; Joo and Qiu, 2009) and deconvolution (e.g., Delaigle and Gijbels, 2006). It is assumed that the pdf $f(x)$ for $x \in I_0$ can be modelled locally as

$$f(x) = g(x) + d_0 \mathbf{1}\{x < t_0\}, \tag{2}$$

where $g(x)$ is a sufficiently smooth function, and $t_0 \in I_0$ is the discontinuity point (or jump location). The jump size is given by

$$d_0 = f\left(t_0^-\right) - f\left(t_0^+\right),$$

where $|d_0| \in (0, \infty)$ is assumed throughout.

Our problem is how to estimate the jump location t_0 nonparametrically. Suppose that there are n *i.i.d.* observations $\{X_i\}_{i=1}^n$ at hand. If t_0 were located in the bulk region, as in the existing literature, splitting the entire sample into two sub-samples near t_0 would cause no serious issue. Two sample sizes are roughly the same, and thus both left and right limits of a density can be estimated equally well. This is clearly not the case in our problem. Because t_0 is located in the tail region, sample-splitting near t_0 results in imbalance in sample sizes of two sub-samples and an imprecise density estimate from the right sub-sample. In view of this, we use the entire sample to estimate both limits of the pdf. To do so, we introduce 'shifted' gamma kernels $K_{G(x,b;\pm\Delta)}(\cdot)$, which are defined as pdfs of gamma distributions $G\{(x \pm \Delta)/b + 1, b\}$, i.e.,

$$K_{G(x,b;\pm\Delta)}(u) := \frac{u^{(x\pm\Delta)/b} \exp(-u/b)}{b^{(x\pm\Delta)/b+1}\Gamma\{(x \pm \Delta)/b + 1\}} \mathbf{1}\{u \geq 0\},$$

where $b\,(= b_n > 0)$ is the smoothing parameter, $\Delta\,(= \Delta_n > 0)$ plays the role of a shift parameter, and each parameter shrinks towards zero at a certain rate. Obviously, $K_{G(x,b;\pm\Delta)}(\cdot)$ collapse to Chen's (2000) original gamma kernel (1) when $\Delta = 0$. The kernels can be interpreted as those designed to smooth the data off the target design point x by a margin of Δ. In addition, they put the maximum weight at sightly left or right of x because they have their modes at $x \pm \Delta$.

We exploit as the source of our jump detection the difference between two density estimates that is created by the shift parameter Δ. Let the shifted density estimators be

$$\hat{f}^{\pm}(x) := \hat{f}^{\pm}_{b,\Delta}(x) := \frac{1}{n}\sum_{i=1}^{n} K_{G(x,b;\pm\Delta)}(X_i).$$

Also define

$$\hat{J}(x) := \hat{f}^{-}(x) - \hat{f}^{+}(x),$$

where a single, common value is chosen for the smoothing parameter b in both density estimates, as in Chu and Cheng (1996), Couallier (1999) and Huh (2002). Following these articles, we also utilize $\left|\hat{J}(x)\right|$ as the diagnostic function for threshold detection. The estimator of the jump location t_0, denoted as $\hat{t}$, is defined as the maximizer of $\left|\hat{J}(x)\right|$ on $x \in I_0$, i.e.,

$$\hat{t} := \arg\max_{x \in I_0} \left|\hat{J}(x)\right|.$$

2.2 Recommended Estimation Procedure

Monte Carlo results in Section 4 indicate that $\hat{t}$ tends to underestimate the jump location. However, it turns out that the negative bias can be alleviated substantially, with no additional cost of spread, by an elementary bias correction. In practice, we recommend the bias-corrected version of the jump location estimator

$$\tilde{t} := \hat{t} + b$$

for a suitably chosen smoothing parameter b. A theoretical foundation of the bias correction can be found in Remark 2, and its necessity is also visualized in Figure 1. Superior finite-sample properties of $\tilde{t}$ over $\hat{t}$ are confirmed in Section 4.

We can compute $\tilde{t}$ in the following steps:

1. Prespecify the interval I_0 that is likely to cover the jump location t_0.

2. Put the shift parameter $\Delta = b^{\alpha}$ for $\alpha = 0.70$ and select the smoothing parameter b via the modified likelihood cross-validation method given by (10)-(11). Both the exponent α in Δ and the choice method for b are based on our judgments from the Monte Carlo study and real data examples.

3. Find the maximizer of $\left|\hat{J}(x)\right|$ on $x \in I_0$ and take it as $\hat{t}$.

4. Obtain the bias-corrected estimate $\tilde{t} = \hat{t} + b$ using the value of b selected in Step 2.

3 Large-Sample Properties of the Jump Location Estimator

In this section convergence properties of the jump location estimator $\hat{t}$ are documented. Our particular focus is on its consistency and asymptotic normality. In the course of this, we demonstrate existence of a unique maximum in a certain uniform approximation to $\left|\hat{J}(x)\right|$ on $x \in I_0$, which constitutes a key condition for consistency of $\hat{t}$. The asymptotic distribution of $\hat{t}$ also hints that a simple form of its leading bias enables us to derive the bias-corrected estimator $\tilde{t}$.

3.1 Regularity Conditions

Convergence results below rely on the fact that $\left|\hat{J}(x)\right|$ can be approximated by the difference between two incomplete gamma functions. To deliver the results, we impose the following regularity conditions.

Assumption 1. $\{X_i\}_{i=1}^n \in \mathbb{R}_+$ are *i.i.d.* random variables.

Assumption 2.

(i) The pdf $f(x)$ is uniformly bounded on $x \in \mathbb{R}_+$.

(ii) The local structure (2) holds and $g^{(3)}(x)$ is Lipschitz continuous and bounded on $x \in I_0$.

Assumption 3. Tuning parameters b and Δ satisfy $b, \Delta \to 0$,

$$\frac{b^{3/4}}{\Delta} + \frac{\Delta}{b^{1/2}} + \frac{b^{1/2}\ln n}{n\Delta^2} \to 0, \tag{3}$$

and

$$\frac{\ln n}{nb^{3/2-\kappa}} = O(1) \tag{4}$$

for some $\kappa \in [0, 1)$, as $n \to \infty$.

All these assumptions are standard for uniform approximations to asymmetric kernel estimators. Similar conditions can be found, for example, in Hirukawa et al. (2022). It follows from Assumption 2(ii) that $f^{(1)}(t_0^-) = f^{(1)}(t_0^+)$. Boundedness of the third-order derivative of the smoothed component $g^{(3)}(\cdot)$ on I_0 is also an important ingredient for approximations to $E\left\{\hat{J}^{(p)}(x)\right\}$ for $p = 0, 1, 2$. This type of condition has been often imposed in simulation studies on jump detection (e.g., Wu and Chu, 1993a,b; Chu and Cheng, 1996).

Assumption 3 controls the shrinkage rates of tuning parameters b and Δ. The condition (3) draws the following important conclusions: (i) $b = o(\Delta)$; (ii) $b^{1/2} = o(\Delta^2/b)$; and (iii) $\Delta = o\left(\Delta^3/b^{3/2}\right)$. These are frequently used to controls remainder terms in the asymptotic expansions. It also follows from $b = o(\Delta)$ and $\Delta = o\left(b^{1/2}\right)$

that although the shift parameter Δ should shrink to zero more slowly than the smoothing parameter b, the convergence rate of Δ must not be too slow (or must be faster than $b^{1/2}$, to be more precise). Couallier (1999), for instance, also imposes a similar rate requirement. The condition (3) also implies that $\ln n / \left(nb^{1/2}\right) \to 0$ holds. The other condition (4) is an additional technical requirement for strong uniform consistency of $\hat{t}$.

3.2 Consistency

Below asymptotic properties of the jump location estimator $\hat{t}$ are explored. Our analysis starts from a uniform approximation to $\hat{f}^{\pm}(x)$ on I_0, which is documented in the next proposition. To save space, we adopt the following shorthand notation whenever no confusion may arise: $K_x^{\pm}(u) = K_{G(x,b;\pm\Delta)}(u)$; $K_x(u) = K_{G(x,b)}(u)$; $a^{\pm} = (x \pm \Delta)/b$; and $z_0 = t_0/b$.

Proposition 1. *If Assumptions 1-3 hold, then*

$$E\left\{\hat{f}^{\pm}(x)\right\} = g(x) + d_0 \int_0^{t_0} K_x^{\pm}(u)\, du \pm g^{(1)}(x)\,\Delta + \left\{g^{(1)}(x) + \frac{x}{2} g^{(2)}(x)\right\} b + o(b) \tag{5}$$

for every $x \in I_0$, and

$$\sup_{x \in I_0} \left| \hat{f}^{\pm}(x) - E\left\{\hat{f}^{\pm}(x)\right\} \right| = O\left(\sqrt{\frac{\ln n}{nb^{1/2}}}\right) \ a.s., \tag{6}$$

as $n \to \infty$.

A direct outcome from Proposition 1 is that

$$\sup_{x \in I_0} \left| \hat{J}(x) - E\left\{\hat{J}(x)\right\} \right| = O\left(\sqrt{\frac{\ln n}{nb^{1/2}}}\right) \ a.s.$$

It also follows from $\left|\left|\hat{J}(x)\right| - \left|E\left\{\hat{J}(x)\right\}\right|\right| \le \left|\hat{J}(x) - E\left\{\hat{J}(x)\right\}\right|$ that

$$\left|\hat{J}(x)\right| = \left|E\left\{\hat{J}(x)\right\}\right| + O\left(\sqrt{\frac{\ln n}{nb^{1/2}}}\right) \ a.s.$$

uniformly on I_0. This reveals that $\left|E\left\{\hat{J}(x)\right\}\right|$ constitutes the dominant term in $\left|\hat{J}(x)\right|$. In other words, the effect of the jump location on the magnitude of $\left|\hat{J}(x)\right|$

appears only in the value of $\left|E\left\{\hat{J}(x)\right\}\right|$ in a first-order asymptotic sense. This also plays a key role in the proof of Theorem 1 below; see the Appendix for more details.

By (5) and Assumption 2(ii), $\left|E\left\{\hat{J}(x)\right\}\right|$ can be further approximated by $\left|E\left\{\hat{J}(x)\right\}\right| := |d_0| J(x) + O(\Delta)$ uniformly on I_0, where

$$\begin{aligned} J(x) &= \left|\int_0^{t_0} K_x^-(u)\,du - \int_0^{t_0} K_x^+(u)\,du\right| \\ &= \int_0^{t_0} K_x^-(u)\,du - \int_0^{t_0} K_x^+(u)\,du \\ &= P\left(a^- + 1, z_0\right) - P\left(a^+ + 1, z_0\right), \end{aligned}$$

and $P(a,z) := \gamma(a,z)/\Gamma(a)$ is a normalized version of the lower incomplete gamma function $\gamma(a,z) = \int_0^z t^{a-1}\exp(-t)\,dt$ for $a, z > 0$. The reasons why $P(a^- + 1, z_0) \geq P(a^+ + 1, z_0)$ holds are that $P(a^\pm + 1, z_0) = \Pr(Y^\pm \leq z_0)$ for $Y^\pm \stackrel{d}{=} G(a^\pm + 1, 1)$ (i.e., $P(a^\pm + 1, z_0)$ are cumulative distribution functions (cdfs) of $Y^\pm$ evaluated at z_0) and that the larger the shape parameter is, the flatter the gamma distribution becomes.

We are about to demonstrate strong consistency of $\hat{t}$. Before proceeding, it is curious whether $J(x)$ on $x \in I_0$ indeed has a unique maximum at t_0 (or within a shrinking neighborhood of t_0 even if it is not maximized exactly at this point). In reality, however, it is quite cumbersome to look into the local property of $J(x)$ analytically. Fortunately, several approximations to the incomplete gamma function are available, and we rely on one of them. More specifically, we employ equation (1) of Pagurova (1965) to approximate the normalized lower incomplete gamma functions $P(a^\pm + 1, z_0)$ around the standard normal cdf. The next proposition refers to properties of the approximation and the maximizer of the approximated function.

Proposition 2. *If Assumption 3 holds, then the followings hold true.*

(i) *$J(x)$ can be further approximated by $J(x) = Q(x)\left(\Delta/b^{1/2}\right) + O\left(\Delta^3/b^{3/2}\right)$ uniformly on I_0 as $n \to \infty$, where*

$$Q(x) := \left(\frac{x + t_0}{x^{3/2}}\right)\phi\left(\frac{x - t_0}{\sqrt{bx}}\right)$$

and $\phi(\cdot)$ is the pdf of $N(0,1)$.

(ii) *$Q(x)$ on I_0 has a unique maximum at $x = t^* \in (t_0 - b, t_0)$.*

FIGURE 1 ABOUT HERE

Before establishing strong consistency of $\hat{t}$, we show by some numerical illustration that maximizing $\left|\hat{J}(x)\right|$ is a well-defined problem. A discontinuous density $f(x)$ and the diagnostic function $\left|\hat{J}(x)\right|$ are drawn in Panel (a) of Figure 1. Model 1-A in Section 4 is chosen for this illustration. The density is discontinuous at $t_0 = 4$ with the jump size $d_0 = 0.15$. The diagnostic function is computed from a Monte Carlo sample of sample size 500 under tuning parameters $b = 0.05$ and $\Delta = b^{0.70}$.

The uniform approximation $\left|\hat{J}(x)\right| \sim |d_0| J(x) \sim |d_0| Q(x) \left(\Delta/b^{1/2}\right)$ also tempts us to make a visual inspection of shapes of these three curves. The curves around the true threshold $t_0 = 4$ are plotted in Panel (b) of Figure 1. Notice that the panel magnifies the area surrounding t_0 to visualize preciseness of the approximations. It can be immediately found that all three curves are single-peaked around the true threshold, which confirms well-definedness of the optimization problem. Approximating $\left|\hat{J}(x)\right|$ by $|d_0| J(x)$ looks decent, whereas the discrepancy between the two curves suggests that the approximation errors which are asymptotically negligible may not be ignored in finite samples. Furthermore, $|d_0| Q(x) \left(\Delta/b^{1/2}\right)$ approximates $|d_0| J(x)$ quite well; rather, they are almost indistinguishable. For reference, Panel (b) also indicates maximizers of $\left|\hat{J}(x)\right|$ and $Q(x)$ are $\hat{t} \approx 3.8693$ and $t^* \approx 3.9502$, respectively. It follows from $b = 0.05$ that the latter confirms Proposition 2(ii).

The function $Q(x)$ has the following properties. Observe that

$$b^{-1/2} Q(x) = \sqrt{\frac{t_0^2/b}{2\pi x^3}} \exp\left\{-\frac{(t_0^2/b)(x-t_0)^2}{2t_0^2 x}\right\} \left(\frac{x+t_0}{t_0}\right),$$

where $\sqrt{(t_0^2/b)/(2\pi x^3)} \exp\left\{-(t_0^2/b)(x-t_0)^2/(2t_0^2 x)\right\}$ is the pdf of the inverse Gaussian distribution $IG(t_0, t_0^2/b)$. Because the shape parameter of this distribution $t_0^2/b \to \infty$, the pdf is close to a normal one for a sufficiently small $b > 0$. In addition, the distribution has mean t_0 and variance bt_0. It follows that the pdf roughly behaves like $N(t_0, bt_0)$, and thus, heuristically, the shape of $Q(x)$ also looks like a bell curve centered around t_0.

The theorem below delivers strong consistency of $\hat{t}$ for t_0. The proof of this theorem closely follows that of Theorem 1 in Chu and Cheng (1996); see the Appendix for more details.[3]

[3] In principle, weak consistency $\hat{t} \xrightarrow{p} t_0$ suffices, and some readers may wonder why Theorem 2.1 of Newey and McFadden (1994) is not employed. $Q(x)$ is uniquely maximized at $x = t^*$, I_0 is compact, and $Q(x)$ is continuous. The problem is that while $\left(b^{1/2}/\Delta\right)\left|\hat{J}(x)\right|$ is uniformly approximated by $|d_0| Q(x)$, $Q(x)$ still depends on n through b. Therefore, it is not possible to rely on the theorem directly.

Theorem 1. *If Assumptions 1-3 hold, then* $\left|\hat{t}-t_0\right| = O(b)$ *a.s., as* $n \to \infty$.

3.3 Asymptotic Normality

The theorem below documents asymptotic normality of $\hat{t}$. The asymptotic distribution is derived indirectly, as in Chu and Cheng (1996, Theorem 1), Couallier (1999, Théorème 2) and Delaigle and Gijbels (2006, Theorem 3.1). The indirect derivation comes from the fact that $\hat{t}$ solves the first-order condition $\hat{J}^{(1)}\left(\hat{t}\right) = 0$. Then, a mean-value expansion of the left-hand side around $\hat{t} = t_0$ is made, and suitable approximations to the incomplete gamma, digamma and polygamma functions are utilized in the expansion; see the Supplemental Material for more details. This is possible because unlike $f(x)$, its estimates $\hat{f}^{\pm}(x)$ are smooth functions even at t_0 due to differentiability of shifted gamma kernels $K_x^{\pm}(\cdot)$ with respect to x.

Theorem 2. *If Assumptions 1-3 hold, then*

$$\sqrt{\frac{n}{b^{1/2}}}\left\{\hat{t}-t_0-(-b)\right\} \stackrel{d}{\to} N(0,V_0) := N\left(0, \frac{3\sqrt{\pi}t_0^{1/2}}{4d_0^2}\left\{\frac{f(t_0^-)+f(t_0^+)}{2}\right\}\right)$$

as $n \to \infty$.

Remark 1. While it is difficult to obtain asymptotic bias and variance of $\hat{t}$ in light of the indirect nature, the asymptotic distribution in Theorem 2 implies the first two moments of $\hat{t}$. The dominant bias term of $\hat{t}$ is $-b$ regardless of the position of t_0. The expression of the term is much simpler than what is obtained by Couallier (1999, Théorème 2). The difference arises from different assumptions on the local structure of the pdf f on I_0; invoke that our Assumption 2(ii) follows the setup by Chu and Cheng (1996). As discussed in Remark 2 below, simplicity of the dominant bias term enables us to make the bias correction of $\hat{t}$ straightforward. Moreover, V_0, the coefficient of the dominant variance term, suggests that the larger the magnitude of discontinuity $|d_0|$, the easier the estimation of t_0. It can be also recognized that the farther t_0 moves away from the origin, the less precise its estimator becomes. Finally, both bias and variance terms are free of the shift parameter Δ. A similar result is obtained in Théorème 2 of Couallier (1999); to put it another way, Δ does not affect convergence properties of $\hat{t}$ in a first-order asymptotic sense.

Remark 2. As will be seen in the next section, $\hat{t}$ tends to yield negative biases in finite samples, i.e., it is likely to underestimate the jump location, which coincides with what Theorem 2 predicts. However, the theorem also suggests that the bias can be corrected straightforwardly by adding b to $\hat{t}$. This is the foundation of the bias-corrected estimator $\tilde{t} = \hat{t} + b$ described in Section 2.2. Indeed, the proofs of Lemma A6 and Theorem 2 jointly imply that the leading bias of $\tilde{t}$ is $O(\Delta^2)$, whereas its variance is still $O\left(b^{1/2}/n\right)$. It will be confirmed in the Monte Carlo study shortly that $\tilde{t}$ is a remedy for better finite-sample properties.

Remark 3. An approximation to the mean squared error (AMSE) of $\hat{t}$ is

$$AMSE\left(\hat{t}\right) = b^2 + \frac{b^{1/2}}{n}V_0 = O\left(b^2 + \frac{b^{1/2}}{n}\right), \tag{7}$$

where $O(b^2)$ and $O\left(b^{1/2}/n\right)$ terms are leading squared bias and variance of $\hat{t}$, respectively. The AMSE for the jump location parameter implied by Théorème 2 of Couallier (1999) is in the form of $O(h^4 + h/n)$, where h is the the bandwidth for standard symmetric kernels. It can be found that this AMSE and (7) are of the same order of magnitude by recognizing that $b \asymp h^2$. Furthermore, it follows from (7) that no bias-variance trade-off occurs, because a smaller b makes both squared bias and variance terms smaller.

Remark 4. Some readers may wonder how to pick b and Δ for super-consistency of $\hat{t}$. Then, put $\Delta \asymp b^{\alpha}$ for some $\alpha \in (1/2, 3/4)$ and $b \asymp n^{-\beta}$ for some $\beta \in (0, 1/(2\alpha - 1/2))$. Such Δ and b jointly satisfy (3) and draw the following three conclusions on the convergence rate of $\hat{t}$:

1. For any $\alpha \in (1/2, 3/4)$, we can always put $\beta > 1/2$. In this case, $AMSE\left(\hat{t}\right) = o(n^{-1})$, or $\hat{t}$ becomes super-consistent.

2. Again for any $\alpha \in (1/2, 3/4)$, we may even choose $\beta = 2/3$, which balances orders of magnitude in the squared bias and variance so that $O(b^2) = O\left(b^{1/2}/n\right) = O\left(n^{-4/3}\right)$. As a consequence, $AMSE\left(\hat{t}\right) = O\left(n^{-4/3}\right)$. It is also clear that the AMSE convergence rate of $\hat{t}$ is determined by the exponent β. $AMSE\left(\hat{t}\right) = O(b^2)$ (i.e., the squared bias dominates) for $\beta \le 2/3$, and $AMSE\left(\hat{t}\right) = O\left(b^{1/2}/n\right)$ (i.e., the squared bias becomes asymptotically negligible) otherwise. The latter case corresponds to an 'undersmoothing' scenario so that $nb^{3/2} \to 0$ holds. As a consequence, the asymptotic normality statement in Theorem 2 reduces to $\sqrt{n/b^{1/2}}\left(\hat{t} - t_0\right) \xrightarrow{d} N(0, V_0)$.

3. The best possible rate is $AMSE\left(\hat{t}\right)=O\left(n^{-2+\varepsilon}\right)$ for an arbitrarily small $\varepsilon>0$. The rate can be attained by setting α and β slightly above $1/2$ and slightly below 2, respectively. Chu and Cheng (1996) and Couallier (1999) also report that their jump location estimators can attain the same convergence rate under the best case scenario.

Furthermore, it is not hard to see that for Δ and b defined above, we can always find some $\kappa\in[0,1)$ satisfying (4). To see this, observe that (4) holds if $nb^{3/2-\kappa}\to\infty$ at a polynomial rate. The rate requirement is attained for cases 1 and 2, for example, by setting $\beta=2/3$ and choosing any $\kappa\in(0,1)$. For case 3, β slightly below 2 and κ slightly below 1 can jointly establish a polynomial divergence of $nb^{3/2-\kappa}$.

Remark 5. As long as $f(x)$ can be locally modelled as or well-approximated by (2), both $\hat{t}$ and $\tilde{t}$ become super-consistent when implemented as in Remark 4. It follows that regardless of whether a parametric (e.g., the GPD) or nonparametric model (e.g., Markovitch and Krieger, 2000) is fitted to the tail part, our jump location estimator can be safely used as a threshold estimate without deteriorating the convergence rate for the tail model. In addition, Table 1 of Clauset et al. (2009) lists examples of non-power law distributions that behave like the GPD. Clauset et al. (2009) even argue that fitting a power law distribution in their procedure has nothing to do with a plausible match of the distribution with the data, and they recommend a goodness-of-fit test as a post-estimation analysis. Super-consistency of our estimators does no harm to convergence rates of the test statistics, either.

4 Finite-Sample Performance

4.1 Monte Carlo Design

We consider two alternative models in the simulation study below. In the first case, the univariate random variable $X\in\mathbb{R}_+$ is drawn from a *log-normal* distribution, which is an example of a heavy-tailed distribution and known as its power law-like behavior (Clauset et al., 2009). What differs from a usual log-normal distribution is that a quadratic term is added to the pdf on the interval $[0,t_0)=[0,4)$. Specifically, the pdf $f(x)$ is

$$f(x)=\left\{\frac{1}{1+(2/3)\,Dt_0}\right\}\left[\frac{1}{x\sigma\sqrt{2\pi}}\exp\left\{-\frac{(\ln x-\mu)^2}{2\sigma^2}\right\}+\delta(x)\right],\ (\mu,\sigma)=\left(\frac{1}{5},\frac{3}{4}\right),$$

where $\delta(x) := D\left[1 - \{(x - t_0)/t_0\}^2\right] 1\{x < t_0\}$ and the jump size $d_0 = f\left(t_0^-\right) - f\left(t_0^+\right) = D/\{1 + (2/3) D t_0\}$. The shift parameter D takes two values, and $D = 1/4, 3/22$ yield $d_0 = 0.15, 0.10$, respectively. The former and latter cases are labelled as "Model 1-A" and "Model 1-B". It also follows from $f^{(1)}(t_0^-) = f^{(1)}(t_0^+)$ that (2) is satisfied in the neighborhood of the jump location $t_0 = 4$.

This design is in some sense similar to that of Chu and Cheng (1996), who consider the density $f(x)$ constructed by splicing left and right sides of two normal distributions with zero mean but different variances. In their design, $f(0^-) \neq f(0^+)$ but $f^{(1)}(0^-) = f^{(1)}(0^+)$, and thus (2) holds in the neighborhood of the origin (aside from the fact that the jump occurs in the middle part).

In the second case, the nonnegative random variable X is generated by some distribution spliced at $t_0 = 4$. The pdf $f(x)$ in this case takes the general form

$$f(x) = f_L(x)\,\mathbf{1}\{x < t_0\} + (1 - c_L)\, f_R(x)\,\mathbf{1}\{x \geq t_0\},$$

where $f_L(x)$ is some density function truncated at t_0, $f_R(x)$ is another density function with support on $[t_0, \infty)$, and $c_L := \int_0^{t_0} f(x)\,dx = \int_0^{t_0} f_L(x)\,dx$ ensures unity of the integral of $f(x)$ over its entire support $\mathbb{R}_+$; in other words, $f_L(x)$ and $f_R(x)$ represent bulk and tail models, respectively. This scenario is labelled as "Model 2".

Throughout the *Weibull* distribution with density

$$f_L(x) = \frac{\kappa}{\lambda}\left(\frac{x}{\lambda}\right)^{\kappa-1} \exp\left\{-\left(\frac{x}{\lambda}\right)^{\kappa}\right\},\ (\kappa, \lambda) = \left(3, \frac{11}{4}\right).$$

is considered as the bulk part of Model 2. Densities of the following distributions are examined for the tail part, and two cases are denoted as "Model 2-A" and "Model 2-B", depending on the tail model:

$$f_R(x) = \begin{cases} \frac{1}{s}\left\{1 + \frac{\xi(x-t_0)}{s}\right\}^{-(1+1/\xi)} \mathbf{1}\{x \geq t_0\}, (\xi, s) = \left(\frac{1}{4}, 4\right) & \text{[A: } \textit{GPD}\text{]} \\ \frac{k}{\ell}\exp\left\{\left(\frac{1}{\ell}\right)^k\right\}\left(\frac{x-t_0+1}{\ell}\right)^{k-1} \exp\left\{-\left(\frac{x-t_0+1}{\ell}\right)^k\right\} \mathbf{1}\{x \geq t_0 - 1\}, & \\ \qquad (k, \ell) = \left(\frac{1}{4}, 1\right) & \text{[B: } \textit{Translated Weibull}\text{]} \end{cases},$$

where $f\left(t_0^-\right) = f_L(3)$ and $f\left(t_0^+\right) = (1 - c_L)\, f_R(3)$. Drees et al. (2022, p.83) argue that discontinuity of the density at the threshold is an easy scenario for the threshold detection method by Clauset et al. (2009). Model 2-A is most favorable to existing threshold detection methods because of a power law distribution for the tail modelling as well as the discontinuity. The tail part in Model 2-B, also known as a stretched exponential distribution, reflects that again it seemingly behaves like a power law distribution (Clauset et al., 2009). Because this pdf is unbounded at the boundary of the support $t_0 - 1$, it is truncated at t_0. Moreover, these models violate (2) by construction but may be more realistic, because it is hard to judge whether the local structure (2) indeed holds in real data. Our aim is to examine how our approach can behave under such unfavorable circumstances. We simulate 1000 Monte Carlo replications of $\{X_i\}_{i=1}^n$ with sample size $n \in \{250, 500\}$ from each model.

Table 1 presents the mode of the distribution of X, the constant c_L, left and right limits of the density at the discontinuity point $f\left(t_0^{\pm}\right)$, and the jump size $d_0 = f\left(t_0^-\right) - f\left(t_0^+\right)$. In each model, more than 95% of observations concentrate on the interval $[0, t_0)$ (i.e., in the bulk region), and the density has a long tail with a polynomial decay rate. These features reasonably mimic properties of cost distributions.

TABLE 1 ABOUT HERE

Our estimation procedure for t_0 is implemented as follows. There are two optimizations required, namely, (i) the one for tuning parameters (b, Δ) and (ii) the other for jump location search. For (i), Remark 4 suggests $\alpha \in (1/2, 3/4)$, and thus we restrict our attention to four values, namely, $\alpha \in \{0.55, 0.60, 0.65, 0.70\}$. A few cross-validation (CV) methods for b are investigated, and their details are deferred to the next section. For each CV method, candidates of b are taken from 100 equally-spaced grids over the interval $[0.005, 0.500]$. For (ii), after (b, Δ) are determined, the jump location is searched via a numerical optimization routine for the diagnostic function $\left|\hat{J}(x)\right|$ on the interval $I_0 = [3, 5]$.[4] Once the jump location estimator $\hat{t}$ is obtained as the maximizer of $\left|\hat{J}(x)\right|$, the bias-corrected estimator is computed as $\tilde{t} = \hat{t} + b$.

Finite-sample performances of $\hat{t}$ and $\tilde{t}$ are compared with those of existing (i) kernel-smoothed jump location estimation procedures and (ii) automated threshold detection methods. For (i), we focus on the procedure by Chu and Cheng (1996) [CC]. The CC diagnostic function is $\left|\hat{J}_{CC}(x)\right| = \left|\hat{f}_1(x) - \hat{f}_2(x)\right|$, where $\hat{f}_j(x) = (nh)^{-1} \sum_{i=1}^n K_j\{(X_i - x)/h\}$, $j = 1, 2$, for the kernels K_1 and K_2 to be specified shortly and a common bandwidth $h\,(> 0)$. As in our method, the maximizer of $\left|\hat{J}_{CC}(x)\right|$ on I_0 is defined as the jump location estimator. The kernels K_1 and K_2

[4] The reason why different algorithms are utilized for (i) and (ii) is as follows. While a numerical optimization routine substantially reduces computation time for (i), it often finds local extrema and corner solutions because of a high degree of nonlinearity in CV criteria. A grid search can circumvent these issues. In contrast, $\left|\hat{J}(x)\right|$ is concave on I_0, and a numerical optimization routine helps expedite computation for (ii).

are fourth-order polynomial ones. These are

$$K_1(u) = \left(0.4857 - 3.8560u + 2.8262u^2 + 19.1631u^3 + 11.9952u^4\right) \times \mathbf{1}\{u \in [-1, 0.2012]\},$$

and $K_2(u) = K_1(-u)$ for all u. These kernels are also employed for jump location detection in nonparametric regression curves by Wu and Chu (1993a,b). The CC procedure is implemented as in ours. After the bandwidth value is found via grid search for a CV criterion in the next section, the jump location is searched via numerical optimization for the diagnostic function $\left|\hat{J}_{CC}(x)\right|$ on I_0.

For (ii), we investigate the followings: (a) the minimum KS distance procedure between the empirical and GPD-based distribution functions by Clauset et al. (2009) [KS]; (b) the minimum quantile discrepancy criterion for the mean absolute deviation between empirical and GPD-based quantiles by Danielsson et al. (2019) [Q-MAD]; (c) the minimum quantile discrepancy criterion for the sup-norm between empirical and GPD-based quantiles by Danielsson et al. (2019) [Q-SUP]; and (d) the automated Eye-Balling method based on tail index estimates by Danielsson et al. (2019) [AEB].

All simulations are conducted on R. In particular, R-packages "`poweRlaw`" and "`tea`" are employed to implement automated threshold detection methods (a) and (b)-(d), respectively.

4.2 Smoothing Parameter Selection

Selecting the smoothing parameter b is the most important practical issue. In our context, values of (b, Δ) must be determined before jump location search so that the diagnostic function can be fixed on I_0. However, Remark 3 does not help resolve this issue. There is no optimal choice for b on the basis of the bias-variance trade-off. Theorem 2 provides no guidance for Δ, either, because it does not automatically guarantee that any Δ satisfying (3) works equally well in finite samples. Furthermore, to the best of our knowledge, there is no decisive conclusion on selecting the tuning parameter in the context of jump location estimation; in fact, Chu and Cheng (1996) adopt fixed bandwidths in their Monte Carlo study.

Taking the dependence of Δ on b into account, we tailor Huh's (2012) approach to construct a few CV criterion functions. Before proceeding, put $\Delta = b^{\alpha}$ for a given α. Accordingly, $\hat{f}^{\pm}(x)$ are rewritten as $\hat{f}_b^{\pm}(x; \alpha)$, which signify the dependence of density estimates on (b, α). Also let

$$\hat{f}_{b,-i}^{\pm}(x; \alpha) := \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} K_x^{\pm}(X_j)$$

be density estimates using the sample with the ith observation eliminated. Finally, denote the number of observations falling into I_0 as $n_0 := \sum_{i=1}^{n} \mathbf{1}\{X_i \in I_0\}$.

The CV criterion function by Huh (2012) is defined as the sum of CV criteria for two density estimates that construct the diagnostic function for jump location estimation. We incorporate this idea into three CV criterion functions. The minimizer of each criterion function is taken as the corresponding CV smoothing parameter. The first one is the least-squares cross-validation (LSCV) criterion. It is defined as

$$CV_{LS}(b;\alpha) = CV_{LS}^{-}(b;\alpha) + CV_{LS}^{+}(b;\alpha), \tag{8}$$

where

$$CV_{LS}^{\pm}(b;\alpha) := \int_{I_0} \left\{\hat{f}_b^{\pm}(x;\alpha)\right\}^2 dx - \frac{2}{n_0} \sum_{i:X_i \in I_0} \hat{f}_{b,-i}^{\pm}(X_i;\alpha). \tag{9}$$

The remaining two criteria are likelihood-based ones. One is the simple likelihood cross-validation (LCV) criterion, which is analogous to $\hat{L}_2(h)$ of Marron (1985) and equation (2.1) of Van Es (1991). It is given by

$$CV_L(b;\alpha) = CV_L^{-}(b;\alpha) + CV_L^{+}(b;\alpha),$$

where

$$CV_L^{\pm}(b;\alpha) := -\sum_{i:X_i \in I_0} \ln\left\{\hat{f}_{b,-i}^{\pm}(X_i;\alpha)\right\}$$

is the negative log-likelihood. The other is the modified LCV criterion, denoted as MLCV, which corresponds to $\hat{L}_5(h)$ of Marron (1985) and equation (2.2) of Van Es (1991). It takes the form of

$$CV_{ML}(b;\alpha) = CV_{ML}^{-}(b;\alpha) + CV_{ML}^{+}(b;\alpha), \tag{10}$$

where

$$\begin{aligned} CV_{ML}^{\pm}(b;\alpha) &:= -\left[\sum_{i:X_i \in I_0} \ln\left\{\hat{f}_{b,-i}^{\pm}(X_i;\alpha)\right\} - \sum_{i=1}^{n} \int_{I_0} K_{X_i}^{\pm}(u)\,du\right] \\ &= -\left[\sum_{i:X_i \in I_0} \ln\left\{\hat{f}_{b,-i}^{\pm}(X_i;\alpha)\right\}\right. \\ &\quad \left.-\sum_{i=1}^{n}\left\{P\left(\frac{X_i \pm \Delta}{b}+1, \frac{\overline{t}}{b}\right) - P\left(\frac{X_i \pm \Delta}{b}+1, \frac{\underline{t}}{b}\right)\right\}\right], \end{aligned} \tag{11}$$

and the second term is intended to eliminate the endpoint effect of the interval $I_0 = \left[\underline{t}, \overline{t}\right]$. Corresponding jump location estimates are labelled "SG-LS", "SG-L" and "SG-ML", where "SG" abbreviates "shifted gamma". For their bias-corrected versions, we put "-BC" at the end.

Finally, choosing the bandwidth h is also required to implement the CC procedure. The LSCV analogous to (8)-(9) (i.e., finding a minimizer of the sum of two LSCV criteria for $\hat{f}_1(x)$ and $\hat{f}_2(x)$) is adopted.

TABLE 2 ABOUT HERE

4.3 Results

Table 2 presents several performance measures of jump location estimators. These include the bias ("Bias"), standard deviation ("SD") and root mean squared error ("RMSE") of each threshold estimator over 1000 Monte Carlo samples. In addition, for CC and SG estimators, Monte Carlo averages and standard deviations (in parentheses) of CV tuning parameters are reported for reference. Furthermore, there is no guarantee that automated threshold detection methods necessarily yield threshold estimates falling into the interval I_0. For these methods, counts of threshold estimates inside I_0 out of 1000 Monte Carlo samples ("$\#\{I_0\}$") are provided.

We start from examining the results from automated threshold detection methods. It can be found that Q-MAD generates the smallest variance among all four automated methods, resulting in the smallest RMSEs. In addition, more than 90% of estimates from Q-MAD are inside I_0 for each model and sample size, despite no restriction on the parameter space. While Models 2-A and 2-B are thought to be more favorable than 1-A and 1-B for these automated methods, Q-MAD is comparable to SG methods in terms of RMSE for the latter case. There are also general tendencies of underestimation by KS and Q-MAD and overestimation by AEB. In particular, the degree of overestimation by AEB is considerable.

CC also looks comparable to SG methods. It consistently overestimates the jump location. However, the bias reduces with the sample size, and jointly by the decrease in dispersion, its RMSE becomes smaller as the sample size is larger.

We can find a general tendency in the results from SG methods. It can be immediately found that the initial estimate $\hat{t}$ tends to be negatively biased, as Theorem 2 predicts. Moreover, a short b makes both the bias and variance small, as suggested in Remark 3. Accordingly, it can be reasonably conjectured that a CV algorithm that can generate a small smoothing parameter value $\hat{b}$ will contribute to the initial estimate $\hat{t}$ with good quality. Clearly, MLCV alone fulfills this requirement. In contrast, LSCV and LCV consistently give a large value of $\hat{b}$, which leads to a considerably biased $\hat{t}$. This is the source of their poor performance, and the sizable bias cannot be corrected completely even after adding a large $\hat{b}$ to $\hat{t}$.

In what follows, we look into SG-ML and SG-ML-BC more carefully. Although the shift parameter Δ does not enter the asymptotic normality result in Theorem 2, the choice of Δ (or the exponent α, to be more precise) matters in finite samples. For each model and sample size, smoothing parameter values via MLCV do not vary much across four values of α. On the other hand, $\alpha = 0.70$ (i.e., the smallest shift) always results in the smallest RMSE among four initial SG-ML estimates. Because bias correction is made by simply adding $\hat{b}$ to the initial estimate $\hat{t}$, additional variability through bias correction (i.e., estimation error of the correction term) is not introduced. As a consequence, superiority of the smallest shift is maintained after the bias correction. Indeed SG-ML-BC with $\alpha = 0.70$ outperforms others in terms of RMSE for Models 1-A and 1-B. Its RMSE is not the smallest for Models 2-A and 2-B; an exception is Model 2-B with $n = 500$, to be more precise. For these cases, however, the RMSE is simply larger than those from other bias corrected SG methods, and it is still smaller than those from CC and four automated methods.

Invoke that the local structure (2) holds in the neighborhood of the true jump location for Models 1-A and 1-B, whereas it is violated for remaining two models. It can be confirmed from Monte Carlo results that SG-ML-BC with $\alpha = 0.70$ is of most practical relevance and importance, because its RMSE is by far the smallest in favorable scenarios and less than those from other competing methods even in unfavorable scenarios with deviation from (2).

5 Real Data Examples

5.1 Data Description

In this section our jump location estimation procedure is applied to two datasets on cost variables that are made publicly available. The one is of actuarial losses and the other of wages. Below each dataset is discussed in detail.

The first dataset is of Danish fire insurance claims. Since the seminal analysis by McNeal (1997), the dataset has been arguably most popularly chosen for empirical studies on non-life insurance. We extract the one named "`danish`" in R-package "`SMPracticals`". The dataset contains $2,492$ fire insurance losses denominated in millions of Danish kroner from 1980 to 1990.

The second dataset is taken from Merged Outgoing Rotation Group Earnings Data of the US Current Population Survey (CPS), also known as CPS Labor Extract, which is available on the webpage of the National Bureau of Economic Research. We extract hourly wages (the variable "`earnhre`") earned by males from the dataset in 1979. Before proceeding, all observations denominated originally in US cents are converted into US dollars. The original sample size is $54,769$. Considering the computation burden of kernel smoothing, we downsize the dataset to a sub-sample of sample size $2,709$ by random sampling, which roughly accounts for 5% of the original.

Table 3 reports summary statistics of the datasets. Each distribution is right-skewed with a long tail and thus reasonably represents stylized facts of a 'cost' distribution. It can be also found that the distribution of the downsized dataset of US male wages well represents that of the original.

TABLES 3-4 ABOUT HERE

5.2 Estimation Details

Based on simulation results in Section 4, we compare SG-ML(-BC) with $\alpha = 0.70$ with KS, Q-MAD, Q-SUP, AEB, and CC. The tuning parameter for each of SG-ML(-BC) and CC is taken from 100 equally-spaced grids over the interval $[0.005, 0.500]$, and then the threshold estimate $\hat{t}$ is searched via a numerical optimization routine of the corresponding diagnostic function.

The prespecified interval I_0 for each dataset roughly covers the tail part of the distribution. In addition, Cooray and Ananda (2005) and Scollnik and Sun (2012) estimate several different parametric composite models from the Danish fire insurance dataset and obtain threshold estimates ranging roughly from 1 through 3. Using the same dataset, Reynkens et al. (2017) report the threshold estimate of 17 via a graphical method. The interval I_0 for this dataset also incorporates these empirical findings.

5.3 Results

Table 4 presents the estimation results. For the Danish fire insurance data, there is no consensus in threshold estimates among four GPD-based detection methods. The estimation result from SG-ML is fairly close to the one from KS, so is the one from its bias-corrected version SG-ML-BC. In contrast, the result from CC is problematic and recognized as an example of estimation failure in that the threshold estimate is a corner solution.

On the other hand, threshold estimates from four GPD-based methods for the US male wage data are similar one another. SG-ML and SG-ML-BC yield threshold estimates near 10, which is considerably close to the results from Q-MAD and Q-SUP. Once again, CC may have failed to estimate the threshold, because its estimate lies almost on the boundary.

As the location of the threshold is unknown in each distribution, it is hard to judge among all threshold estimators considered. Nonetheless, it is safe to say that SG-ML(-BC) can serve as a good alternative to existing threshold detection methods.

6 Conclusion

It is widely recognized that a single model cannot describe the while range of a cost distribution well. Accordingly, it is of growing importance and interest to find the location of a threshold at which two different models are spliced, whereas this estimation problem is known to be notoriously difficult. This paper has explored a method of estimating the threshold of a cost distribution nonparametrically. Development of our method starts from tailoring the existing techniques for change point or jump location detection in statistics to stylized facts of cost distributions. The diagnostic function is the absolute difference of two kernel density estimates, and the maximizer of the function over a prespecified interval is defined as the jump location estimator. Because the threshold is located in the right tail region, we propose to compute two density estimates using the entire sample and shifted gamma kernels. Our jump location estimator is shown to be super-consistent and asymptotically normal when suitably implemented. The proof strategy is also new in that approximations to the incomplete gamma, digamma and polygamma functions are utilized. Since the estimator tends to underestimate the jump location and its dominant bias term takes a simple form, we advocate correcting its bias in a straightforward manner. It is confirmed in the Monte Carlo study that as a result of the bias correction, the negative bias is substantially eliminated with no additional cost of spread. Several data-based methods of choosing the smoothing parameter are also assessed via simulations. Our proposal is finally applied to three kinds of real world datasets. Judging from simulation results and real data examples, we recommend SG-ML-BC for practical use.

A Appendix

The Appendix provides technical proofs of theorems and propositions. To save space, we defer proofs of lemmata to the Supplemental Material. Before proceeding, additional short-handed notation is introduced, and a few useful formulae related to the gamma function are presented.

A.1 Additional Notation

$\Psi(x) = d\ln\Gamma(x)/dx = \Gamma^{(1)}(x)/\Gamma(x)$ and $\Psi^{(m)}(x) = d^m\Psi(x)/dx^m$ signify the digamma and polygamma functions, respectively. In addition, the following notation is adopted in the proofs: $R(a,z) = z^a\exp(-z)/\Gamma(a+1)$ for $a, z > 0$; $\dot{K}_c^{\pm}(u) = \partial K_x^{\pm}(u)/\partial x|_{x=c}$; $\ddot{K}_c^{\pm}(u) = \partial^2 K_x^{\pm}(u)/\partial x^2|_{x=c}$; $\dddot{K}_c^{\pm}(u) = \partial^3 K_x^{\pm}(u)/\partial x^3|_{x=c}$; $H_i = \dot{K}_{t_0}^{-}(X_i) - \dot{K}_{t_0}^{+}(X_i)$; and $z^{\pm} = (t_0 \pm \Delta)/b$.

A.2 Useful Formulae on the Gamma Function

Stirling's Formula.

$$\Gamma(a+1) = \sqrt{2\pi}a^{a+1/2}\exp(-a)\left\{1 + \frac{1}{12a} + O\left(a^{-2}\right)\right\} \text{ as } a \to \infty. \tag{A1}$$

Recursive Formula for the Lower Incomplete Gamma Function.

$$\gamma(a+1,z) = a\gamma(a,z) - z^a\exp(-z) \text{ for } a, z > 0. \tag{A2}$$

Recursive Formula for the Polygamma Function.

$$\Psi^{(m)}(a+1) = \Psi^{(m)}(a) + \frac{(-1)^m m!}{a^{m+1}} \text{ for } a > 0 \text{ and } m \in \{0,1,2,...\}. \tag{A3}$$

A.3 Proof of Proposition 1

The proof requires the following lemmata.

Lemma A1. *As $b \to 0$,*

$$\sup_{x\in I_0}\left|\Psi\left(\frac{x}{b}+1\right) - \left\{\ln\left(\frac{x}{b}\right) + \frac{b}{2x}\right\}\right| = O\left(b^2\right),$$

$$\sup_{x\in I_0}\left|\Psi^{(1)}\left(\frac{x}{b}+1\right) - \left(\frac{b}{x} - \frac{b}{2x^2}\right)\right| = O\left(b^3\right), \text{ and}$$

$$\sup_{x\in I_0}\left|\Psi^{(m)}\left(\frac{x}{b}+1\right)\right| = O\left(b^m\right) \text{ for } m \geq 2.$$

Lemma A2. *Let* $C_0 := 2\max\left\{\underline{t}^{-3/2}, \bar{t}^{1/2}\right\} + \underline{t}^{-1/2}$. *Then, as* $n \to \infty$,

$$\sup_{(x,u)\in I_0\times\mathbb{R}_+}\left|\dot{K}_x^{\pm}(u)\right| \le b^{-3/2}\sqrt{\frac{2}{\pi}}C_0.$$

Lemma A3. (Van der Vaart and Wellner, 1996, Lemma 2.2.9) *Let* $X_1,\ldots,X_n$ *be independent random variables with bounded ranges* $[-M,M]$ *and zero means. Then,*

$$\Pr\left(\left|\sum_{i=1}^{n}X_i\right| > x\right) \le 2\exp\left\{-\frac{x^2}{2(v+Mx/3)}\right\}$$

for all x *and* $v \ge Var\left(\sum_{i=1}^{n}X_i\right)$.

A.3.1 Proof of Proposition 1

Proof of (5). Because

$$E\left\{\hat{f}^{\pm}(x)\right\} = \int_0^\infty K_x^{\pm}(u)\,g(u)\,du + d_0\int_0^{t_0}K_x^{\pm}(u)\,du,$$

it suffices to show that

$$\int_0^\infty K_x^{\pm}(u)\,g(u)\,du = g(x) \pm g^{(1)}(x)\,\Delta + \left\{g^{(1)}(x) + \frac{x}{2}g^{(2)}(x)\right\}b + o(b). \qquad \text{(A4)}$$

Observe that the left-hand side of (A4) can be expressed as $E\{g(X^{\pm})\}$ for $X^{\pm} \overset{d}{=} G(a^{\pm}+1,b)$. A second-order Taylor expansion of $g(X^{\pm})$ around $X^{\pm}=x$ yields

$$E\left\{g\left(X^{\pm}\right)\right\} = g(x) + g^{(1)}(x)\,E\left(X^{\pm}-x\right) + \frac{1}{2}g^{(2)}(x)\,E\left(X^{\pm}-x\right)^2 + R_g, \qquad \text{(A5)}$$

where $\bar{x}^{\pm}$ is between $X^{\pm}$ and x, and $R_g := (1/2)\left\{g^{(2)}(\bar{x}^{\pm}) - g^{(2)}(x)\right\}E\left(X^{\pm}-x\right)^2$ is the remainder term. Then, by the property of gamma random variables and Assumption 3, $E\left(X^{\pm}-x\right) = \pm\Delta + b$ and $E\left(X^{\pm}-x\right)^2 = xb + \Delta^2 \pm 3\Delta b + 2b^2 = xb + o(b)$. $R_g = o(b)$ can be also shown by using Lipschitz continuity of $g^{(2)}(\cdot)$. Then, (A4) is demonstrated by substituting these results into (A5).

Proof of (6). The proof for this part closely follows the one for Theorem 2 of Hirukawa et al. (2022). For ease of exposition, let $a_n := \sqrt{\ln n/(nb^{1/2})}$, $\varsigma_{in}^{\pm}(x) := (1/n)\left[K_x^{\pm}(X_i) - E\left\{K_x^{\pm}(X_i)\right\}\right]$, $N_n := n^{1+\epsilon}b^{-3/2}$ for a sufficiently small $\epsilon > 0$, and $|I_0| := \bar{t} - \underline{t}$. The proof for this part takes the following two steps.

1. Split the interval I_0 into N_n equally-spaced grids to create N_n sub-intervals with length $N_n^{-1}|I_0|$, and replace the supremum with a maximization over finite N_n sub-intervals.

2. Employ Lemma A3 (Bernstein's inequality) to bound the remainder term.

Step 1. Let I_j, $j = 1, \ldots, N_n$, be the jth sub-interval. Also let x_j be the right-most point in I_j with $x_0 \equiv \underline{t}$ and $x_{N_n} \equiv \overline{t}$. Suppose that the design point x falls into I_j. By the mean-value theorem Lemma A2 and Assumption 3,

$$\begin{aligned}\left| K_x^{\pm}(u) - K_{x_j}^{\pm}(u) \right| &\leq \sup_{(x,u)\in I_0\times\mathbb{R}_+} \left| \dot{K}_x^{\pm}(u) \right| \sup_{x\in I_j} |x - x_j| \\ &\leq O\left(b^{-3/2}\right) O\left(N_n^{-1}\right) = O\left\{n^{-(1+\epsilon)}\right\} \leq O(a_n).\end{aligned}$$

It follows from C_r-inequality that

$$\max_{1\leq j\leq N_n} \sup_{x\in I_j} \left| \sum_{i=1}^{n} \varsigma_{in}^{\pm}(x) - \sum_{i=1}^{n} \varsigma_{in}^{\pm}(x_j) \right| = O(a_n). \tag{A6}$$

Step 2. Before employing Bernstein's inequality in Lemma A3, we must determine two bounds M and v. First, it holds that for a given $x > \Delta\,(> 0)$,

$$K_x^{\pm}(u) \leq \frac{b^{-1/2}(x \pm \Delta)^{-1/2}\{1 + o(1)\}}{\sqrt{2\pi}}. \tag{A7}$$

This can be confirmed by recognizing that $K_x^{\pm}(u)$ is maximized at $u = x \pm \Delta$ and employing (A1). Then, by (A7),

$$\sup_{(x,u)\in I_0\times\mathbb{R}_+} \left| K_x^{\pm}(u) \right| \leq b^{-1/2}\sqrt{\frac{2}{\pi\underline{t}}},$$

and thus, by C_r-inequality,

$$\left| \varsigma_{in}^{\pm}(x) \right| \leq 2\sqrt{\frac{2}{\pi\underline{t}}} n^{-1} b^{-1/2} = 2\sqrt{\frac{2}{\pi\underline{t}}} \frac{a_n^2}{\ln n} =: M.$$

Second, Assumption 2(i) implies that there is some constant $\overline{C} \in (0, \infty)$ so that $\sup_{x\in\mathbb{R}_+} f(x) \leq \overline{C}$. Then,

$$\begin{aligned} Var\left\{ \sum_{i=1}^{n} \varsigma_{in}^{\pm}(x) \right\} &= \sum_{i=1}^{n} Var\left\{ \varsigma_{in}^{\pm}(x) \right\} \\ &= \sum_{i=1}^{n} E\left\{ \varsigma_{in}^{\pm}(x) \right\}^2 \\ &\leq \sum_{i=1}^{n} E\left\{ \frac{1}{n} K_x^{\pm}(X_i) \right\}^2 \\ &\leq \frac{\overline{C}}{n} \int_0^{\infty} \left\{ K_x^{\pm}(u) \right\}^2 du. \end{aligned}$$

Also let

$$A^{\pm}(x) := \frac{b^{-1}\Gamma(2a^{\pm}+1)}{2^{2a^{\pm}+1}\Gamma^2(a^{\pm}+1)}. \tag{A8}$$

Observe that

$$\int_0^{\infty}\left\{K_x^{\pm}(u)\right\}^2 du = A^{\pm}(x)\int_0^{\infty}\frac{u^{2a^{\pm}}\exp\{-u/(b/2)\}}{(b/2)^{2a^{\pm}+1}\Gamma(2a^{\pm}+1)}du = A^{\pm}(x),$$

because the integrand in the middle term is the pdf of $G(2a^{\pm}+1, b/2)$. Using (A1), $(x \pm \Delta)^{-1/2} = x^{-1/2}\{1+o(1)\}$ and $o(1) \leq 1$ for a sufficiently large n gives

$$A^{\pm}(x) = \frac{b^{-1/2}(x \pm \Delta)^{-1/2}\{1+o(1)\}}{2\sqrt{\pi}} \leq \frac{b^{-1/2}}{\sqrt{\pi \underline{t}}}$$

uniformly on I_0. In sum,

$$Var\left\{\sum_{i=1}^{n}\varsigma_{in}^{\pm}(x)\right\} \leq \frac{\overline{C}}{\sqrt{\pi \underline{t}}}n^{-1}b^{-1/2} = \frac{\overline{C}}{\sqrt{\pi \underline{t}}}\frac{a_n^2}{\ln n} =: v.$$

Lemma A3 establishes that for such M and v and an arbitrarily chosen $K > 0$,

$$\Pr\left\{\left|\sum_{i=1}^{n}\varsigma_{in}^{\pm}(x)\right| > K\sqrt{\frac{\overline{C}}{\sqrt{\pi \underline{t}}}}a_n\right\} \leq 2\exp\left\{-\frac{K^2 \ln n}{2\left(1+\frac{2}{3}\sqrt{\frac{2}{\pi \underline{t}}}Ka_n/\sqrt{\frac{\overline{C}}{\sqrt{\pi \underline{t}}}}\right)}\right\}.$$

It follows from $a_n = o(1)$ that $(2/3)\sqrt{2/(\pi \underline{t})}Ka_n/\sqrt{\overline{C}/\sqrt{\pi \underline{t}}} \leq 1$ holds for a sufficiently large n. Accordingly,

$$\Pr\left\{\left|\sum_{i=1}^{n}\varsigma_{in}^{\pm}(x)\right| > K\sqrt{\frac{\overline{C}}{\sqrt{\pi \underline{t}}}}a_n\right\} \leq 2\exp\left\{-\frac{K^2 \ln n}{2(1+1)}\right\} = 2n^{-\frac{K^2}{4}}.$$

In the end,

$$\begin{aligned}
&\Pr\left\{\max_{1\leq j\leq N_n}\left|\sum_{i=1}^{n}\varsigma_{in}^{\pm}(x_j)\right| > K\sqrt{\frac{\overline{C}}{\sqrt{\pi \underline{t}}}}a_n\right\} \\
&\leq \sum_{i=1}^{N_n}\max_{1\leq j\leq N_n}\Pr\left\{\left|\sum_{i=1}^{n}\varsigma_{in}^{\pm}(x_j)\right| > K\sqrt{\frac{\overline{C}}{\sqrt{\pi \underline{t}}}}a_n\right\} \\
&\leq N_n \cdot \max_{1\leq j\leq N_n}\Pr\left\{\left|\sum_{i=1}^{n}\varsigma_{in}^{\pm}(x_j)\right| > K\sqrt{\frac{\overline{C}}{\sqrt{\pi \underline{t}}}}a_n\right\} \\
&= O\left(N_n n^{-K^2/4}\right).
\end{aligned}$$

Moreover, (4) implies that

$$b^{-3/2} = O\left\{ n^{\frac{1}{1-\kappa}} \left(\frac{b^{\frac{\kappa}{2}}}{\ln n} \right)^{\frac{1}{1-\kappa}} \right\} \leq O\left(n^{\frac{1}{1-\kappa}} \right),$$

where the last inequality holds because $b^{\kappa/2}/\ln n$ is convergent. Then, picking $K = 2\sqrt{2(1+\epsilon) + 1/(1-\kappa)}$ yields $N_n n^{-K^2/4} = O\left\{ n^{-(1+\epsilon)} \right\}$ so that

$$\sum_{n=1}^{\infty} \Pr\left\{ \max_{1\leq j\leq N_n} \left| \sum_{i=1}^{n} \varsigma_{in}^{\pm}(x_j) \right| > K\sqrt{\frac{\overline{C}}{\sqrt{\pi}\underline{t}}} a_n \right\} \leq \sum_{n=1}^{\infty} O\left(\frac{1}{n^{1+\epsilon}} \right) < \infty.$$

Therefore, by the Borel-Cantelli lemma,

$$\max_{1\leq j\leq N_n} \left| \sum_{i=1}^{n} \varsigma_{in}^{\pm}(x_j) \right| = O(a_n) \;\; a.s. \tag{A9}$$

It follows from (A6) and (A9) that

$$\begin{aligned} & \sup_{x\in I_0} \left| \hat{f}^{\pm}(x) - E\left\{ \hat{f}^{\pm}(x) \right\} \right| \\ & \leq \max_{1\leq j\leq N_n} \left| \sum_{i=1}^{n} \varsigma_{in}^{\pm}(x_j) \right| + \max_{1\leq j\leq N_n} \sup_{x\in I_j} \left| \sum_{i=1}^{n} \varsigma_{in}^{\pm}(x) - \sum_{i=1}^{n} \varsigma_{in}^{\pm}(x_j) \right| \\ & = O(a_n) \;\; a.s. \end{aligned}$$

This completes the proof. ■

A.4 Proof of Proposition 2

Proof of (i). Dividing both sides of (A2) by $\Gamma(a^{\pm}+1)$ and using $\Gamma(a^{\pm}+1) = a^{\pm}\Gamma(a^{\pm})$, we have

$$P\left(a^{\pm}+1, z_0\right) = P\left(a^{\pm}, z_0\right) - R\left(a^{\pm}, z_0\right).$$

Then, $J(x)$ can be rewritten as

$$J(x) = \left\{ P\left(a^{-}, z_0\right) - P\left(a^{+}, z_0\right) \right\} - \left\{ R\left(a^{-}, z_0\right) - R\left(a^{+}, z_0\right) \right\}. \tag{A10}$$

We start from approximating $P(a^{\pm}, z_0)$. Let $y^{\pm}$ solve the equation $z_0 = a^{\pm} + \sqrt{a^{\pm}} y^{\pm}$. Then,

$$y^{\pm} = \frac{z_0 - a^{\pm}}{\sqrt{a^{\pm}}} = \frac{t_0 - (x \pm \Delta)}{\sqrt{b(x\pm\Delta)}}.$$

It follows from equation (1) of Pagurova (1965) that

$$P\left(a^{\pm}, z_0\right) = P\left(a^{\pm}, a^{\pm} + \sqrt{a^{\pm}} y^{\pm}\right) = \Phi\left(y^{\pm}\right) - \frac{1}{3\sqrt{a^{\pm}}} \Phi^{(3)}\left(y^{\pm}\right) + R_{a^{\pm}}, \tag{A11}$$

where $\Phi(\cdot)$ is the cdf of $N(0,1)$, and the remainder term $R_{a^{\pm}}$ is $(1/a^{\pm})$ times some linear combination of higher-order derivatives of $\Phi(\cdot)$ evaluated at $y^{\pm}$. Now, a third-order Taylor expansion of $\Phi(y^{\pm})$ around $\Delta = 0$ yields

$$\begin{aligned}
&\Phi\left(y^{\pm}\right) \\
&= \Phi\left(\frac{t_0 - x}{\sqrt{bx}}\right) \mp \left(\frac{t_0 + x}{2x^{3/2}}\right) \phi\left(\frac{t_0 - x}{\sqrt{bx}}\right)\left(\frac{\Delta}{b^{1/2}}\right) \\
&+ \frac{1}{2}\left\{\left(\frac{t_0 + x}{4x^3}\right) \phi^{(1)}\left(\frac{t_0 - x}{\sqrt{bx}}\right)\left(\frac{\Delta^2}{b}\right)\right. \\
&\left. + \left(\frac{3t_0 + x}{4x^{5/2}}\right) \phi\left(\frac{t_0 - x}{\sqrt{bx}}\right)\left(\frac{\Delta^2}{b^{1/2}}\right)\right\} + O\left(\frac{\Delta^3}{b^{3/2}}\right),
\end{aligned} \tag{A12}$$

where the $O\left(\Delta^3/b^{3/2}\right)$ rate is uniform on I_0. It follows that

$$\Phi\left(y^{-}\right) - \Phi\left(y^{+}\right) = \left(\frac{t_0 + x}{x^{3/2}}\right) \phi\left(\frac{t_0 - x}{\sqrt{bx}}\right)\left(\frac{\Delta}{b^{1/2}}\right) + O\left(\frac{\Delta^3}{b^{3/2}}\right). \tag{A13}$$

Next we work on the term involving $\Phi^{(3)}\left(y^{\pm}\right)/\sqrt{a^{\pm}}$. Notice that

$$\frac{1}{\sqrt{a^{\pm}}} = \frac{b^{1/2}}{\sqrt{x}}\left\{1 \mp \frac{\Delta}{2x} + O\left(\Delta^2\right)\right\}, \tag{A14}$$

where the $O\left(\Delta^2\right)$ rate is uniform on I_0. Moreover, a second-order Taylor expansion of $\Phi^{(3)}\left(y^{\pm}\right)$ around $\Delta = 0$ yields

$$\Phi^{(3)}\left(y^{\pm}\right) = \Phi^{(3)}\left(\frac{t_0 - x}{\sqrt{bx}}\right) \mp \left(\frac{t_0 + x}{2x^{3/2}}\right) \Phi^{(4)}\left(\frac{t_0 - x}{\sqrt{bx}}\right)\left(\frac{\Delta}{b^{1/2}}\right) + O\left(\frac{\Delta^2}{b}\right),$$

where the $O\left(\Delta^2/b\right)$ rate is again uniform on I_0. By (3), we have

$$\begin{aligned}
&\left|\frac{1}{\sqrt{a^{-}}} \Phi^{(3)}\left(y^{-}\right) - \frac{1}{\sqrt{a^{+}}} \Phi^{(3)}\left(y^{+}\right)\right| \\
&\leq O\left(b^{1/2}\right) O\left(\frac{\Delta}{b^{1/2}}\right) = O(\Delta) = o\left(\frac{\Delta^3}{b^{3/2}}\right)
\end{aligned} \tag{A15}$$

uniformly on I_0.

Furthermore, using the above argument leads to

$$\left|R_{a^{-}} - R_{a^{+}}\right| \leq O(b)\, O\left(\frac{\Delta}{b^{1/2}}\right) = O\left(\Delta b^{1/2}\right) = o\left(\frac{\Delta^3}{b^{3/2}}\right) \tag{A16}$$

again uniformly on I_0. Combining (A11), (A13), (A15) and (A16) gives

$$P\left(a^{-},z_0\right)-P\left(a^{+},z_0\right)=\left(\frac{t_0+x}{x^{3/2}}\right)\phi\left(\frac{t_0-x}{\sqrt{bx}}\right)\left(\frac{\Delta}{b^{1/2}}\right)+O\left(\frac{\Delta^3}{b^{3/2}}\right), \tag{A17}$$

where the $O\left(\Delta^3/b^{1/2}\right)$ rate is uniform on I_0.

We should also evaluate the order of magnitude in $R\left(a^{-},z_0\right)-R\left(a^{+},z_0\right)$. Using (A1) yields

$$\begin{aligned} R\left(a^{\pm},z_0\right) &= \frac{z_0^{a^{\pm}}\exp\left(-z_0\right)}{\Gamma\left(a^{\pm}+1\right)} \\ &= \left\{\frac{1+O\left(1/a^{\pm}\right)}{\sqrt{2\pi}}\right\}\frac{1}{\sqrt{a^{\pm}}}\left\{\left(\frac{z_0}{a^{\pm}}\right)^{a^{\pm}}\exp\left(a^{\pm}-z_0\right)\right\}, \end{aligned} \tag{A18}$$

where $\left\{1+O\left(1/a^{\pm}\right)\right\}/\sqrt{2\pi}=O(1)$ uniformly on I_0. In addition,

$$\left(\frac{z_0}{a^{\pm}}\right)^{a^{\pm}}\exp\left(a^{\pm}-z_0\right)=:\exp\left\{a^{\pm}\left(\ln\rho^{\pm}+1-\rho^{\pm}\right)\right\} \tag{A19}$$

for $\rho^{\pm}:=z_0/a^{\pm}>0$. It follows from $\ln\rho^{\pm}+1-\rho^{\pm}\leq 0$ that $\exp\left\{a^{\pm}\left(\ln\rho^{\pm}+1-\rho^{\pm}\right)\right\}\leq 1\leq O(1)$. Then, by (A14), we have

$$\left|R\left(a^{-},z_0\right)-R\left(a^{+},z_0\right)\right|\leq O\left(\Delta b^{1/2}\right)=o\left(\frac{\Delta^3}{b^{3/2}}\right) \tag{A20}$$

uniformly on I_0. Part (i) is established by substituting (A17) and (A20) into (A10) and then denoting $Q(x):=\left\{\left(t_0+x\right)/x^{3/2}\right\}\phi\left\{\left(t_0-x\right)/\sqrt{bx}\right\}$.

Proof of (ii). Notice that

$$Q^{(1)}(x)=\frac{q(x)}{2\sqrt{2\pi}bx^{7/2}}\exp\left\{-\frac{\left(x-t_0\right)^2}{2bx}\right\}, \tag{A21}$$

where $q(x):=-bx\left(x+3t_0\right)-\left(x-t_0\right)\left(x+t_0\right)^2$. Heuristically, $q(x)\approx-\left(x-t_0\right)\left(x+t_0\right)^2$ for $b\approx 0$, and we can see that $Q(x)$ is maximized at $x\approx t_0$. To be more precise, the maximizer of $Q(x)$ is given by the solution of $q(x)=0$ or $-bx\left(x+3t_0\right)=\left(x-t_0\right)\left(x+t_0\right)^2$. It is not hard to see that for a sufficiently small $b>0$, the latter has a unique solution on $\mathbb{R}_+$.

Let t^* be the solution. Our argument so far suggests that $t^*\approx t_0$ for a sufficiently small $b>0$. Now we even conjecture that $t^*=t_0+cb$ for some $|c|<\infty$. To verify this conjecture, consider that $0=q\left(t^*\right)=-b\left\{\left(t_0+cb\right)\left(4t_0+cb\right)+c\left(2t_0+cb\right)^2\right\}$. Because $b>0$, c solves $\left(t_0+cb\right)\left(4t_0+cb\right)+c\left(2t_0+cb\right)^2=0$. This equation can be further rewritten as $b(c+1)\left(c+2t_0/b\right)^2=-t_0c$. Observe that the left- and right-hand sides are cubic and downward sloping linear functions of c, respectively. For a sufficiently small $b>0$, $-2t_0/b<-1$ holds, and in this case we can recognize that the equation has a unique solution $c\in(-1,0)$. This completes the proof. ■

A.5 Proof of Theorem 1

Because $\left|\hat{t}-t_0\right| \leq \left|\hat{t}-t^*\right| + |t^*-t_0|$ and we have already known in Proposition 2(ii) that $|t^*-t_0| \leq b$, it suffices to show that

$$\left|\hat{t}-t^*\right| = O\left(b\right) \;\; a.s. \tag{A22}$$

The proof for (A22) closely follows the one for Theorem 1 of Chu and Cheng (1996). We keep using the same notation as in the proof of (6) in Proposition 1. In addition, define $\Upsilon := \{x : x \in I_0, |x-t^*| > b\,|I_0|\}$. Also let w be an element of $E_n := \{x_0, x_1, \ldots, x_{N_n}\}$ that is closest to t^*, i.e.,

$$|w-t^*| = \min_{0\leq j\leq N_n} |x_j - t^*| \Leftrightarrow w = \arg\min_{s\in E_n} |s-t^*|\,.$$

To establish (A22), we need to show that $\Pr\left\{\hat{t}\in\Upsilon\right\} = 0$. However, the set $\left\{\sup_{x\in\Upsilon}\left|\hat{J}\left(x\right)\right| \geq \left|\hat{J}\left(w\right)\right|\right\}$ is larger than the set $\left\{\hat{t}\in\Upsilon\right\}$, and thus $\Pr\left\{\hat{t}\in\Upsilon\right\} \leq \Pr\left\{\sup_{x\in\Upsilon}\left|\hat{J}\left(x\right)\right| \geq \left|\hat{J}\left(w\right)\right|\right\}$ is the case. In what follows, we demonstrate that

$$\Pr\left\{\sup_{x\in\Upsilon}\left|\hat{J}\left(x\right)\right| \geq \left|\hat{J}\left(w\right)\right|\right\} = 0. \tag{A23}$$

Let $x^* := \arg\sup_{x\in\Upsilon}\left|E\left\{\hat{J}\left(x\right)\right\}\right|$. Then, by (5) and Proposition 2,

$$\left|E\left\{\hat{J}\left(w\right)\right\}\right| - \sup_{x\in\Upsilon}\left|E\left\{\hat{J}\left(x\right)\right\}\right| = |d_0|\left\{Q\left(w\right)-Q\left(x^*\right)\right\}\left(\frac{\Delta}{b^{1/2}}\right) + O\left(\frac{\Delta^3}{b^{3/2}}\right).$$

For a sufficiently large n, w is closer to t^* than x^*. Then, $Q\left(w\right) - Q\left(x^*\right) > 0$ is the case. In addition, $Q\left(w\right) - Q\left(x^*\right) = Q^{(1)}\left(\bar{w}\right)\left(w-x^*\right)$ for some $\bar{w}$ between w and x^* by the mean value theorem. Combining these, we may write $Q\left(w\right) - Q\left(x^*\right) = \left|Q^{(1)}\left(\bar{w}\right)\right||w-x^*|$. Now, by (A21), $\left|Q^{(1)}\left(x\right)\right| = O\left(b^{-1}\right)$ uniformly on $x\in I_0$. Because $\left|Q^{(1)}\left(\bar{w}\right)\right| > 0$, there is some constant $\underline{C} > 0$ so that $b\left|Q^{(1)}\left(\bar{w}\right)\right| \geq \underline{C}$, or $\left|Q^{(1)}\left(\bar{w}\right)\right| \geq \underline{C}/b$ holds. Moreover,

$$|w-x^*| \geq |x^*-t^*| - |w-t^*| \geq b\,|I_0| - N_n^{-1}\,|I_0| = b\left\{1-n^{-(1+\epsilon)}b^{1/2}\right\}|I_0|\,.$$

Since $n^{-(1+\epsilon)}b^{1/2}$ is convergent, we may put $n^{-(1+\epsilon)}b^{1/2} \leq 1/2$ for a sufficiently large n so that $|w-x^*| \geq (1/2)\,b\,|I_0|$. Therefore,

$$\left|E\left\{\hat{J}\left(w\right)\right\}\right| - \sup_{x\in\Upsilon}\left|E\left\{\hat{J}\left(x\right)\right\}\right| \geq 3L + O\left(\frac{\Delta^3}{b^{3/2}}\right),$$

where

$$L = L_n := \frac{1}{6}\,|d_0|\,\underline{C}\,|I_0|\left(\frac{\Delta}{b^{1/2}}\right).$$

By this result and a straightforward calculation,

$$\sup_{x\in\Upsilon}\left|\hat{J}(x)\right|-\left|\hat{J}(w)\right|\leq 2\sup_{x\in I_0}\left|\hat{J}(x)-E\left\{\hat{J}(x)\right\}\right|-3L+O\left(\frac{\Delta^3}{b^{3/2}}\right).$$

However, combining the definition of L with (3) and (6) implies that

$$2\left[\sup_{x\in I_0}\left|\hat{J}(x)-E\left\{\hat{J}(x)\right\}\right|-L\right]+\left\{O\left(\frac{\Delta^3}{b^{3/2}}\right)-L\right\}<0$$

with probability one. Therefore, $\Pr\left\{\sup_{x\in\Upsilon}\left|\hat{J}(x)\right|<\left|\hat{J}(w)\right|\right\}=1$, and (A23) is established. This completes the proof. ■

A.6 Proof of Theorem 2

The proof requires the following lemmata.

Lemma A4. *For $a,z,\lambda>0$ and $m\in\{0,1,2,\ldots\}$,*

$$\lambda^m\frac{\gamma(a+m+1,z)}{\Gamma(a+1)}=:p_mP(a,z)-r_mR(a,z),$$

where $p_{m+1}=\lambda(a+m+1)p_m$, $r_{m+1}=\lambda(a+m+1)r_m+(\lambda z)^{m+1}$, and $p_0=r_0=1$.

Lemma A5. *As $n\to\infty$,*

$$P\left(z^{\pm},z_0\right)=\frac{1}{2}\mp\frac{1}{\sqrt{2\pi}\sqrt{t_0}}\left(\frac{\Delta}{b^{1/2}}\right)+O\left(\frac{\Delta^2}{b^{1/2}}\right),\text{ and}$$

$$R\left(z^{\pm},z_0\right)=\frac{b^{1/2}}{\sqrt{2\pi}\sqrt{t_0}}\left[1\mp\frac{\Delta}{2t_0}-\frac{1}{2t_0}\left(\frac{\Delta^2}{b}\right)+O\left\{\max\left(b,\frac{\Delta^3}{b}\right)\right\}\right].$$

Lemma A6. *As $n\to\infty$,*

$$E\left\{\left(\frac{b^{1/2}}{\Delta}\right)\hat{J}^{(1)}(t_0)\right\}\to-\sqrt{\frac{2}{\pi}}\frac{d_0}{t_0^{3/2}}.$$

Lemma A7. *As $n\to\infty$,*

$$Var\left\{\sqrt{\frac{nb^{5/2}}{\Delta^2}}\hat{J}^{(1)}(t_0)\right\}\to\frac{3}{2\sqrt{\pi}t_0^{5/2}}\left\{\frac{f(t_0^-)+f(t_0^+)}{2}\right\}.$$

Lemma A8. *If $\left|\hat{t}-t_0\right|=o_p\left(b^{1/2}\right)$, then, as $n\to\infty$,*

$$\left(\frac{b^{3/2}}{\Delta}\right)\hat{J}^{(2)}(\xi)\xrightarrow{p}-\sqrt{\frac{2}{\pi}}\frac{d_0}{t_0^{3/2}}$$

for any ξ between $\hat{t}$ and t_0.

Lemma A9. *As* $n \to \infty$, $E|H_i|^3 = O(\Delta^3/b^4)$.

A.6.1 Proof of Theorem 2

By a mean-value expansion of the first-order condition $\hat{J}^{(1)}\left(\hat{t}\right) = 0$, we have

$$\begin{aligned} 0 &= \hat{J}^{(1)}(t_0) + \hat{J}^{(2)}\left(\mathring{t}\right)\left(\hat{t} - t_0\right) \\ &= E\left\{\hat{J}^{(1)}(t_0)\right\} + \left[\hat{J}^{(1)}(t_0) - E\left\{\hat{J}^{(1)}(t_0)\right\}\right] + \hat{J}^{(2)}\left(\mathring{t}\right)\left(\hat{t} - t_0\right) \end{aligned} \tag{A24}$$

for some $\mathring{t}$ between $\hat{t}$ and t_0. Theorem 1 indicates that the condition for Lemma A8 is satisfied.

Rearranging (A24), we obtain

$$\begin{aligned} &\sqrt{\frac{n}{b^{1/2}}}\left[\hat{t} - t_0 - \left\{-\frac{E\left(\hat{J}^{(1)}(t_0)\right)}{\hat{J}^{(2)}\left(\mathring{t}\right)}\right\}\right] \\ &= -\sqrt{\frac{n}{b^{1/2}}}\left[\frac{\hat{J}^{(1)}(t_0) - E\left\{\hat{J}^{(1)}(t_0)\right\}}{\hat{J}^{(2)}\left(\mathring{t}\right)}\right]. \end{aligned} \tag{A25}$$

Then, by Lemmata A6 and A8, the leading bias term for $\hat{t}$ becomes

$$-\frac{E\left\{\hat{J}^{(1)}(t_0)\right\}}{\hat{J}^{(2)}\left(\mathring{t}\right)} = -b + o_p(b).$$

To demonstrate asymptotic normality of $\sqrt{n/b^{1/2}}\left(b^{3/2}/\Delta\right)\left[\hat{J}^{(1)}(t_0) - E\left\{\hat{J}^{(1)}(t_0)\right\}\right]$, we also check Lyapunov's condition. This quantity can be expressed as

$$\sum_{i=1}^{n} Y_i := \sum_{i=1}^{n}\sqrt{\frac{b^{5/2}}{n\Delta^2}}\left\{H_i - E(H_i)\right\}.$$

Then, by C_r-inequality, Jensen's inequality (due to the convexity of y^3 for $y \geq 0$) and Lemma A9,

$$E|Y_i|^3 \leq 8\left(\frac{b^{5/2}}{n\Delta^2}\right)^{3/2} E|H_i|^3 = O\left(n^{-3/2}b^{-1/4}\right).$$

It also follows from Lemma A7 that $Var(Y_i) = O(n^{-1})$. Therefore,

$$\frac{\sum_{i=1}^{n} E|Y_i|^3}{\left\{\sum_{i=1}^{n} Var(Y_i)\right\}^{3/2}} = O\left(\frac{1}{\sqrt{nb^{1/2}}}\right) \to 0,$$

and Lyapunov's condition is established.

Now we are allowed to employ a central limit theorem, in conjunction with Lemma A7, to obtain

$$\sqrt{\frac{n}{b^{1/2}}}\left(\frac{b^{3/2}}{\Delta}\right)\left[\hat{J}^{(1)}(t_0)-E\left\{\hat{J}^{(1)}(t_0)\right\}\right]\stackrel{d}{\to}N\left(0,\frac{3}{2\sqrt{\pi}t_0^{5/2}}\left\{\frac{f(t_0^-)+f(t_0^+)}{2}\right\}\right).$$

Then, it holds that for the right-hand side of (A25),

$$-\sqrt{\frac{n}{b^{1/2}}}\left[\frac{\left(b^{3/2}/\Delta\right)\left\{\hat{J}^{(1)}(t_0)-E\left(\hat{J}^{(1)}(t_0)\right)\right\}}{\left(b^{3/2}/\Delta\right)\hat{J}^{(2)}\left(\mathring{t}\right)}\right]\stackrel{d}{\to}N\left(0,\frac{3\sqrt{\pi}t_0^{1/2}}{4d_0^2}\left\{\frac{f(t_0^-)+f(t_0^+)}{2}\right\}\right)$$

by Lemma A8. This completes the proof. ■

Funding

This work was supported by the Japan Society for the Promotion of Science under grant numbers 19K01595 and 23K01340 (M. Hirukawa).

Declaration of Interest

The authors report that there are no competing interests to declare.

Declaration of Generative AI in Scientific Writing

The authors do not use generative artificial intelligence (AI) or AI-assisted technologies in the writing process.

Data Statement

The datasets used in Section 5 are openly available; see Section 5.1 for more details.

References

[1] Bader, B., J. Yan, and X. Zhang (2018): "Automated Threshold Selection for Extreme Value Analysis via Ordered Goodness-of-Fit Tests with Adjustment for False Discovery Rate," *Annals of Applied Statistics*, 12, 310 - 329.

[2] Chen, S. X. (2000): "Probability Density Function Estimation Using Gamma Kernels," *Annals of the Institute of Statistical Mathematics*, 52, 471 - 480.

[3] Chu, C. K., and P. E. Cheng (1996): "Estimation of Jump Points and Jump Values of a Density Function," *Statistica Sinica*, 6, 79 - 95.

[4] Clauset, A., C. R. Shalizi, and M. E. J. Newman (2009): "Power-Law Distributions in Empirical Data," *SIAM Review*, 51, 661 - 703.

[5] Cooray, K., and M. M. A. Ananda (2005): "Modeling Actuarial Data with a Composite Lognormal-Pareto Model," *Scandinavian Actuarial Journal*, 5, 321 - 334.

[6] Couallier, V. (1999): "Estimation Non Paramétrique d'une Discontinuité dans une Densité," *Comptes Rendus de l'Académie des Sciences - Série I*, 329, 633 - 636.

[7] Cowell, F. A., F. H. G. Ferreira, and J. A. Litchfield (1998): "Income Distribution in Brazil 1981-1990: Parametric and Non-Parametric Approaches," *Journal of Income Distribution*, 8, 63 - 76.

[8] Cunnane, C. (1973): "A Particular Comparison of Annual Maxima and Partial Duration Series Methods of Flood Frequency Prediction," *Journal of Hydrology*, 18, 257 - 271.

[9] Danielsson, J., L. M. Ergun, L. de Haan, and C. G. de Vries (2019): "Tail Index Estimation: Quantile-Driven Threshold Selection," Bank of Canada Staff Working Paper 2019-28.

[10] Davison, A. C., and R. L. Smith (1990): "Models for Exceedances over High Thresholds," *Journal of the Royal Statistical Society, Series B*, 52, 393 - 442.

[11] Delaigle, A., and I. Gijbels (2006): "Estimation of Boundary and Discontinuity Points in Deconvolution Problems," *Statistica Sinica*, 16, 773 - 788.

[12] Desmet, L., I. Gijbels, and A. Lambert (2010): "Estimation of Irregular Probability Densities," in J. Antoch, M. Hušková and P. K. Sen (eds.), *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in Honor of Professor Jana Jurečková.* Beachwood, OH: Institute of Mathematical Statistics, 46 - 61.

[13] Drees, H., A. Janßen, S. I. Resnick, and T. Wang (2020): "On a Minimum Distance Procedure for Threshold Selection in Tail Analysis," *SIAM Journal on Mathematics of Data Science*, 2, 75 - 102.

[14] DuMouchel, W. H. (1983): "Estimating the Stable Index α in Order to Measure Tail Thickness: A Critique," *Annals of Statistics*, 11, 1019 - 1031.

[15] Funke, B., and M. Hirukawa (2019): "Nonparametric Estimation and Testing on Discontinuity of Positive Supported Densities: A Kernel Truncation Approach," *Econometrics and Statistics*, 9, 156 - 170.

[16] Funke, B., and M. Hirukawa (2023): "Density Derivative Estimation Using Asymmetric Kernels," *Journal of Nonparametric Statistics*, forthcoming.

[17] Hirukawa, M., Murtazashvili I., and A. Prokhorov (2022): "Uniform Convergence Rates for Nonparametric Estimators Smoothed by the Beta Kernel," *Scandinavian Journal of Statistics*, 49, 1353 - 1382.

[18] Huh, J. (2002): "Nonparametric Discontinuity Point Estimation in Density or Density Derivatives," *Journal of the Korean Statistical Society*, 31, 261 - 276.

[19] Huh, J. (2012): "Bandwidth Selections Based on Cross-Validation for Estimation of a Discontinuity Point in Density" [in Korean], *Journal of the Korean Data & Information Science Society*, 23, 765 - 775.

[20] James, R., H. Leung, and A. Prokhorov (2023): "A Machine Learning Attack on Illegal Trading," *Journal of Banking and Finance*, 148, Article No.106735.

[21] Jenkins, S. P. (2017): "Pareto Models, Top Incomes and Recent Trends in UK Income Inequality," *Economica*, 84, 261 - 289.

[22] Joo, J.-H., and P. Qiu (2009): "Jump Detection in a Regression Curve and Its Derivative," *Technometrics*, 51, 289 - 305.

[23] Klugman, S. A., H. H. Panjer, and G. E. Willmot (2019): *Loss Models : From Data to Decisions,* 5th Edition. Hoboken, NJ: John Wiley & Sons.

[24] Loretan, M., and P. C. B. Phillips (1994): "Testing the Covariance Stationarity of Heavy-Tailed Time Series: An Overview of the Theory with Applications to Several Financial Datasets," *Journal of Empirical Finance*, 1, 211 - 248.

[25] Markovitch, N. M., and U. R. Krieger (2000): "Nonparametric Estimation of Long-Tailed Density Functions and Its Application to the Analysis of World Wide Web Traffic," *Performance Evaluation*, 42, 205 - 222.

[26] Marron, J. S. (1985): "An Asymptotically Efficient Solution to the Bandwidth Problem of Kernel Density Estimation," *Annals of Statistics*, 13, 1011 - 1023.

[27] McNeal, A. J. (1997): "Estimating the Tails of Loss Severity Distributions Using Extreme Value Theory," *ASTIN Bulletin*, 27, 117 - 137.

[28] Newey, W. K., and D. L. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in R. F. Engle and D. L. McFadden (eds.), *Handbook of Econometrics, Volume IV.* Amsterdam: Elsevier, 2111 - 2245.

[29] Pagurova, V. I. (1965): "An Asymptotic Formula for the Incomplete Gamma Function," *USSR Computational Mathematics and Mathematical Physics*, 5, 162 - 166.

[30] Resnick, S. I. (1997): "Discussion of the Danish Data on Large Fire Insurance Losses," *ASTIN Bulletin*, 27, 139 - 151.

[31] Reynkens, T., R. Verbelen, J. Beirlant, and K. Antonio (2017): "Modelling Censored Losses Using Splicing: A Global Fit Strategy with Mixed Erlang and Extreme Value Distributions," *Insurance: Mathematics and Economics*, 77, 65 - 77.

[32] Sasaki, Y., and Y. Wang (2023): "Extreme Changes in Changes," *Journal of Business & Economic Statistics*, forthcoming.

[33] Scarrott, C., and A. MacDonald (2012): "A Review of Extreme Value Threshold Estimation and Uncertainty Quantification," *REVSTAT - Statistical Journal*, 10, 33 - 60.

[34] Scollnik, D. P. M., and C. Sun (2012): "Modeling with Weibull-Pareto Models," *North American Actuarial Journal*, 16, 260 - 272.

[35] Van der Vaart, A. W., and J. A. Wellner (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics.* New York: Springer-Verlag.

[36] Van Es, B. (1991): "Likelihood Cross-Validation Bandwidth Selection for Nonparametric Kernel Density Estimators," *Journal of Nonparametric Statistics*, 1, 83 - 110.

[37] Wu, J. S., and C. K. Chu (1993a): "Kernel-Type Estimators of Jump Points and Jump Values of a Regression Function," *Annals of Statistics*, 21, 1545 - 1566.

[38] Wu, J. S., and C. K. Chu (1993b): "Nonparametric Function Estimation and Bandwidth Selection for Discontinuous Regression Functions," *Statistica Sinica*, 3, 557 - 576.

Figure 1: The Diagnostic Function and Its Approximations

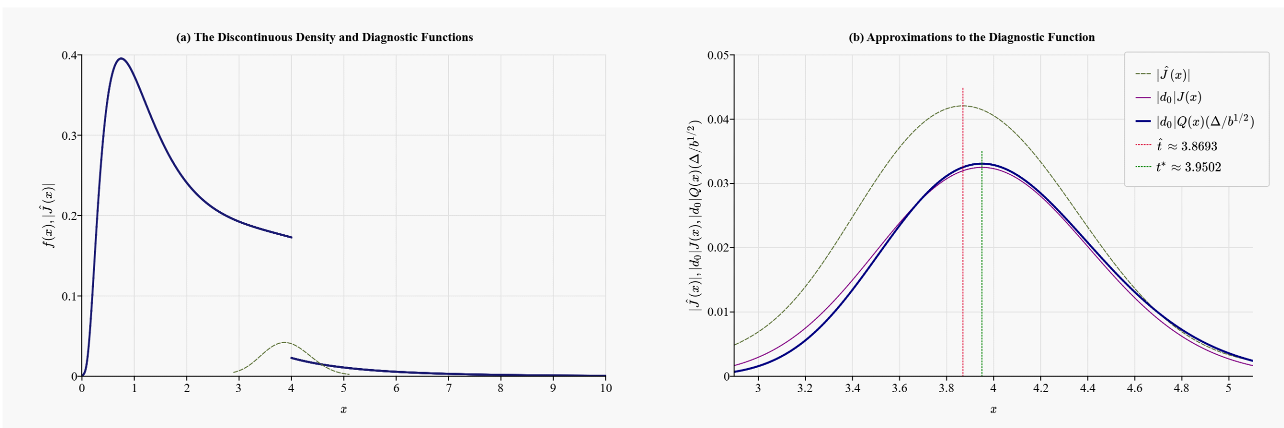

Table 1: Characteristic Numbers of Underlying Distributions

Model		Distribution	Mode	c_L	$f\left(t_0^-\right)$	$f\left(t_0^+\right)$	d_0
1	A	*Log-Normal + Quadratic* $(D = 1/4)$	0.6959	0.9659	0.1728	0.0228	0.1500
	B	*Log-Normal + Quadratic* $(D = 3/22)$	0.6959	0.9583	0.1279	0.0279	0.1000
2	A	*Splicing with Weibull & GPD*	2.4023	0.9539	0.1063	0.0115	0.0948
	B	*Splicing with Weibull & Translated Weibull*	2.4023	0.9539	0.1063	0.0115	0.0948

Table 2: Monte Carlo Results

		n = 250						n = 500					
Estimator	α	Bias	SD	RMSE	$\tilde{h}$ or $\tilde{b}$		$\#\{I_0\}$	Bias	SD	RMSE	$\tilde{h}$ or $\tilde{b}$		$\#\{I_0\}$
					Model 1-A: Log-Normal + Quadratic ($D = 1/4$)								
CC	–	0.2260	0.6340	0.6731	0.0116	(0.0092)	–	0.0487	0.5775	0.5796	0.0073	(0.0037)	–
KS	–	−0.8770	0.5392	1.0295	–	(–)	644	−0.5234	0.7046	0.8778	–	(–)	707
Q-MAD	–	−0.2990	0.1569	0.3376	–	(–)	1000	−0.3229	0.1962	0.3778	–	(–)	997
Q-SUP	–	−0.1039	0.4014	0.4146	–	(–)	976	0.0800	0.6708	0.6755	–	(–)	913
AEB	–	0.6615	0.8305	1.0618	–	(–)	729	0.7181	0.5745	0.9197	–	(–)	708
SG-LS	0.55	−0.6680	0.4375	0.7985	0.3302	(0.2182)	–	−0.7461	0.4079	0.8503	0.3716	(0.2027)	–
	0.60	−0.6578	0.4423	0.7927	0.3263	(0.2206)	–	−0.7351	0.4149	0.8441	0.3666	(0.2066)	–
	0.65	−0.6501	0.4444	0.7874	0.3221	(0.2220)	–	−0.7185	0.4243	0.8345	0.3585	(0.2113)	–
	0.70	−0.6333	0.4497	0.7767	0.3141	(0.2243)	–	−0.7053	0.4310	0.8266	0.3518	(0.2147)	–
SG-LS-BC	0.55	−0.3378	0.2250	0.4058	–	(–)	–	−0.3745	0.2078	0.4283	–	(–)	–
	0.60	−0.3315	0.2273	0.4020	–	(–)	–	−0.3685	0.2107	0.4245	–	(–)	–
	0.65	−0.3280	0.2282	0.3996	–	(–)	–	−0.3600	0.2155	0.4196	–	(–)	–
	0.70	−0.3192	0.2313	0.3942	–	(–)	–	−0.3535	0.2192	0.4160	–	(–)	–
SG-L	0.55	−0.5628	0.4462	0.7182	0.2725	(0.2076)	–	−0.6063	0.4269	0.7415	0.2914	(0.1987)	–
	0.60	−0.5118	0.4494	0.6811	0.2449	(0.2052)	–	−0.5363	0.4320	0.6886	0.2548	(0.1989)	–
	0.65	−0.4758	0.4468	0.6527	0.2262	(0.2011)	–	−0.4898	0.4288	0.6510	0.2311	(0.1939)	–
	0.70	−0.4590	0.4433	0.6381	0.2160	(0.1966)	–	−0.4608	0.4219	0.6247	0.2148	(0.1877)	–
SG-L-BC	0.55	−0.2904	0.2474	0.3814	–	(–)	–	−0.3149	0.2357	0.3933	–	(–)	–
	0.60	−0.2668	0.2546	0.3688	–	(–)	–	−0.2814	0.2417	0.3710	–	(–)	–
	0.65	−0.2497	0.2575	0.3586	–	(–)	–	−0.2587	0.2437	0.3554	–	(–)	–
	0.70	−0.2430	0.2598	0.3557	–	(–)	–	−0.2460	0.2437	0.3463	–	(–)	–
SG-ML	0.55	0.1375	0.4524	0.4728	0.0315	(0.0210)	–	0.3265	0.5146	0.6094	0.0249	(0.0183)	–
	0.60	0.0465	0.3836	0.3864	0.0335	(0.0196)	–	0.1547	0.4346	0.4613	0.0297	(0.0173)	–
	0.65	−0.0035	0.3204	0.3204	0.0351	(0.0188)	–	0.0608	0.3497	0.3550	0.0326	(0.0166)	–
	0.70	−0.0352	0.2737	0.2759	0.0364	(0.0187)	–	0.0101	0.2831	0.2832	0.0342	(0.0157)	–
SG-ML-BC	0.55	0.1690	0.4378	0.4693	–	(–)	–	0.3514	0.4993	0.6105	–	(–)	–
	0.60	0.0800	0.3714	0.3799	–	(–)	–	0.1845	0.4216	0.4602	–	(–)	–
	0.65	0.0316	0.3098	0.3114	–	(–)	–	0.0933	0.3390	0.3516	–	(–)	–
	0.70	0.0011	0.2639	0.2639	–	(–)	–	0.0443	0.2741	0.2777	–	(–)	–
					Model 1-B: Log-Normal + Quadratic ($D = 3/22$)								
CC	–	0.3200	0.6223	0.6998	0.0139	(0.0126)	–	0.1409	0.5942	0.6107	0.0086	(0.0056)	–
KS	–	−0.9832	0.6012	1.1525	–	(–)	594	−0.6664	0.6955	0.9632	–	(–)	660
Q-MAD	–	−0.3227	0.2688	0.4200	–	(–)	993	−0.3638	0.2716	0.4540	–	(–)	995
Q-SUP	–	−0.0720	0.6380	0.6421	–	(–)	925	0.1549	0.9658	0.9782	–	(–)	837
AEB	–	0.9418	0.8936	1.2983	–	(–)	602	1.0100	0.6164	1.1832	–	(–)	535
SG-LS	0.55	−0.9517	0.2014	0.9727	0.4740	(0.1070)	–	−0.9940	0.0711	0.9965	0.4967	(0.0392)	–
	0.60	−0.9472	0.2109	0.9703	0.4717	(0.1115)	–	−0.9923	0.0808	0.9956	0.4958	(0.0444)	–
	0.65	−0.9363	0.2304	0.9642	0.4658	(0.1220)	–	−0.9914	0.0856	0.9951	0.4953	(0.0468)	–
	0.70	−0.9303	0.2398	0.9607	0.4625	(0.1273)	–	−0.9856	0.1130	0.9921	0.4925	(0.0590)	–
SG-LS-BC	0.55	−0.4777	0.0973	0.4875	–	(–)	–	−0.4973	0.0321	0.4983	–	(–)	–
	0.60	−0.4755	0.1029	0.4865	–	(–)	–	−0.4965	0.0367	0.4979	–	(–)	–
	0.65	−0.4705	0.1122	0.4837	–	(–)	–	−0.4961	0.0390	0.4976	–	(–)	–
	0.70	−0.4679	0.1165	0.4822	–	(–)	–	−0.4931	0.0545	0.4961	–	(–)	–
SG-L	0.55	−0.9171	0.2572	0.9525	0.4554	(0.1321)	–	−0.9809	0.1295	0.9894	0.4897	(0.0660)	–
	0.60	−0.8958	0.2857	0.9403	0.4437	(0.1469)	–	−0.9652	0.1737	0.9807	0.4819	(0.0875)	–
	0.65	−0.8789	0.3064	0.9308	0.4347	(0.1556)	–	−0.9605	0.1840	0.9780	0.4789	(0.0939)	–
	0.70	−0.8711	0.3138	0.9259	0.4304	(0.1592)	–	−0.9510	0.2031	0.9724	0.4740	(0.1029)	–
SG-L-BC	0.55	−0.4617	0.1294	0.4794	–	(–)	–	−0.4912	0.0649	0.4955	–	(–)	–
	0.60	−0.4521	0.1439	0.4745	–	(–)	–	−0.4833	0.0878	0.4912	–	(–)	–
	0.65	−0.4443	0.1573	0.4713	–	(–)	–	−0.4816	0.0925	0.4904	–	(–)	–
	0.70	−0.4407	0.1616	0.4694	–	(–)	–	−0.4770	0.1029	0.4880	–	(–)	–
SG-ML	0.55	0.0586	0.5252	0.5285	0.0379	(0.0278)	–	0.2704	0.5556	0.6179	0.0308	(0.0246)	–
	0.60	−0.0557	0.4497	0.4532	0.0423	(0.0279)	–	0.0642	0.4729	0.4772	0.0398	(0.0253)	–
	0.65	−0.1146	0.3854	0.4021	0.0450	(0.0275)	–	−0.0434	0.3983	0.4007	0.0454	(0.0259)	–
	0.70	−0.1498	0.3557	0.3859	0.0468	(0.0275)	–	−0.0943	0.3433	0.3560	0.0489	(0.0260)	–
SG-ML-BC	0.55	0.0965	0.5064	0.5155	–	(–)	–	0.3012	0.5360	0.6149	–	(–)	–
	0.60	−0.0133	0.4332	0.4334	–	(–)	–	0.1040	0.4549	0.4667	–	(–)	–
	0.65	−0.0696	0.3706	0.3771	–	(–)	–	0.0020	0.3822	0.3822	–	(–)	–
	0.70	−0.1030	0.3408	0.3560	–	(–)	–	−0.0455	0.3280	0.3311	–	(–)	–

Table 2: *(continued)*

Estimator	α	$n=250$ Bias	SD	RMSE	$\hat{h}$ or $\hat{b}$		$\#\{I_0\}$	$n=500$ Bias	SD	RMSE	$\hat{h}$ or $\hat{b}$		$\#\{I_0\}$
Model 2-A: Splicing with Weibull & GPD													
CC	–	0.3261	0.6679	0.7432	0.0104	(0.0228)	–	0.1999	0.6443	0.6746	0.0060	(0.0023)	–
KS	–	−1.3273	0.9856	1.6533	–	(–)	91	−0.6556	1.8320	1.9458	–	(–)	162
Q-MAD	–	−0.2051	0.7303	0.7585	–	(–)	975	0.1488	1.4884	1.4958	–	(–)	905
Q-SUP	–	0.3085	1.6298	1.6588	–	(–)	864	1.0775	2.9302	3.1221	–	(–)	784
AEB	–	4.3704	2.9114	5.2514	–	(–)	81	4.2648	2.0070	4.7135	–	(–)	29
SG-LS	0.55	−0.6238	0.1250	0.6362	0.1503	(0.0644)	–	−0.6202	0.0914	0.6269	0.1528	(0.0450)	–
	0.60	−0.6268	0.1260	0.6393	0.1550	(0.0675)	–	−0.6237	0.0914	0.6304	0.1578	(0.0473)	–
	0.65	−0.6290	0.1272	0.6417	0.1583	(0.0705)	–	−0.6265	0.0919	0.6333	0.1613	(0.0495)	–
	0.70	−0.6308	0.1282	0.6437	0.1608	(0.0728)	–	−0.6285	0.0935	0.6354	0.1635	(0.0514)	–
SG-LS-BC	0.55	−0.4736	0.0848	0.4811	–	(–)	–	−0.4674	0.0615	0.4714	–	(–)	–
	0.60	−0.4717	0.0841	0.4792	–	(–)	–	−0.4659	0.0601	0.4698	–	(–)	–
	0.65	−0.4708	0.0840	0.4782	–	(–)	–	−0.4652	0.0592	0.4690	–	(–)	–
	0.70	−0.4700	0.0839	0.4775	–	(–)	–	−0.4649	0.0592	0.4687	–	(–)	–
SG-L	0.55	−0.6196	0.1186	0.6309	0.1848	(0.0668)	–	−0.6224	0.0831	0.6279	0.1862	(0.0470)	–
	0.60	−0.6216	0.1233	0.6337	0.1767	(0.0660)	–	−0.6242	0.0871	0.6303	0.1779	(0.0466)	–
	0.65	−0.6231	0.1282	0.6362	0.1718	(0.0648)	–	−0.6259	0.0898	0.6323	0.1728	(0.0459)	–
	0.70	−0.6239	0.1346	0.6383	0.1692	(0.0638)	–	−0.6275	0.0915	0.6342	0.1701	(0.0452)	–
SG-L-BC	0.55	−0.4348	0.0715	0.4407	–	(–)	–	−0.4362	0.0491	0.4389	–	(–)	–
	0.60	−0.4449	0.0759	0.4513	–	(–)	–	−0.4464	0.0522	0.4494	–	(–)	–
	0.65	−0.4514	0.0811	0.4586	–	(–)	–	−0.4531	0.0547	0.4564	–	(–)	–
	0.70	−0.4547	0.0881	0.4632	–	(–)	–	−0.4574	0.0566	0.4609	–	(–)	–
SG-ML	0.55	−0.4484	0.3354	0.5600	0.0333	(0.0082)	–	−0.4259	0.3665	0.5619	0.0331	(0.0079)	–
	0.60	−0.4518	0.3036	0.5443	0.0332	(0.0074)	–	−0.4568	0.2998	0.5464	0.0332	(0.0064)	–
	0.65	−0.4506	0.2968	0.5395	0.0335	(0.0073)	–	−0.4697	0.2632	0.5384	0.0337	(0.0057)	–
	0.70	−0.4334	0.3124	0.5342	0.0340	(0.0075)	–	−0.4733	0.2508	0.5356	0.0343	(0.0057)	–
SG-ML-BC	0.55	−0.4152	0.3313	0.5311	–	(–)	–	−0.3928	0.3612	0.5336	–	(–)	–
	0.60	−0.4186	0.3005	0.5153	–	(–)	–	−0.4237	0.2964	0.5171	–	(–)	–
	0.65	−0.4171	0.2940	0.5102	–	(–)	–	−0.4360	0.2609	0.5081	–	(–)	–
	0.70	−0.3994	0.3092	0.5051	–	(–)	–	−0.4390	0.2487	0.5046	–	(–)	–
Model 2-B: Splicing with Weibull & Translated Weibull													
CC	–	0.3302	0.6788	0.7548	0.0116	(0.0351)	–	0.1609	0.6716	0.6906	0.0058	(0.0020)	–
KS	–	−1.4714	0.9325	1.7420	–	(–)	124	−1.0546	3.0694	3.2455	–	(–)	173
Q-MAD	–	−0.2656	1.3378	1.3639	–	(–)	972	0.4715	4.2932	4.3190	–	(–)	936
Q-SUP	–	0.1266	1.8973	1.9015	–	(–)	921	2.7196	16.7906	17.0094	–	(–)	801
AEB	–	6.4959	13.3971	14.8889	–	(–)	398	3.7522	5.4395	6.6081	–	(–)	376
SG-LS	0.55	−0.6123	0.1278	0.6255	0.1384	(0.0591)	–	−0.5991	0.0968	0.6068	0.1364	(0.0427)	–
	0.60	−0.6154	0.1283	0.6287	0.1425	(0.0619)	–	−0.6024	0.0974	0.6103	0.1405	(0.0448)	–
	0.65	−0.6175	0.1296	0.6309	0.1453	(0.0645)	–	−0.6050	0.0982	0.6130	0.1434	(0.0467)	–
	0.70	−0.6193	0.1307	0.6329	0.1473	(0.0665)	–	−0.6069	0.0993	0.6150	0.1453	(0.0484)	–
SG-LS-BC	0.55	−0.4739	0.0899	0.4824	–	(–)	–	−0.4626	0.0670	0.4675	–	(–)	–
	0.60	−0.4730	0.0888	0.4812	–	(–)	–	−0.4620	0.0658	0.4667	–	(–)	–
	0.65	−0.4721	0.0883	0.4803	–	(–)	–	−0.4616	0.0652	0.4662	–	(–)	–
	0.70	−0.4720	0.0881	0.4801	–	(–)	–	−0.4616	0.0649	0.4662	–	(–)	–
SG-L	0.55	−0.6036	0.1328	0.6181	0.1502	(0.0621)	–	−0.5977	0.0968	0.6055	0.1477	(0.0443)	–
	0.60	−0.6028	0.1397	0.6188	0.1428	(0.0614)	–	−0.5969	0.1013	0.6054	0.1404	(0.0438)	–
	0.65	−0.6022	0.1461	0.6196	0.1386	(0.0605)	–	−0.5969	0.1041	0.6059	0.1364	(0.0430)	–
	0.70	−0.6023	0.1518	0.6211	0.1368	(0.0596)	–	−0.5976	0.1056	0.6068	0.1346	(0.0423)	–
SG-L-BC	0.55	−0.4534	0.0857	0.4615	–	(–)	–	−0.4500	0.0611	0.4541	–	(–)	–
	0.60	−0.4600	0.0925	0.4692	–	(–)	–	−0.4565	0.0656	0.4612	–	(–)	–
	0.65	−0.4635	0.0993	0.4741	–	(–)	–	−0.4605	0.0687	0.4656	–	(–)	–
	0.70	−0.4654	0.1061	0.4774	–	(–)	–	−0.4630	0.0706	0.4683	–	(–)	–
SG-ML	0.55	−0.4056	0.3411	0.5300	0.0325	(0.0087)	–	−0.3769	0.3688	0.5273	0.0350	(0.0101)	–
	0.60	−0.4099	0.3227	0.5217	0.0326	(0.0082)	–	−0.4048	0.3125	0.5114	0.0355	(0.0090)	–
	0.65	−0.3956	0.3177	0.5074	0.0329	(0.0078)	–	−0.4224	0.2595	0.4958	0.0359	(0.0080)	–
	0.70	−0.3742	0.3280	0.4976	0.0330	(0.0077)	–	−0.4118	0.2559	0.4848	0.0359	(0.0078)	–
SG-ML-BC	0.55	−0.3732	0.3370	0.5028	–	(–)	–	−0.3418	0.3625	0.4982	–	(–)	–
	0.60	−0.3774	0.3194	0.4944	–	(–)	–	−0.3693	0.3079	0.4808	–	(–)	–
	0.65	−0.3628	0.3147	0.4802	–	(–)	–	−0.3865	0.2567	0.4640	–	(–)	–
	0.70	−0.3412	0.3249	0.4711	–	(–)	–	−0.3758	0.2532	0.4532	–	(–)	–

Note: "Bias", "SD" and "RMSE" are biases, standard deviations and root mean squared errors of jump location or threshold estimates over 1000 Monte Carlo samples, respectively. For kernel-smoothed jump location estimation procedures, Monte Carlo averages and standard deviations (in parentheses) of CV tuning parameters are presented under the heading "$\hat{h}$ or $\hat{b}$". For automated threshold detection methods, counts of threshold estimates falling into I_0 are reported under the heading "$\#\{I_0\}$".

Table 3: Descriptive Statistics of Datasets

Data	n	Mean	SD	SK	Min.	Q1	Q2	Q3	90%	95%	99%	Max.
(A) Danish Fire Insurance Losses (in millions of Danish kroner)												
Original	2,492	3.063	7.975	19.884	0.313	1.157	1.634	2.646	5.080	8.406	24.614	263.250
(B) US Male Hourly Wages (in US dollars)												
Original	54,769	6.136	2.972	3.013	0.500	3.850	5.630	8.000	9.980	11.000	14.793	99.990
Downsized	2,709	6.080	2.858	1.525	1.000	3.750	5.550	8.000	9.954	11.000	14.640	40.000

Note: n = sample size; Mean = average; SD = standard deviation; SK = skewness; Min. = minimum value; Q1 = first quartile; Q2 = median (i.e., second quartile); Q3 = third quartile; 90% = 90% quantile; 95% = 95% quantile; 99% = 99% quantile; and Max. = maximum value.

Table 4: Results of Threshold Estimation

Data	Estimator	Estimate of t_0	Estimator	α	I_0	Estimate of t_0	$\hat{h}$ or $\hat{b}$
(A) Danish Fire Insurance Losses (in millions of Danish kroner)							
Original	KS	1.375	CC	–	$[1, 30]$	30.000	0.005
	Q-MAD	29.037					
	Q-SUP	11.123	SG-ML	0.70	$[1, 30]$	1.861	0.235
	AEB	25.288	SG-ML-BC	–	–	2.096	–
(B) US Male Hourly Wages (in US dollars)							
Downsized	KS	10.400	CC	–	$[5, 15]$	14.998	0.005
	Q-MAD	10.000					
	Q-SUP	10.000	SG-ML	0.70	$[5, 15]$	9.524	0.140
	AEB	13.500	SG-ML-BC	–	–	9.664	–