The Econometrics Journal

Econometrics Journal (2013), volume **16**, pp. 430–462.
doi: 10.1111/ectj.12012

Asymptotics for threshold regression under general conditions

PING YU[†] AND YONGQIANG ZHAO[‡]

[†]*Department of Economics, University of Auckland, Auckland 1010, New Zealand.*
E-mail: p.yu@auckland.ac.nz

[‡]*Department of Mathematics, University of Wisconsin-Madison, Madison, WI 53706, USA.*
E-mail: yqzhao@math.wisc.edu

First version received: March 2012; final version accepted: June 2013

Summary The inference of the threshold point in threshold models critically depends on the assumption that the density of the threshold variable at the true threshold point is continuous and bounded away from zero and infinity. However, violation of this assumption may arise in several econometric contexts such as treatment effects evaluation. This paper presents a thorough characterisation of the asymptotic distributions in the least-squares estimation of such abnormal cases. First, the asymptotic results on the threshold point are different from the conventional case. For example, any convergence rate between zero and infinity is possible; the asymptotic distribution can be discrete, continuous or a mixture of discrete and continuous; the weak limits of the localised objective functions can be non-homogeneous instead of homogeneous compound Poisson processes. Second, this paper distinguishes threshold regression from structural change models by studying a problem unique in threshold regression. Third, the asymptotic distributions of regular parameters are not affected by estimation of the threshold point irrespective of the density of the threshold variable. Numerical calculations and simulation results confirm the theoretical analysis, and the density of the threshold variable in an application is checked to illustrate the relevance of the study in this paper.

Keywords: *Discrete asymptotic distribution, Extreme value type III distribution, Local information, Non-homogeneous Poisson process, Rapidly varying, Regularly varying, Slowly varying, Stochastic equicontinuity failure, Strong identification, Threshold regression, Treatment effect, Weak identification.*

1. INTRODUCTION

Since the pioneering work by Tong (1978, 1983), threshold models get much popularity in current applied statistical and econometric practice. An encyclopedic survey on this kind of models is available in Tong (1990) and a selective review of the history of threshold models is given by Tong (2011); see also Lee and Seo (2008) and Hansen (2011) for a summary of applications especially in economics. Threshold models are naturally motivated in regression analysis when we suspect the population can be divided into different groups whose regression coefficients are different. Sometimes these groups are selected on categorical variables, such as gender or race, while in other cases they are selected based on continuous variables, such as income or GDP levels. Such continuous variables are called threshold variables in threshold regression.

The typical setup of threshold regression is

$$y = \begin{cases} x'\beta_1 + \sigma_1 e, & q \le \gamma; \\ x'\beta_2 + \sigma_2 e, & q > \gamma; \end{cases} \tag{1.1}$$

where q is the threshold variable with pdf $f(q)$ and cdf $F(q)$, γ is the unknown threshold point, $x \in \mathbb{R}^k$ includes characteristics of the population, $\beta \equiv (\beta_1', \beta_2')' \in \mathbb{R}^{2k}$ and $\sigma \equiv (\sigma_1, \sigma_2)'$ are threshold parameters on the mean and variance in the two groups. We set $E[e^2] = 1$ as a normalisation of the error variance and allow for conditional heteroscedasticity. All the other variables have the same definitions as in the linear regression framework. Usually, the parameters of interest are $\theta \equiv (\beta', \gamma)'$. Under the assumption that $E[e|x, q] = 0$, a popular estimator is the least-squares estimator (LSE). For notational simplicity, we use $(\widehat{\beta}_1', \widehat{\beta}_2', \widehat{\gamma})'$ to denote the LSE of θ.

All existing literature on threshold regression assumes that $f(q)$ is continuous and $0 < \underline{f} \le f(q) \le \overline{f} < \infty$ for q in a neighbourhood of γ_0. Researchers use this assumption for a number of reasons (e.g. it is natural), an important technical one among which is that the asymptotic distribution of the estimator of γ is relatively easy to derive under this assumption. It is well known that the inference on γ critically depends on the local behaviour of $f(\cdot)$ around γ_0. This point can be easily observed in two popular frameworks of threshold effects. The first framework is introduced by Chan (1993) in non-linear time series environments, where $(\beta_1', \sigma_1)' - (\beta_2', \sigma_2)'$ is a fixed constant. In this framework, the convergence rate of $\widehat{\gamma}$ is n, and the asymptotic distribution is continuous and related with a two-sided homogeneous compound Poisson process with intensity $f(\gamma_0)$. The second framework is introduced by Hansen (2000), where no threshold effect on variance exists and the threshold effect in mean diminishes asymptotically. In this framework, the convergence rate of $\widehat{\gamma}$ is slower than n, while the asymptotic distribution is still continuous and related with a two-sided Brownian motion with a scale factor $f(\gamma_0)^{-1}$. So in both frameworks, the asymptotic distribution of $\widehat{\gamma}$ is fragile to the density of q around γ_0. To identify γ and make the inference on γ non-degenerate, the assumption that $f(q)$ is continuous and bounded from zero and infinity seems natural and necessary although this assumption is likely to be violated in practice. Also, in both frameworks, the inference on regular parameters β is not affected by the estimation of γ, but it is unknown whether this still holds when the assumption on $f(q)$ fails.

In this paper, we consider the asymptotic behaviour of the LSE when the density of the threshold variable is not continuous and/or not between zero and infinity. While these cases may seem pathological, they are, in fact, far from it. The following example illustrates this point. Suppose we want to measure the treatment effects based on (1.1). Without loss of generality, we assume that $q > \gamma$ is associated with the treated group and $q \le \gamma$ the control group. Usually, we are interested in two pieces of information. The first piece of information is related with the treatment assignment, that is, the value of γ. The second piece of information relates to the treatment effects. For example, the population-average treatment effect $PATE \equiv E[Y_1 - Y_0]$ can be estimated by the sample-average treatment effect $SATE \equiv n^{-1} \sum_{i=1}^n x_i'(\widehat{\beta}_2 - \widehat{\beta}_1)$, where Y_d is the potential outcome under treatment d for $d = 0$ and 1; the population-average treatment effect for the treated $PATT \equiv E[Y_1 - Y_0|D = 1]$ can be estimated by the sample-average treatment effect for the treated $SATT \equiv n^{-1} \sum_{i=1}^n x_i'(\widehat{\beta}_2 - \widehat{\beta}_1)\mathbf{1}(q_i > \widehat{\gamma})$, where D is the treatment status, and $\mathbf{1}(\cdot)$ is the indicator function. The treatment assignment may depend on various considerations. For example, if it is a beneficial treatment, such as a job training programme based on the income level, the policy maker may choose γ such that locally (around

γ_0) most people benefit to popularise the programme; in other words, γ is close to the mode of q's distribution. If it is a hurtful treatment, such as a tax increase plan, the policy maker may choose γ such that locally few people are hurt to avoid objection of the plan. In the former case, we can approximate the model by assuming $f(\gamma_0) = \infty$, and in the latter case, by assuming $f(\gamma_0) = 0$. Even if we assume $f(\cdot)$ is bounded from zero and infinity in a neighbourhood of γ_0, it may still be discontinuous at γ_0 due to, say, the policy leak about the threshold.

We briefly discuss three other reasons for the discontinuity of $f(q)$ at γ_0. First, non-response and attrition. A specific example is Hoekstra (2009) who estimates the effects of college quality on earnings. In this scenario, q is the admission score and y is the earnings many years later. His dataset, however, does not permit to trace individuals who moved out of the state either for study or for work so that the information of those individuals is missing. Differential attrition below γ_0 and above γ_0 may induce the discontinuity of $f(q)$ at γ_0. Second, difference in data collection schemes. Different data collection schemes may have been used for individuals above γ_0 and below γ_0. Suppose individuals with a certain medical condition, for example, heart attack, register at a hospital and have their medical status q measured. Those with $q > \gamma_0$ are hospitalised and remain under intensive surveillance for several days, while those with $q \leq \gamma_0$ receive a certain prescription drug, but are not hospitalised. To estimate the effects of hospitalisation on health status we need follow-up data on both groups. Whereas extensive data are routinely collected in the hospital for the in-patient group, data collection for the out-patient group requires expensive home visits. To reduce data collection costs, one might join in a multi-purpose survey, which over-samples individuals with high-risk conditions such that $\lim_{q \downarrow \gamma_0} f(q) > \lim_{q \uparrow \gamma_0} f(q)$.[1] Third, manipulation or sorting of q. For example, McCrary (2008) studies the roll call voting in the House of Representatives, where sorting is both expected and found such that $f(q)$ is discontinuous at γ_0. He also provides a test of the continuity of $f(q)$ at γ_0 when $0 < \underline{f} \leq f(q) \leq \overline{f} < \infty$ for q in a neighbourhood of γ_0. In all cases, one may want to know how the local behaviour of $f(q)$ around γ_0 affects estimation of the treatment assignment, and how this estimation affects the treatment evaluation.

Similar problems as in the above examples may happen in all applications of threshold regression. Actually, similar problems may arise as long as the asymptotic distribution is related with the value of a density at a fixed point. For example, it is well known that in quantile regression, the asymptotic distribution of the estimator involves the error density at zero. Knight (1997, 1998, 2008) discusses the L_1 asymptotics when the error density takes various shapes, especially, when it is zero or infinity or discontinuous at the fixed point 0; see also Rogers (2001) and Cho et al. (2010) for more discussions in the time series context. Han et al. (2011) develop a test for the presence of an infinite density at the median and apply it to stock returns of leading companies across major US industry groups. Their results confirm the presence of infinite density at the median as a new significant empirical evidence for stock return distributions.

The contributions of this paper are threefold. First, it provides a thorough characterisation of the asymptotic behaviour of $\widehat{\gamma}$ when the usual assumption on $f(q)$ fails. We use the discontinuous framework of Chan (1993) with i.i.d. data. The results are very different from those in the literature. For example, the convergence rate can be different from n; the asymptotic distribution may not be continuous; and the weak limits of the localised objective functions need not be homogeneous compound Poisson processes. Second, our paper distinguishes threshold regression from structural change models. It is commonly believed that threshold regression and

[1] See Appendix A of Frölich (2007) for related discussions in regression discontinuity designs (RDDs).

structural change models are closely related. For example, the asymptotic distributions of $\widehat{\gamma}$ in the structural change literature such as Bai (1997) and the threshold regression literature such as Chan (1993) and Hansen (2000) are similar. This is not surprising since structural change models can be essentially treated as threshold regression with q following a uniform distribution on [0, 1] and being independent of (x, e). This work complements this general view by studying a problem unique in threshold regression. Third, our paper shows that the inference on β is the same as in the case where γ_0 is known even under the non-standard assumptions on $f(q)$. An immediate corollary of this result is that we can safely conduct inference on functionals of β without worrying about the effect of non-standard asymptotic behaviours of $\widehat{\gamma}$. This is essentially because only local information around the threshold point is informative to γ, while the inference on β relies on the global information such as the sample mean which is independent of the local information in some sense. Yu (2008, 2012) provides a detailed analysis on this point under usual assumptions on $f(q)$, and this paper extends that result to more general cases.

The rest of this paper is organised as follows. Because the main difference of this paper from the existing literature lies in the specification of $f(q)$ around γ_0, Section 2 discusses the effects of $f(q)$ on the asymptotic distribution of $\widehat{\gamma}$ in a simple setup where the only randomness is from q. Section 3 presents the asymptotic distribution of the LSE of γ and β. The latter is shown to be independent of and not affected by $\widehat{\gamma}$. Section 4 includes some numerical examples, Monte Carlo simulation results and application in an economic growth model. Finally, Section 5 concludes and discusses some unsolved problems, such as the confidence interval construction and specification testing. All proofs and lemmas are given in Appendices A and B, respectively, and Supporting Information is available with the Online version of the article.

A word on notations and terms: $\mathbb{Z}_+$ is the set of non-negative integers, and $\mathbb{N} = \mathbb{Z}_+ \backslash \{0\}$. $\|\cdot\|$ is the Euclidean norm. The letter C is used to denote a generic positive constant, which need not be the same in each occurrence. For two real numbers x and y, $x \vee y$ is the larger of them, and $x \wedge y$ is the smaller one. $[x]$ is the largest integer no greater than x. ℓ is always used for indicating the two regimes in (1.1), so is not written out explicitly as '$\ell = 1, 2$' throughout the paper. The case with the specification of $f(\cdot)$ in the existing literature is referred to as 'the classical case'.

2. THE SETUP AND THE LOCAL BEHAVIOUR OF THE THRESHOLD VARIABLE

In this section, we first define the LSE in threshold regression, then show the importance of $f(q)$ around γ_0 on the asymptotic distribution of $\widehat{\gamma}$ by a simple threshold model and conclude with a general classification theorem of the local behaviour of $f(q)$ around γ_0.

2.1. The Least Squares Estimator

For a random sample $\{w_i\}_{i=1}^n$ with $w_i = (y_i, x_i', q_i)'$, the LSE of γ is usually defined by a profiled procedure:

$$\widehat{\gamma} = \arg\min_{\gamma} Q_n(\gamma),$$

where

$$Q_n(\gamma) = \min_{\beta_1,\beta_2} n P_n(m(\cdot|\theta)) = \min_{\beta_1,\beta_2} \sum_{i=1}^{n} m(w_i|\theta),$$

with P_n being the empirical measure of the original data, and

$$m(w|\theta) = \left(y - x'\beta_1 \mathbf{1}(q \le \gamma) - x'\beta_2 \mathbf{1}(q > \gamma)\right)^2.$$

Usually, there is an interval of γ minimising this objective function. Most literature in threshold regression takes the left endpoint of the interval as the minimiser and calls the estimator as the left-endpoint LSE (LLSE). Yu (2008) suggests to use the middle point LSE (MLSE) and shows that in most cases with $0 < \underline{f} \le f(q) \le \overline{f} < \infty$, the MLSE is more efficient than the LLSE. This paper will explore the asymptotic properties of both estimators. To express the model in matrix notation, define the $n \times 1$ vector Y by stacking the variables y_i, and the $n \times k$ matrices X, $X_{\le\gamma}$ and $X_{>\gamma}$ by stacking the vectors x_i', $x_i'\mathbf{1}(q_i \le \gamma)$ and $x_i'\mathbf{1}(q_i > \gamma)$, respectively. Let

$$\begin{pmatrix} \widehat{\beta}_1(\gamma) \\ \widehat{\beta}_2(\gamma) \end{pmatrix} \equiv \arg\min_{\beta_1,\beta_2} \sum_{i=1}^{n} m(w_i|\theta) = \begin{pmatrix} \left(X_{\le\gamma}'X_{\le\gamma}\right)^{-1} X_{\le\gamma}'Y \\ \left(X_{>\gamma}'X_{>\gamma}\right)^{-1} X_{>\gamma}'Y \end{pmatrix},$$

then the LSE of β is defined as $\widehat{\beta} = (\widehat{\beta}_1', \widehat{\beta}_2')' \equiv (\widehat{\beta}_1'(\widehat{\gamma}), \widehat{\beta}_2'(\widehat{\gamma}))'$. In the classical case, $\widehat{\beta} - \beta_0 = O_P(n^{-1/2})$, and $\widehat{\gamma} - \gamma_0 = O_P(n^{-1})$.

2.2. *No error term: an illustration*

A simple threshold model is considered here to illustrate the effect of $f(q)$ on the inference of γ_0. A similar example is used in Section 2 of Yu (2012). The model is

$$y = \mathbf{1}(q \le \gamma), \tag{2.1}$$

where γ is the only parameter of interest, and q has a cdf $F(\cdot)$ that represents the only randomness in this simple model. For now, suppose q has a density that is continuous and positive on $B(\gamma_0)\backslash\{\gamma_0\}$, where $B(\gamma_0)$ is an open ball centred at γ_0. This model can be viewed as a treatment rule in RDDs. Usually, γ_0 is known in RDDs, but it may also be unknown as analysed in Porter and Yu (2012). So this simple model is relevant in practice. We will use this simple model to thoroughly characterise the local behaviour of $f(q)$ around γ_0, which is key to the inference of γ_0. In the following discussion, $q_{(m)}$ denotes the mth order statistic of $\{q_i\}_{i=1}^n$.

A simple calculation shows that the LSE of γ is the interval $[q_{(m)}, q_{(m+1)})$ with $\gamma_0 \in [q_{(m)}, q_{(m+1)})$, where m y_i's are 1 and the remaining y_i's are 0. Suppose the LLSE, $\widehat{\gamma}_{\text{LLSE}}$, is used as the estimator of γ, then $\widehat{\gamma}_{\text{LLSE}} = q_{(m)}$, which is the q_i that is closest to γ_0 from the left. For $t \le 0$ and some sequence of constants a_n,

$$\begin{aligned} P(a_n(\widehat{\gamma}_{\text{LLSE}} - \gamma_0) \le t) &= P\left(q_i \notin \left(\gamma_0 + \frac{t}{a_n}, \gamma_0\right] \text{ for all } i\right) \\ &= \left(1 - \frac{1}{n} \cdot n\left(F(\gamma_0) - F\left(\gamma_0 + \frac{t}{a_n}\right)\right)\right)^n \to e^{-\Lambda_1(t)}, \end{aligned} \tag{2.2}$$

where

$$\Lambda_1(t) = \lim_{n\to\infty} \Lambda_{1n}(t) \equiv \lim_{n\to\infty} n\left(F(\gamma_0) - F\left(\gamma_0 + \frac{t}{a_n}\right)\right) \in [0,\infty],\ t \le 0$$

is the average number of data points falling into $(\gamma_0 + t/a_n, \gamma_0]$ when n data points are randomly sampled, and a_n is selected such that $\Lambda_1(t)$ does not degenerate to 0 or ∞ on $(-\infty, 0)$.[2] When $f(\cdot)$ in the left neighbourhood of γ_0 gets larger, a_n gets larger as a smaller neighbourhood $(\gamma_0 + t/a_n, \gamma_0]$ can have enough samples falling in it. From the assumptions on $f(\cdot)$, a_n diverges to infinity but is not always infinity, and $\Lambda_1(\cdot)$ is decreasing with $\Lambda_1(-\infty) = \infty$ and $\Lambda_1(0) = 0$. As a result, $a_n(\widehat{\gamma}_{\text{LLSE}} - \gamma_0) = O_p(1)$. It can be seen that only the information in the left neighbourhood of γ_0 is used in $\widehat{\gamma}_{\text{LLSE}}$. Suppose $\Lambda_2(t)$ is similarly defined but using the information in the right neighbourhood of γ_0:

$$\Lambda_2(t) = \lim_{n\to\infty} \Lambda_{2n}(t) \equiv \lim_{n\to\infty} n\left(F\left(\gamma_0 + \frac{t}{b_n}\right) - F(\gamma_0)\right) \in [0,\infty],\ t > 0,$$

where b_n is selected such that $\Lambda_2(t)$ does not degenerate to 0 or ∞ on $(0,\infty)$. $\Lambda(t)$ is defined as $\Lambda_1(t)$ when $t \le 0$ and $\Lambda_2(t)$ when $t > 0$.

If the MLSE, $\widehat{\gamma}_{\text{MLSE}}$, is used, then

$$\widehat{\gamma}_{\text{MLSE}} - \gamma_0 = \frac{q_{(m)} + q_{(m+1)}}{2} - \gamma_0 = \frac{q_{(m)} - \gamma_0}{2} + \frac{q_{(m+1)} - \gamma_0}{2},$$

and the information in both the left and right neighbourhoods of γ_0 is used. It can be shown that the convergence rate of $\widehat{\gamma}_{\text{MLSE}}$ is determined by $a_n \wedge b_n$ by the following arguments. For $t > 0$,

$$\begin{aligned} P\left(b_n(q_{(m+1)} - \gamma_0) > t\right) &= P\left(q_i \notin \left[\gamma_0, \gamma_0 + \frac{t}{b_n}\right] \text{ for all } i\right) \\ &= \left(1 - \frac{1}{n}\cdot n\left(F\left(\gamma_0 + \frac{t}{b_n}\right) - F(\gamma_0)\right)\right)^n \to e^{-\Lambda_2(t)}, \end{aligned}$$

so $b_n(q_{(m+1)} - \gamma_0) = O_p(1)$. If $a_n/b_n \to 0$, then $a_n(q_{(m+1)} - \gamma_0) = o_p(1)$. By Slutsky's theorem, $a_n(\widehat{\gamma}_{\text{MLSE}} - \gamma_0)$ has the same asymptotic distribution as $a_n\frac{q_{(m)}-\gamma_0}{2}$, which is $O_p(1)$. However, $b_n(q_{(m)} - \gamma_0)$ diverges to $-\infty$ almost surely, so $b_n(\widehat{\gamma}_{\text{MLSE}} - \gamma_0)$ diverges to $-\infty$ and $b_n(\widehat{\gamma}_{\text{MLSE}} - \gamma_0)$ cannot be $O_p(1)$. Other cases of a_n and b_n can be similarly analysed. When a_n/b_n converges to a positive number $C \in (0,\infty)$, it can be made to converge to 1 by rescaling a_n and b_n. In summary, the asymptotic distribution of $\widehat{\gamma}_{\text{MLSE}}$ falls into one of the three cases:

1. If $a_n/b_n \to 1$, then

$$P(a_n \wedge b_n(\widehat{\gamma}_{\text{MLSE}} - \gamma_0) \le t) \to -\int_{(2t)\vee 0}^{\infty} e^{-\Lambda_1(2t-s)} de^{-\Lambda_2(s)}. \tag{2.3}$$

2. If $a_n/b_n \to 0$, then

$$P(a_n \wedge b_n(\widehat{\gamma}_{\text{MLSE}} - \gamma_0) \le t) \to e^{-\Lambda_1(2t)} \text{ for } t \le 0.$$

[2] Note that if a_n is the appropriate normalising rate, so is Aa_n for any $0 < A < \infty$. Correspondingly, $\Lambda_1(t)$ becomes $\Lambda_1(\frac{t}{A})$. So we call $\Lambda^{(1)}(\cdot)$ and $\Lambda^{(2)}(\cdot)$ to be *of the same type* if for some $A > 0$, $\Lambda^{(2)}(t) = \Lambda^{(1)}(At)$.

3. If $a_n/b_n \to \infty$, then

$$P(a_n \wedge b_n(\widehat{\gamma}_{\text{MLSE}} - \gamma_0) > t) \to e^{-\Lambda_2(2t)} \text{ for } t > 0.$$

Interestingly, when $a_n/b_n \to \infty$, $\widehat{\gamma}_{\text{LLSE}}$ and $\widehat{\gamma}_{\text{MLSE}}$ have different convergence rates. The lesson here is that when both neighbourhoods are involved in the estimation, the convergence rate is determined by the neighbourhood with 'less' data.

For both $\widehat{\gamma}_{\text{LLSE}}$ and $\widehat{\gamma}_{\text{MLSE}}$, the distribution of q in the neighbourhood of γ_0 determines their asymptotic distributions by determining Λ. This insight is still true when an error term is added in, i.e.

$$y = \mathbf{1}(q \le \gamma) + e. \tag{2.4}$$

It can be shown that for the LSE $\widehat{\gamma}$ and a convergence rate c_n,

$$c_n(\widehat{\gamma} - \gamma_0) = \arg\min_v D_n(v),$$

where

$$D_n(v) = \sum_{i=1}^{n}(1 + 2e_i)\mathbf{1}\left(\gamma_0 + \frac{v}{c_n} < q_i \le \gamma_0\right) + \sum_{i=1}^{n}(1 - 2e_i)\mathbf{1}\left(\gamma_0 < q_i \le \gamma_0 + \frac{v}{c_n}\right)$$

is a step function which simplifies to $\sum_{i=1}^{n}\mathbf{1}(\gamma_0 + \frac{v}{c_n} < q_i \le \gamma_0) + \sum_{i=1}^{n}\mathbf{1}(\gamma_0 < q_i \le \gamma_0 + \frac{v}{c_n})$ in (2.1) with $c_n = a_n$ for $\widehat{\gamma}_{\text{LLSE}}$ and $c_n = a_n \wedge b_n$ for $\widehat{\gamma}_{\text{MLSE}}$. Obviously, when an error term is added in, only the jump magnitudes in $D_n(v)$ change from 1 to $1 \pm 2e_i$, while the jump locations are the same. However, the jump locations are completely determined by $f(q)$ around γ_0 in finite samples and by Λ asymptotically, so the simple model (2.1) embodies the key difference between this paper and the literature. This implies that we can study the general model by only adding in random jumps. Of course, there is indeed some speciality in this simple model: in the general model, the data points in both neighbourhoods of γ_0 are used even in $\widehat{\gamma}_{\text{LLSE}}$, so the convergence rates for both $\widehat{\gamma}_{\text{LLSE}}$ and $\widehat{\gamma}_{\text{MLSE}}$ are $a_n \wedge b_n$, which is denoted as c_n hereafter.

2.3. Characterisation of the local behaviour of q

This subsection provides a thorough characterisation of Λ which is determined solely by the local behaviour of $f(q)$ around γ_0 and is a key building block in deriving the asymptotic distribution of $\widehat{\gamma}$ in the next section. This characterisation is obtained by abstracting two examples. Readers may jump to Theorem 2.1 directly and then check these examples to see what specification of $f(q)$ around γ_0 will generate the corresponding Λ. For simplicity, suppose $\gamma_0 = 0$. The specifications of $F(x) - F(0)$ in the first two examples are essentially borrowed from Knight (1998).

EXAMPLE 2.1. Suppose

$$F(x) - F(0) = \lambda \text{sign}(x)|x|^{\alpha}L(|x|) \tag{2.6}$$

for x in a neighbourhood of 0, where $\alpha > 0$, $\lambda > 0$, and L is a *slowly varying* function at 0, denoted as $L \in RV_0$ as $x \to 0$. In other words, $F(x) - F(0)$ is a *regularly varying* function at zero, denoted as $F(x) - F(0) \in RV_\alpha$ as $x \to 0$, and α is called the *exponent of variation*.[3] See

[3] A positive locally integrable function $L : (0, \infty) \to (0, \infty)$ is called *slowly varying* at zero if $\lim_{x \downarrow 0}\frac{L(kx)}{L(x)} = 1$, for any $k > 0$. If this limit is finite but non-zero for any $k > 0$, then L is called *regularly varying* at zero. A function $L(x)$

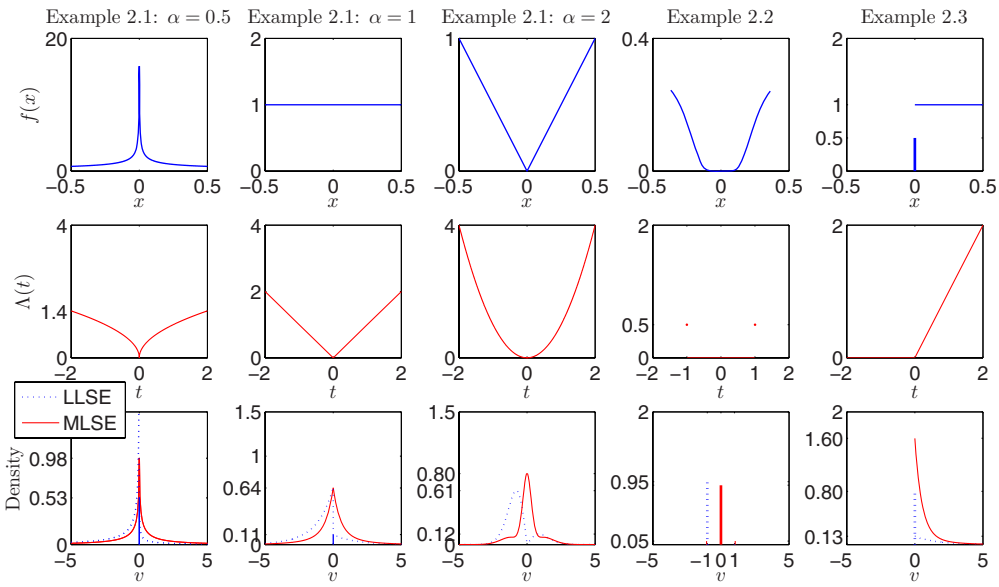

Figure 1. $\Lambda(t)$ and LSE asymptotic distributions.

Section 0.4.2 of Resnick (1987) and Seneta (1976) for more details on this type of functions. In this case,

$$a_n = b_n = n^{1/\alpha} L^*(n), \text{ and } \Lambda(t) = \lambda |t|^\alpha,$$

where L^* is a slowly varying function at infinity satisfying $L^*(n) = L(1/a_n)^{1/\alpha}$.

Two typical examples of L are as follows. (a) If $L(x)$ is a constant function (for x close to 0), then $L^*(n) = 1$. When $\alpha = 1$, $f(\cdot)$ reduces to the classical case and λ plays a similar role as $f(\gamma_0)$, so the convergence rate is n. When $\alpha < 1$, $f(0) = \infty$, so the convergence rate is $n^{\frac{1}{\alpha}}$ which is faster than n. When $\alpha > 1$, $f(0) = 0$, so the convergence rate is $n^{\frac{1}{\alpha}}$ which is slower than n. (b) If $L(|x|) = \ln(|x|^{-1})$ (for x close to 0), then $L^*(n) = (\ln n)^{1/\alpha}$. When $\alpha = 1$, $a_n = n \ln n$, which is greater than n since $f(\cdot)$ diverges to infinity at the rate $-\ln|x|$ for x close to 0. When $\alpha < 1$, $a_n > n$, and when $\alpha > 1$, $a_n \in (n^{\frac{1}{\alpha}}, n)$.

Figure 1 shows $\Lambda(t)$ with $L(\cdot) = 1$, $\lambda = 1$ and different α values. It can be seen that $\Lambda(t)$ is very different from that in the classical case where $\alpha = 1$ and $\Lambda(t) = \lambda|t|$.

When $f(\cdot)$ is continuous in the neighbourhood of γ_0, $\Lambda_{\ell n}(\cdot)$ is continuous. One may also expect $\Lambda_\ell(\cdot)$ to be continuous, but this is not true as shown in the following example.

is slowly (regularly) varying at infinity iff $L(1/x)$ is slowly (regularly) varying at zero. One typical example of slowly varying functions is the logarithm; other examples are the powers and the iterations of the logarithm, e.g. $\ln^\alpha$, $\alpha \in \mathbb{R}$ and $\ln\ln$. The function $L(x) = x$ is not slowly varying, neither is $L(x) = x^\alpha$ for any real $\alpha \neq 0$. They are regularly varying functions. Any regularly varying function is of the form $x^\alpha L(x)$ where $\alpha \geq 0$ and L is a slowly varying function.

EXAMPLE 2.2. Suppose

$$f(x) = \frac{1}{2x^2}\exp(-|x|^{-1})$$

and its cdf

$$F(x) = \frac{1}{2}(1 + \text{sign}(x)\exp(-|x|^{-1})).$$

It is easy to show that

$$\Lambda(t) = \lim_{n\to\infty} n\left|F\left(\frac{t}{\ln n}\right) - F(0)\right| = \begin{cases} 0, & \text{if } |t| < 1; \\ 1/2, & \text{if } |t| = 1; \\ \infty, & \text{if } |t| > 1; \end{cases}$$

which is discontinuous although $f(\cdot)$ is continuous. So $a_n = b_n = \ln n$ is very slow.[4] In the extreme case where $f(x) = 0$ for x in a neighbourhood of 0, γ_0 cannot be identified, corresponding to zero convergence rate. $\Lambda(t)$ is shown in Figure 1.

The form of $F(x)$ is well known in calculus for the fact that any of its finite-order derivatives at γ_0 is zero. de Haan (1970) calls $F(\gamma_0 + x) - F(\gamma_0)$ as *rapidly varying* at $x = 0$, and Resnick (1987) gives a name as *regularly varying with index* ∞. Two intuitive notations used in the literature are $F(\gamma_0 + x) - F(\gamma_0) \sim |x|^{\infty}L(|x|)$ or $F(\gamma_0 + x) - F(\gamma_0) \in RV_{\infty}$ as $x \to 0$. Rigorously, if for $x > 0$,

$$\lim_{t\downarrow 0}\frac{F(\gamma_0 + tx) - F(\gamma_0)}{F(\gamma_0 + t) - F(\gamma_0)} = x^{\infty} \equiv \begin{cases} 0, & \text{if } x < 1; \\ 1, & \text{if } x = 1; \\ \infty, & \text{if } x > 1; \end{cases}$$

then $F(\gamma_0 + x) - F(\gamma_0) \in RV_{\infty}$ as $x \downarrow 0$. A similar definition applies to $F(\gamma_0) - F(\gamma_0 - x)$ as $x \downarrow 0$. Compared with Example 2.1, such kind of functions can be treated as the limit case when $\alpha \to \infty$.

The above discussion assumes that there is no point mass at γ_0 in $F(\cdot)$. The following example shows what will happen when this assumption is violated.

EXAMPLE 2.3. Suppose there is a point mass 2^{-1} at 0 in $F(\cdot)$, and $f(\cdot) = 1$ in the right neighbourhood of 0. In the simple Example 2.1, $P(\widehat{\gamma}_{\text{LLSE}} = 0) = 1 - (1 - f(0))^n \to 1$. For $t < 0$, $\Lambda_1(t) = \infty$ as long as $a_n \neq \infty$, and for $t > 0$, $a_n = n$ and $\Lambda_2(t) = t$. Consequently, for any $a_n \neq \infty$,

$$a_n(\widehat{\gamma}_{\text{LLSE}} - \gamma_0) \xrightarrow{d} 0, \quad \text{and} \quad n(\widehat{\gamma}_{\text{MLSE}} - \gamma_0) \xrightarrow{d} \text{Exp}(1/2),$$

where $\text{Exp}(\Delta)$ is an exponential distribution with scale Δ. In other words, the convergence rate of $\widehat{\gamma}_{\text{LLSE}}$ is ∞, while the convergence rate of $\widehat{\gamma}_{\text{MLSE}}$ is n. Here, a_n is defined as ∞ when there is a point mass at γ_0 in $F(\cdot)$. If $a_n = \infty$, $\Lambda_1(t) = 0$ for $t < 0$, as shown in Figure 1. When an error term is added in, the convergence rates of both $\widehat{\gamma}_{\text{LLSE}}$ and $\widehat{\gamma}_{\text{MLSE}}$ will be $a_n \wedge b_n = b_n$ which is determined by $f(\cdot)$ in the right neighbourhood of γ_0. So the insight from the last subsection still applies as long as we treat the point mass as the limit case of Example 2.1 with $\alpha \to 0$.

[4] The convergence rate can be arbitrarily slow. For example, $F(x) = \frac{1}{2}(1 + \text{sign}(x)\exp(-\exp(|x|^{-1})))$, then $a_n = \ln\ln n$.

The above examples suggest that any rate between 0 and infinity is possible. In Examples 2.1 and 2.2, we derived $\Lambda_\ell(\cdot)$ for special $f(\cdot)$'s. Actually, these $\Lambda_\ell(\cdot)$'s are the only possible forms. Theorem 2.1 states this result for $\Lambda_1(\cdot)$, and a similar result applies to $\Lambda_2(\cdot)$.

THEOREM 2.1. *Suppose there is no point mass at γ_0, and $f(\cdot)$ is continuous and positive on $(\gamma_0-\epsilon, \gamma_0)$ for some $\epsilon > 0$. If a_n and $\Lambda_1(\cdot)$ exist, then $\Lambda_1(\cdot)$ falls into one of the following two classes:*

(a) $\Lambda_1(t) = |t|^{\alpha_1}$, *for $t < 0$ and some $0 < \alpha_1 < \infty$;*

(b) $\Lambda_1(t) = \begin{cases} 0, & \text{if } t > -1; \\ \kappa_1, & \text{if } t = -1; \\ \infty, & \text{if } t < -1; \end{cases}$ *for some $0 \le \kappa_1 \le \infty$.*

Moreover, $\Lambda_1(\cdot)$ is of the type (a) iff $F(\gamma_0) - F(\gamma_0 - x) \in RV_{\alpha_1}$ as $x \downarrow 0$, and $\Lambda_1(\cdot)$ is of the type (b) iff $F(\gamma_0) - F(\gamma_0 - x) \in RV_\infty$ as $x \downarrow 0$. In both cases,

$$a_n = \frac{1}{\gamma_0 - \gamma_n},$$

where $\gamma_n = (1/(F(\gamma_0) - F(\cdot)))^{\leftarrow}(n)$ with $H^{\leftarrow}(y) = \inf\{s : H(s) \ge y\}$ being the inverse function.

Theorem 2.1 says that when $f(\cdot)$ is very thin in the neighbourhood of γ_0, $\Lambda_\ell(\cdot)$ is discontinuous; otherwise, $\Lambda_\ell(\cdot)$ is continuous. In the former case, the possibilities that $\kappa_\ell = 0$ or ∞ cannot be excluded. In the latter case, more descriptions on $F(\cdot)$ can be found in Corollary 1.14 and Proposition 1.16 of Resnick (1987). For calculating the normalising constant a_n, Proposition 1.19 of Resnick (1987) suggests that we need only calculate this constant for the tail equivalent distribution. The proof of Theorem 2.1 is borrowed from the extreme value theory; see, e.g. Proposition 0.3 of Resnick (1987). One may think that Theorem 2.1 is just a trivial corollary of Proposition 0.3 of Resnick (1987) since γ_0 can be treated as the right endpoint of the distribution $F(\cdot)$ in the extreme value theory, but this is not true. The key point here is that the location (or drift) normalising parameter in Proposition 0.3 of Resnick (1987) can be chosen freely, so $\Lambda_1(t)$ can also take the form of e^{-t}; while in our case, this parameter is fixed, so the non-degenerate $\Lambda_1(t)$ can only take the form $|t|^{\alpha_1}$. Example 2 on page 445 of Gnedenko (1943) illustrates this point. Note that Example 2 there is essentially the same as Example 2.2.

In all the above examples, a_n (b_n) and $\Lambda_\ell(\cdot)$ exist, but this does not always hold. Such an example is given below which is inspired by Remark 3 of Zhu (2009), and can be treated as the continuous version of the example on page 451 of Gnedenko (1943).

EXAMPLE 2.4. Take the cut-off function

$$\chi(x) = \begin{cases} 0, & x \le 1; \\ \kappa(x), & 1 < x < 2; \\ 1, & x \ge 2; \end{cases}$$

such that $\chi(x) \in C^{(\infty)}(\mathbb{R})$ and $\kappa(x) \in (0,1)$. For example, we can take $\kappa(x) = \frac{\int_1^x \exp\{-\frac{1}{0.5^2-(t-1.5)^2}\}dt}{\int_1^2 \exp\{-\frac{1}{0.5^2-(t-1.5)^2}\}dt}$. Now, let $F_1(x) = \frac{1}{2A}\sum_{k=-1}^{\infty} p_k\ \chi(e^k x)$, where $A = \sum_{k=-1}^{\infty} p_k \in (0,\infty)$.

Note that for $x \in [\frac{2}{e^i}, \frac{1}{e^{i-1}}]$, $i \in \mathbb{Z}_+$, $\chi(e^k x) = 0$ when $k < i$, and $\chi(e^k x) = 1$ when $k \geq i$. So

$$F_1(x) = \begin{cases} \dfrac{1}{2A} p_i \kappa(e^i x) + \dfrac{1}{2A} \displaystyle\sum_{k=i+1}^{\infty} p_k, & \text{if } x \in \left[\dfrac{1}{e^i}, \dfrac{2}{e^i}\right] \text{ for some } i \in \mathbb{Z}_+; \\ \dfrac{1}{2A} \displaystyle\sum_{k=i}^{\infty} p_k, & \text{if } x \in \left[\dfrac{2}{e^i}, \dfrac{1}{e^{i-1}}\right] \text{ for some } i \in \mathbb{Z}_+; \\ \dfrac{1}{2A} \displaystyle\sum_{k=0}^{\infty} p_k + \dfrac{1}{2A} p_{-1} \kappa(e^{-1} x), & \text{if } x > e. \end{cases}$$

Let $F_2(x) = \frac{1}{4}(1 + \text{sign}(x) \exp(-\frac{e}{|x|}))$ be a variant of the $F(x)$ in Example 2.2. Define $F(x) = F_1(x) + F_2(x)$, then $F(x)$ is in $C^{(\infty)}(\mathbb{R})$ and has a positive density in the neighbourhood of zero.

Suppose $p_k = e^{-e^k}$, and thus $\lim_{i \to \infty} \frac{\sum_{k=i+1}^{\infty} p_k}{p_i} = 0$; that is, p_i dominates the remaining terms when i is large enough. Now,

$$n \cdot 2A \cdot F_1\left(\frac{t}{b_n}\right) = \begin{cases} n \cdot p_i \kappa\left(\dfrac{e^i}{b_n} t\right) + n \displaystyle\sum_{k=i+1}^{\infty} p_k, & \text{if } t \in \left[\dfrac{b_n}{e^i}, \dfrac{2b_n}{e^i}\right] \text{ for some } i \in \mathbb{Z}_+; \\ n \displaystyle\sum_{k=i}^{\infty} p_k, & \text{if } t \in \left[\dfrac{2b_n}{e^i}, \dfrac{b_n}{e^{i-1}}\right] \text{ for some } i \in \mathbb{Z}_+; \\ n \displaystyle\sum_{k=0}^{\infty} p_k + p_{-1} \kappa\left(\dfrac{1}{b_n e} t\right), & \text{if } t > b_n e; \end{cases}$$

$$= \begin{cases} n \cdot p_{[\ln b_n]+i} \kappa\left(\dfrac{e^{[\ln b_n]+i}}{b_n} t\right) + n \displaystyle\sum_{k=[\ln b_n]+i+1}^{\infty} p_k \\ = n \cdot e^{-e^{[\ln b_n]+i}} \kappa\left(\dfrac{e^{[\ln b_n]+i}}{b_n} t\right) + n \displaystyle\sum_{k=[\ln b_n]+i+1}^{\infty} e^{-e^k}, \\ \text{if } t \in \left[\dfrac{b_n}{e^{[\ln b_n]+i}}, \dfrac{2b_n}{e^{[\ln b_n]+i}}\right] \text{ for some } i \in \mathbb{Z} \text{ such that } [\ln b_n] + i \in \mathbb{Z}_+; \\ n \displaystyle\sum_{k=[\ln b_n]+i}^{\infty} p_k = n \sum_{k=[\ln b_n]+i}^{\infty} e^{-e^k}, \\ \text{if } t \in \left[\dfrac{2b_n}{e^{[\ln b_n]+i}}, \dfrac{b_n}{e^{[\ln b_n]+i-1}}\right] \text{ for some } i \in \mathbb{Z} \text{ such that } [\ln b_n] + i \in \mathbb{Z}_+; \\ n \displaystyle\sum_{k=0}^{\infty} p_k + p_{-1} \kappa\left(\dfrac{1}{b_n e} t\right) = n \sum_{k=0}^{\infty} e^{-e^k} + c_{-1} \kappa\left(\dfrac{1}{b_n e} t\right), \text{ if } t > b_n e; \end{cases}$$

where the last equality is just a shift of the index i. It seems that to make the limit non-degenerate, b_n should be $O(\ln n)$, and i takes -1 or 0. Actually, $n \cdot 2A \cdot F_1(\frac{t}{b_n})$ does not even converge at all given that $\lim_{n \to \infty} n e^{-e^{[\ln \ln n]}}$ does not exist. Nevertheless, we can select a subsequence of n to

make it converge; e.g. when $n_k = [e^{e^k}]$, and $b_{n_k} = e^k$,

$$n_k \cdot 2A \cdot F_1\left(\frac{t}{b_{n_k}}\right) \to 2A \cdot \Lambda_{21}(t) = \begin{cases} 0, & \text{if } t \in (0,1]; \\ \kappa(t), & \text{if } t \in (1,2]; \\ 1, & \text{if } t \in (2,e]; \\ \infty, & \text{if } t \in (e,\infty). \end{cases}$$

From Example 2.2,

$$\Lambda_1(t) = \lim_{n\to\infty} n\left(F_2(0) - F_2\left(\tfrac{t}{\ln n}\right)\right) = \begin{cases} 0, & \text{if } t > -e; \\ 1/4, & \text{if } t = -e; \\ \infty, & \text{if } t < -e; \end{cases}$$

$$\Lambda_{22}(t) \ = \lim_{n\to\infty} n\left(F_2\left(\tfrac{t}{\ln n}\right) - F_2(0)\right) = \begin{cases} 0, & \text{if } t < e; \\ 1/4, & \text{if } t = e; \\ \infty, & \text{if } t > e. \end{cases}$$

Consequently,

$$\Lambda_2(t) = \Lambda_{21}(t) + \Lambda_{22}(t) = \begin{cases} 0, & \\ \dfrac{\kappa(t)}{2A}, & \text{if } t \in (0,1]; \\ \dfrac{1}{2A}, & \text{if } t \in (1,2]; \\ \dfrac{1}{2A}, & \text{if } t \in (2,e); \\ \dfrac{1}{2A} + \dfrac{1}{4}, & \text{if } t = e; \\ \infty, & \text{if } t \in (e,\infty); \end{cases}$$

which does not take the form in Theorem 2.1, so $n \cdot 2A \cdot F_1(\frac{t}{b_n})$ does not converge. Figure 2 shows the density of $F(\cdot)$ and $\Lambda(t)$ under the subsequence n_k. Gnedenko (1943) used the Poisson distribution as the example that is not attracted to any of the three limit laws of the maxima. From Figure 2, we can see that $f_2(x)$ is like swelling a point mass to a smooth density on each interval $[\frac{1}{e^i}, \frac{2}{e^i}]$, $i = -1, 0, 1, \ldots$. In this sense, it is not surprising that $n \cdot 2A \cdot F_1(\frac{t}{b_n})$ does not converge. Another observation is that the non-parametric technique such as kernel smoothing cannot identify the local behaviour of $f(\cdot)$ around zero, and thus in practice, it is hard to determine if c_n exists for a given dataset.

3. ASYMPTOTIC THEORY

After studying the local behaviour of the threshold variable, we are ready to explore the asymptotic theory for the LSEs in the general model (1.1). We begin with some regularity assumptions.

3.1. Regularity assumptions

For convenience, we divide the assumptions into Assumptions 3.1–3.4, which are about the local behaviour of $f(\cdot)$ around γ_0, and Assumption 3.5, which contains all the remaining assumptions.

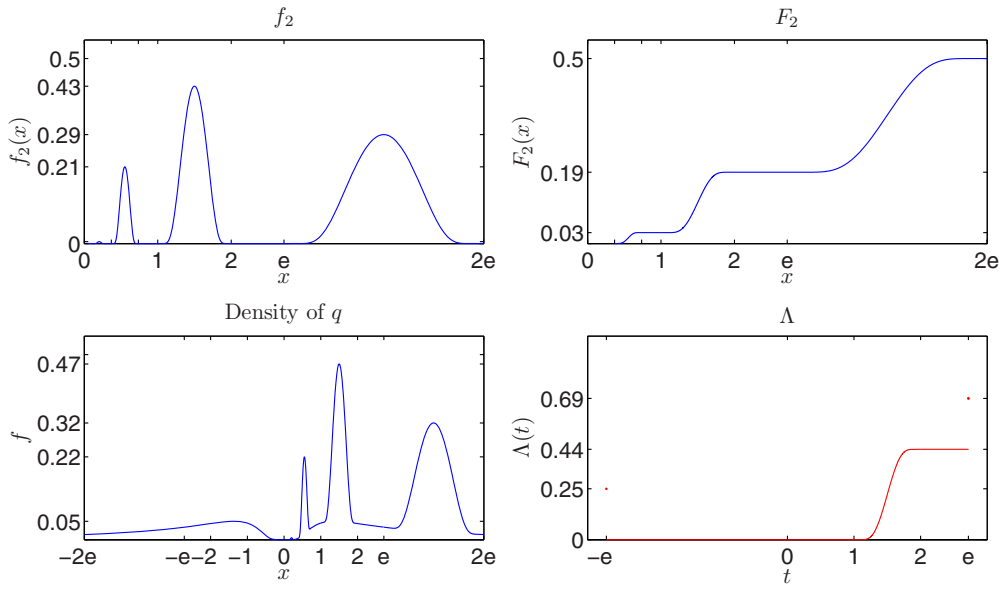

Figure 2. $F_2(x)$, $f_2(x)$, $f(x)$ and $\Lambda(t)$.

ASSUMPTION 3.1. *If there is no point mass at γ_0 in $F(\cdot)$, then $f(\cdot)$ is continuous and positive on $(\gamma_0 - \epsilon, \gamma_0)$ and $(\gamma_0, \gamma_0 + \epsilon)$ for some $\epsilon > 0$. If there is a point mass at γ_0 in $F(\cdot)$, then $f(\cdot)$ needs to be continuous and positive only on $(\gamma_0, \gamma_0 + \epsilon)$ for some $\epsilon > 0$.*

Assumption 3.1 is used to identify γ_0, so it excludes the partial identification case where $f(x) = 0$ in a neighbourhood of γ_0. Since $f(\cdot) \in L^1$ on $(\gamma_0 - \epsilon, \gamma_0)$ and $(\gamma_0, \gamma_0 + \epsilon)$, and continuous functions are dense in L^1 space, Assumption 3.1 is not very restricted.

ASSUMPTION 3.2. *a_n (b_n) and $\Lambda_\ell(\cdot)$ exist.*

This assumption excludes the case such as Example 2.4.

ASSUMPTION 3.3. *If there is no point mass at γ_0 in $F(\cdot)$, then $F(\gamma_0) - F(\gamma_0 - x) \in RV_{\alpha_1}$ as $x \downarrow 0$, and $F(\gamma_0 + x) - F(\gamma_0) \in RV_{\alpha_2}$ as $x \downarrow 0$, where $\alpha_\ell \in (0, \infty)$. If there is a point mass at γ_0 in $F(\cdot)$, then only $F(\gamma_0 + x) - F(\gamma_0) \in RV_{\alpha_2}$ as $x \downarrow 0$.*

Based on Theorem 2.1, Assumption 3.3 implies that $\Lambda_\ell(t) = |t|^{\alpha_\ell}$.[5] From Section 2, c_n is determined by the larger α_ℓ. If $F(\cdot)$ satisfies Assumption 3.3, we can normalise a_n and b_n such that a_n/b_n converges to one of the three values: 0, 1 and ∞.

ASSUMPTION 3.4. *Either $F(\gamma_0) - F(\gamma_0 - x) \in RV_\infty$ as $x \downarrow 0$, or $F(\gamma_0 + x) - F(\gamma_0) \in RV_\infty$ as $x \downarrow 0$, or both are satisfied.*

When only one neighbourhood of $F(\cdot)$ falls in RV_∞, c_n is determined by that neighbourhood; when both neighbourhoods fall in RV_∞, c_n is determined by the neighbourhood with a thinner

[5] Actually, the theorems in Section 3.2 only need the continuity of $\Lambda_\ell(\cdot)$, which is required in proving the uniform convergence of $\Lambda_{\ell n}(\cdot)$ to $\Lambda_\ell(\cdot)$. From 0.1 of Resnick (1987), $\Lambda_{\ell n}(\cdot)$ converges uniformly to $\Lambda_\ell(\cdot)$ on any compact set when $\Lambda_{\ell n}(\cdot)$ is non-decreasing and $\Lambda_\ell(\cdot)$ is continuous.

density. For convenience in stating the theorems in the next subsection, we regularise a_n, b_n and $F(\cdot)$ as follows. When a_n/b_n converges to 0, a_n is selected such that $\Lambda_1(t)$ jumps at -1 with $\Lambda_1(-1) = \kappa_1$; when a_n/b_n diverges to ∞, b_n is selected such that $\Lambda_2(t)$ jumps at 1 with $\Lambda_2(1) = \kappa_2$; when a_n/b_n converges to a positive number $C \in (0, \infty)$, we assume the limit $\Lambda_1(t)$ of $\Lambda_{1n}(t) \equiv n(F(\gamma_0) - F(\gamma_0 + \frac{t}{a_n \wedge b_n}))$ jumps at -1 and the limit $\Lambda_2(t)$ of $\Lambda_{2n}(t) \equiv n(F(\gamma_0 + \frac{t}{a_n \wedge b_n}) - F(\gamma_0))$ jumps at 1. In an RV_∞ neighbourhood, $\Lambda_{\ell n}(\cdot)$ will not uniformly converge to $\Lambda_\ell(\cdot)$ on some compact set. In consequence, the stochastic equicontinuity fails in this case, but a direct calculation can be used to find the asymptotic distribution.

ASSUMPTION 3.5. *(a) $w_i \in \mathbb{W} \equiv \mathbb{R} \times \mathbb{X} \times \mathbb{Q} \subset \mathbb{R}^{k+2}$, $\beta_1 \in B_1 \subset \mathbb{R}^k$, $\beta_2 \in B_2 \subset \mathbb{R}^k$, $0 < \sigma_1 \in \Omega_1 \subset \mathbb{R}$, $0 < \sigma_2 \in \Omega_2 \subset \mathbb{R}$, $\Omega_1 \times \Omega_2$ is compact, $\gamma \in \Gamma = [\underline{\gamma}, \overline{\gamma}] \subset \mathbb{R}$, $\beta_{10} \neq \beta_{20}$, and $\sigma_{10} \neq \sigma_{20}$, where $\neq$ is an element by element operation; (b) $E[\|xe\|^2] < \infty$, and $E[\|x\|^4] < \infty$; (c) $E[xx'] > E[xx'\mathbf{1}(q \leq \gamma)] > 0$ for all $\gamma \in \Gamma$; (d) $E[\|xe\|^2 | q = \gamma] < \infty$, $E[\|x\|^4 | q = \gamma] < \infty$ and $E[.xx'|q = \gamma] > 0$ for γ in a neighbourhood of γ_0; (e) Both z_{1i} and z_{2i} have absolutely continuous distributions, where the distribution of z_{1i} is the limiting conditional distribution of $\overline{z}_{1i}$ given $\gamma_0 + \Delta < q_i \leq \gamma_0$, $\Delta < 0$ as $\Delta \uparrow 0$ with*

$$\overline{z}_{1i} = \{2x_i'(\beta_{10} - \beta_{20})\sigma_{10}e_i + (\beta_{10} - \beta_{20})x_i x_i'(\beta_{10} - \beta_{20})\},$$

and the distribution of z_{2i} is the limiting conditional distribution of $\overline{z}_{2i}$ given $\gamma_0 < q_i \leq \gamma_0 + \Delta$, $\Delta > 0$ as $\Delta \downarrow 0$ with

$$\overline{z}_{2i} = \left\{-2x_i'(\beta_{10} - \beta_{20})\sigma_{20}e_i + (\beta_{10} - \beta_{20})x_i x_i'(\beta_{10} - \beta_{20})\right\}.$$

By Assumption 3.5(a), $y \in \mathbb{R}$, which guarantees that the LLSE and MLSE have the same convergence rate. Assumptions 3.5(b)–3.5(d) are standard and roughly a subset of Assumption 1 in Hansen (2000); see Section 3.1 of Hansen (2000) for more discussions. Assumption 3.5(e) guarantees that the minimiser of the localised objective function is unique. This form of $z_{\ell i}$ is also used in Chan (1993). When $\overline{z}_{\ell i}$ and q_i have a joint continuous distribution, $z_{\ell i}$ follows the conditional distribution of $\overline{z}_{\ell i}$ given $q_i = \gamma_0$.

3.2. Large sample theory for the LSEs

This subsection discusses the large sample properties of the LSEs. First, Theorem 3.1 states the weak limit of the localised objective function when Assumption 3.3 is satisfied, which is critical in deriving the asymptotic distributions of the LSEs given in Theorem 3.2. Define $h = (u_1', u_2', v)' \equiv (u', v)'$ as the local parameter for θ.

THEOREM 3.1. *Suppose Assumptions 3.1–3.3 and 3.5 hold, then*

$$nP_n\left(m\left(\cdot|\beta_0 + \frac{u}{\sqrt{n}}, \gamma_0 + \frac{v}{c_n}\right) - m(\cdot|\beta_0, \gamma_0)\right)$$

converges weakly to

$$u_1' E[x_i x_i' \mathbf{1}(q_i \leq \gamma_0)]u_1 + u_2' E[x_i x_i' \mathbf{1}(q_i > \gamma_0)]u_2 - 2\sigma_{10}u_1'W_1 - 2\sigma_{20}u_2'W_2 + D(v)$$

for h on any compact set in $\mathbb{R}^{2k} \times \Xi$, where $W_1 \sim N(\mathbf{0}, E[xx'e^2\mathbf{1}(q \leq \gamma_0)])$, $W_2 \sim N(\mathbf{0}, E[xx'e^2\mathbf{1}(q > \gamma_0)])$ and $D(v)$ and Ξ depend on the local behaviour of $F(\cdot)$ around γ_0:

$$\text{(a) } \textit{If } \frac{a_n}{b_n} \to 1, \textit{ then } \Xi = (-\infty, \infty), \textit{ and } D(v) = \begin{cases} \sum_{i=1}^{N_1(|v|)} z_{1i} & \textit{if } v \le 0; \\ \sum_{i=1}^{N_2(v)} z_{2i} & \textit{if } v > 0; \end{cases}$$

$$\text{(b) } \textit{If } \frac{a_n}{b_n} \to 0, \textit{ then } \Xi = (-\infty, 0], \textit{ and } D(v) = \sum_{i=1}^{N_1(|v|)} z_{1i} \quad \textit{for } v \le 0;$$

$$\text{(c) } \textit{If } \frac{a_n}{b_n} \to \infty, \textit{ then } \Xi = [0, \infty), \textit{ and } D(v) = \sum_{i=1}^{N_2(v)} z_{2i} \quad \textit{for } v \ge 0.$$

Here, $D(v)$ in all the cases is the cadlag version of a compound Poisson process with $D(0) = 0$, $N_1(|\cdot|)$ is a reverse-time Poisson process with the integrated rate function $\Lambda_1(\cdot)$, and $N_2(\cdot)$ is a Poisson process with the integrated rate function $\Lambda_2(\cdot)$. Furthermore, W_1, W_2, $\{z_{1i}\}_{i\ge1}$, $\{z_{2i}\}_{i\ge1}$, $N_1(|\cdot|)$, $N_2(\cdot)$ are independent of each other.

In the classical case, $N_\ell(|\cdot|)$ is a homogeneous Poisson process with intensity $f(\gamma_0)$. But in general, $N_\ell(|\cdot|)$ can be a non-homogeneous Poisson process with the rate parameter $\lambda_\ell(v) \equiv |\frac{d\Lambda_\ell(v)}{dv}| = \alpha_\ell |v|^{\alpha_\ell - 1}$. The case with a point mass at γ_0 falls into (c) of Theorem 3.1 since we define $a_n = \infty$.

The independence between W_1 (W_2) and $D(\cdot)$ happens in the classical case discussed in Yu (2012). It also happens in the non-standard cases in this paper since γ is essentially a 'middle' boundary of q as shown in Yu (2012), and only the local information around γ_0 is informative to the threshold point. From Lemma 21.19 of Van der Vaart (1998), estimators of a boundary and a regular parameter are asymptotically independent no matter the distribution around the boundary is, so the independence between W_1 (W_2) and $D(\cdot)$ is much expected.

Now, we give the asymptotic distributions of the LSEs of γ_0 when $F(\cdot)$ satisfies Assumption 3.3.

THEOREM 3.2. *Suppose Assumptions 3.1–3.3 and 3.5 hold, then $\widehat{\gamma}_{\text{LLSE}} - \gamma_0 = O_p(c_n^{-1})$, and $\widehat{\gamma}_{\text{MLSE}} - \gamma_0 = O_p(c_n^{-1})$. Furthermore,*

$$c_n(\widehat{\gamma}_{\text{LLSE}} - \gamma_0) \xrightarrow{d} M_- \equiv Z_L, \; c_n(\widehat{\gamma}_{\text{MLSE}} - \gamma_0) \xrightarrow{d} \frac{M_- + M_+}{2} \equiv Z_M,$$

where $[M_-, M_+)$ is the minimising interval of $D(v)$ with $D(v)$ defined in Theorem 3.1.

The explicit forms of the density of Z_L and Z_M are given in the Supplementary Information. According to the algorithms there, Z_M is always continuously distributed, while Z_L is a mixture of continuous and discrete in Theorem 3.1(c). Theorem 3.2 critically relies on the fact that $N_\ell(|\cdot|)$ is a genuine Poisson process. When there is a jump in $\Lambda_\ell(\cdot)$, $N_\ell(|\cdot|)$ is not really a Poisson process. This is because the *orderliness* condition of the Poisson process is violated. The orderliness condition requires that jumps do not occur simultaneously; mathematically, that is, $\lim_{\Delta v \to 0} P(N_\ell(v + \Delta v) - N_\ell(v) > 1 | N_\ell(v + \Delta v) - N_\ell(v) \ge 1) = 0$. This condition is not satisfied when there is a jump in $\Lambda_\ell(\cdot)$.[6] In this case, the following theorem can be applied to find the asymptotic distributions of the LSEs.

[6] Conceptually, $N_\ell(|\cdot|)$ is a Poisson process or Poisson random measure (PRM) with mean intensity measure μ_ℓ, where μ_ℓ is defined by $\mu_1([v, 0)) = \Lambda_1(v)$, $\mu_2((0, v]) = \Lambda_2(v)$. When $\Lambda_\ell(\cdot)$ falls into type (b) of Theorem 2.1, $N_\ell(\cdot)$ is not Radon or simple. So $N_\ell(\cdot)$ is not the PRM defined on page 130 of Resnick (1987). As a result, the criteria for weak convergence in Section 3.5 of Resnick (1987) cannot be applied.

THEOREM 3.3. *Suppose Assumptions 3.1, 3.2, 3.4 and 3.5 hold, then the asymptotic distributions of* $\widehat{\gamma}_{\text{LLSE}}$ *and* $\widehat{\gamma}_{\text{MLSE}}$ *are discrete, and the support and point masses in their asymptotic distributions depend on the behaviour of* $F(\cdot)$ *in the neighbourhood of* γ_0*:*

(a) If $\frac{a_n}{b_n} \to 1$*, then*

$$\begin{aligned}
P\left(c_n\left(\widehat{\gamma}_{\text{LLSE}} - \gamma_0\right) = -1\right) &\to P\left(D(1) \geq \min\left\{D(-1), 0\right\}\right), \\
P\left(c_n\left(\widehat{\gamma}_{\text{LLSE}} - \gamma_0\right) = 1\right) &\to P\left(D(1) < \min\left\{D(-1), 0\right\}\right), \\
P\left(c_n\left(\widehat{\gamma}_{\text{MLSE}} - \gamma_0\right) = -1\right) &\to P\left(D(-1) < \min\left\{D(1), 0\right\}\right), \\
P\left(c_n\left(\widehat{\gamma}_{\text{MLSE}} - \gamma_0\right) = 0\right) &\to P\left(D(1) \geq 0, D(-1) \geq 0\right), \\
P\left(c_n\left(\widehat{\gamma}_{\text{MLSE}} - \gamma_0\right) = 1\right) &\to P\left(D(1) < \min\left\{D(-1), 0\right\}\right).
\end{aligned}$$

(b) If $\frac{a_n}{b_n} \to 0$*, then*

$$\begin{aligned}
P\left(c_n\left(\widehat{\gamma}_{\text{LLSE}} - \gamma_0\right) = -1\right) &\to 1, \\
P\left(c_n\left(\widehat{\gamma}_{\text{MLSE}} - \gamma_0\right) = -1\right) &\to P\left(D(-1) < 0\right), \\
P\left(c_n\left(\widehat{\gamma}_{\text{MLSE}} - \gamma_0\right) = -1/2\right) &\to P\left(D(-1) \geq 0\right).
\end{aligned}$$

(c) If $\frac{a_n}{b_n} \to \infty$*, then*

$$\begin{aligned}
P\left(c_n\left(\widehat{\gamma}_{\text{LLSE}} - \gamma_0\right) = 0\right) &\to P\left(D(1) \geq 0\right), \\
P\left(c_n\left(\widehat{\gamma}_{\text{LLSE}} - \gamma_0\right) = 1\right) &\to P\left(D(1) < 0\right), \\
P\left(c_n\left(\widehat{\gamma}_{\text{MLSE}} - \gamma_0\right) = 1/2\right) &\to P\left(D(1) \geq 0\right), \\
P\left(c_n\left(\widehat{\gamma}_{\text{MLSE}} - \gamma_0\right) = 1\right) &\to P\left(D(1) < 0\right).
\end{aligned}$$

Here, $D(1) = \sum_{i=1}^{P_2(\kappa_2)} z_{2i}$, $D(-1) = \sum_{i=1}^{P_1(\kappa_1)} z_{1i}$, $P_\ell(\kappa_\ell)$ *is a Poisson random variable with mean* κ_ℓ*, and* $\{z_{1i}\}_{i\geq 1}$, $\{z_{2i}\}_{i\geq 1}$, $P_1(\kappa_1)$, $P_2(\kappa_2)$ *are independent of each other. When* $\kappa_1 = 0$, $D(-1) \equiv 0$*; when* $\kappa_1 = \infty$, $D(-1) \equiv \infty$*. Similar conventions apply to* κ_2 *and* $D(1)$*.* a_n *and* b_n *are regularised as mentioned after Assumption 3.4.*

The asymptotic distributions in Theorem 3.1(b), (c) and Theorem 3.3(a)–(c) can be treated as degenerate forms of Theorem 3.1(a) if we define $D(v) = 0$ when $\Lambda_\ell(\cdot) = 0$, and $D(v) = \infty$ when $\Lambda_\ell(\cdot) = \infty$. But it should be noted that the proof in Theorem 3.1(a) cannot be extended to other cases since the stochastic equicontinuity fails when $\Lambda(\cdot)$ is not continuous. Next, we state the asymptotic distributions of $\widehat{\beta}$ and its functionals.

THEOREM 3.4. *Suppose Assumptions 3.1–3.3 (or 3.4) and 3.5 hold, then* $\widehat{\beta}_\ell - \beta_{\ell 0} = O_p(n^{-1/2})$*, and*

$$\begin{aligned}
\sqrt{n}\left(\widehat{\beta}_1 - \beta_{10}\right) &\xrightarrow{d} Z_{\beta_1} \sim E\left[xx'\mathbf{1}\left(q \leq \gamma_0\right)\right]^{-1} \cdot N\left(\mathbf{0}, E\left[xx'\sigma_{10}^2 e^2 \mathbf{1}\left(q \leq \gamma_0\right)\right]\right), \\
\sqrt{n}\left(\widehat{\beta}_2 - \beta_{20}\right) &\xrightarrow{d} Z_{\beta_2} \sim E\left[xx'\mathbf{1}\left(q > \gamma_0\right)\right]^{-1} \cdot N\left(\mathbf{0}, E\left[xx'\sigma_{20}^2 e^2 \mathbf{1}\left(q > \gamma_0\right)\right]\right),
\end{aligned}$$

which are the same as in the case where γ_0 *is known. Also,* $\widehat{\beta}_1$ *and* $\widehat{\beta}_2$ *are asymptotically independent of each other and* $\widehat{\gamma}_{\text{LLSE}}$ *(*$\widehat{\gamma}_{\text{MLSE}}$*). Furthermore, the SATE and SATT have the*

following asymptotic distributions:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} x_i'\left(\widehat{\beta}_2-\widehat{\beta}_1\right)-E\left[x\right]'\left(\beta_{20}-\beta_{10}\right)\right)\xrightarrow{d}E\left[x\right]'\left(Z_{\beta_2}-Z_{\beta_1}\right)+\left(\beta_{20}-\beta_{10}\right)'Z_x,$$

and

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} x_i'\left(\widehat{\beta}_2-\widehat{\beta}_1\right)\mathbf{1}\left(q_i>\widehat{\gamma}\right)-E\left[x\mathbf{1}\left(q>\gamma_0\right)\right]'\left(\beta_{20}-\beta_{10}\right)\right)$$
$$\xrightarrow{d}E\left[x\mathbf{1}\left(q>\gamma_0\right)\right]'\left(Z_{\beta_2}-Z_{\beta_1}\right)+\left(\beta_{20}-\beta_{10}\right)'Z_x^+,$$

where $Z_x\sim N(\mathbf{0},\,Var(x))$, $Z_x^+\sim N(\mathbf{0},\,Var(x\mathbf{1}(q>\gamma_0)))$, *and both* Z_x *and* Z_x^+ *are independent of* Z_{β_1} *and* Z_{β_2}.

From Theorem 3.4, the asymptotic distribution of $\widehat{\beta}_\ell$ is not affected by the estimation of γ_0 no matter what $f(q)$ around γ_0 is. The asymptotic distribution of the SATE and SATT includes two parts. The first part is from the random variation in $\widehat{\beta}_2-\widehat{\beta}_1$, and the second part is from the variation in $n^{-1}\sum_{i=1}^n x_i$ and $n^{-1}\sum_{i=1}^n x_i\mathbf{1}(q_i>\widehat{\gamma})$. For whatever SATE and SATT, $\widehat{\gamma}$ does not affect their asymptotic distributions.

4. NUMERICAL EXAMPLES, SIMULATION RESULTS AND APPLICATION

In this section, we first check the asymptotic distributions of the LSEs of γ in a threshold model with error term. This checking is based on the algorithms in the Supplementary Information where we also compare these asymptotic distributions with those in the simple model (2.1). We then conduct some Monte Carlo simulations to confirm the convergence rate of $\widehat{\gamma}$ stated in Theorems 3.2 and 3.3, and that the estimation of γ does not affect the efficiency of $\widehat{\beta}$ and the SATE and SATT. Finally, we apply the LSE to a real dataset and check whether the threshold variable has a continuous density at the threshold point.

4.1. Numerical examples

Suppose the population model is

$$y=\beta\mathbf{1}(q\le\gamma)+0.5e, \tag{4.1}$$

where $\beta_0=1$ and $\gamma_0=0$ are unknown, q follows the distributions as in the examples of Section 2.3 and $e\sim N(0,1)$ is independent of q. By Theorem 3.2 and 3.3, the asymptotic distributions of the LSEs of γ are

$$c_n\left(\widehat{\gamma}-\gamma_0\right)\xrightarrow{d}\arg\min_v D\left(v\right),$$

where

$$D(v) = \begin{cases} \sum_{i=1}^{N_1(|v|)} z_{1i} = \sum_{i=1}^{N_1(|v|)} \left(1 + e_i^-\right), & \text{if } v \le 0; \\ \sum_{i=1}^{N_2(v)} z_{2i} = \sum_{i=1}^{N_2(v)} \left(1 - e_i^+\right), & \text{if } v > 0; \end{cases}$$

$\{e_i^-, e_i^+, i = 1, 2, \ldots, N_1(\cdot), N_2(\cdot)\}$ are independent of each other, e_i^- and e_i^+ are i.i.d. copies of e, and $N_1(\cdot)$ and $N_2(\cdot)$ are Poisson processes with integrated intensities determined by $F(\cdot)$.

The asymptotic distributions of the LSEs of γ with $f(q)$ specified in Example 2.1, 2.2 and 2.3 are shown in Figure 1. Our first impression is that the asymptotic distribution can be continuous (Example 2.1 and the MLSE in Example 2.3), discrete (Example 2.2) or a mixture of discrete and continuous (the LLSE in Example 2.3). In Example 2.1, we only analyse Example 2.1(a) with $\lambda = 1$, $L(\cdot) = 1$ and different α values. An interesting phenomenon about the asymptotic distribution of $\widehat{\gamma}_{\text{LLSE}}$ when $\alpha = 2$ is that its density at zero is 0. This contrasts the usual asymptotic density whose mode is at zero. In Example 2.2, the asymptotic distribution of the MLSE is symmetric since $D(1)$ and $D(-1)$ have the same distribution in this example. Note also that the asymptotic distributions in Example 2.3 are not specific to the point mass at γ_0. As argued in Section 2.3 the asymptotic distributions are determined by $f(q)$ in the right neighbourhood of γ_0. So as long as $\lim_{q \uparrow 0} f(q) = \infty$, the asymptotic distributions in Example 2.3 are as shown in Figure 1.

4.2. *Simulation results*

Our first simulation is based on (4.1). In Example 2.1, to make sure the form of $F(x)$ in (2.6) is a genuine cdf, the support of q is constrained on $[-2^{-1/\alpha}, 2^{-1/\alpha}]$. In Examples 2.1 and 2.3, the sample size is set as 100 and 400, and in Example 2.2, the sample size is set as 100 and 1000. The number of repetition is 1000 in all examples. The simulation results are summarised in Table 1. $\widehat{\beta}_o$ in Table 1 is the LSE of β_0 when γ_0 is known. Since the RMSE is sensitive to outliers, we also report the median absolute error (MAE) of all estimators.

Table 1 confirms the theoretical analysis in Theorems 3.2, 3.3 and 3.4. First, by comparing the risk of $\widehat{\gamma}$ when $n = 100$ and $n = 400$, we can see that the convergence rate matches the theoretical prediction. Specifically, the convergence rate for Example 2.1 with $\alpha = 1$ and Example 2.3 is n, for Example 2.1 with $\alpha = 0.5$ is n^2, for Example 2.1 with $\alpha = 2$ is $\sqrt{n}$, and for Example 2.2 is $\ln n$. In Example 2.2, the convergence rate of $\widehat{\gamma}_{\text{MLSE}}$ seems faster than $\ln n$. This may be due to the fact that when $n = 100$, the threshold effect is not large enough to apply Theorem 3.3. In other words, the convergence rate when $n = 100$ is supposed to be slower than $\ln n$, while when $n = 1000$, $\ln n$ is a good approximation of the convergence rate. To confirm this result, we also check two other setups $\beta_0 = 0.5$ and 2 in the Supplementary Information. The simulation results there show that the convergence rate when $\beta_0 = 2$ is roughly $\ln n$ while it is not the case when $\beta_0 = 0.5$. Second, the risk of $\widehat{\gamma}_{\text{MLSE}}$ is smaller than that of $\widehat{\gamma}_{\text{LLSE}}$ except in Example 2.3. From the asymptotic distributions of $\widehat{\gamma}_{\text{LLSE}}$ and $\widehat{\gamma}_{\text{MLSE}}$ in Figure 1, this result is obvious. Third, for $\widehat{\beta}$, the convergence rate is obviously $\sqrt{n}$ except in Example 2.2. This means that the imprecision of $\widehat{\gamma}$ in Example 2.2 indeed affects the estimation of β in *finite samples*. Fourth, the risks of $\widehat{\beta}$ and $\widehat{\beta}_o$ in all cases are almost the same, which confirms that $\widehat{\beta}$ is adaptive to the estimation of γ_0. Also, the risks of $\widehat{\beta}$ in all examples (except Example 2.2) are quite

Table 1. Performances of $\widehat{\gamma}$ and $\widehat{\beta}$: $\beta_0 = 1$.

Risk ($\times 10^{-2}$) →	RMSE				MAE			
Examples↓ Estimators→	$\widehat{\gamma}_{\text{LLSE}}$	$\widehat{\gamma}_{\text{MLSE}}$	$\widehat{\beta}$	$\widehat{\beta}_o$	$\widehat{\gamma}_{\text{LLSE}}$	$\widehat{\gamma}_{\text{MLSE}}$	$\widehat{\beta}$	$\widehat{\beta}_o$
$n = 100$								
Example 2.2: $\alpha = 0.5$	0.113	0.106	7.716	7.612	0.009	0.009	5.286	5.169
Example 2.2: $\alpha = 1$	1.973	1.756	7.370	7.203	0.919	0.562	5.200	5.013
Example 2.2: $\alpha = 2$	11.835	8.945	7.137	7.036	9.699	3.738	4.742	4.627
Example 2.2	27.513	20.687	7.326	7.291	24.919	6.728	5.046	5.018
Example 2.3	1.035	1.351	7.315	7.252	0	0.455	4.807	4.711
$n = 400$								
Example 2.2: $\alpha = 0.5$	0.0054	0.0044	3.737	3.726	0.0005	0.0006	2.542	2.537
Example 2.2: $\alpha = 1$	0.510	0.464	3.416	3.427	0.224	0.156	2.286	2.292
Example 2.2: $\alpha = 2$	6.017	4.722	3.552	3.548	4.751	2.148	2.356	2.339
Example 2.2 ($n = 1000$)	16.253	10.549	2.254	2.248	15.951	2.126	1.477	1.472
Example 2.3	0.229	0.315	3.688	3.675	0	0.120	2.459	2.468

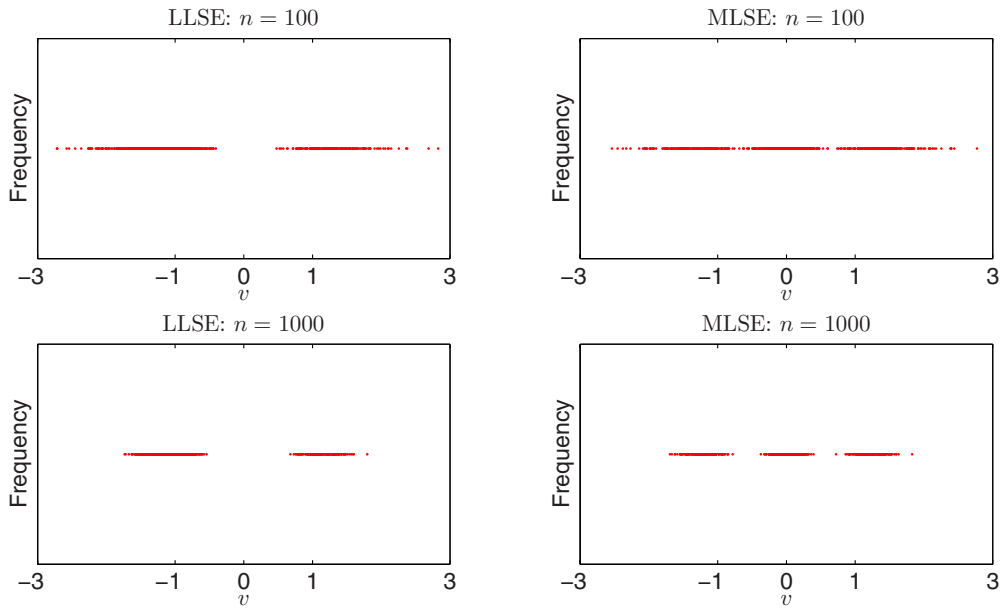

Figure 3. Scatterplot of $\ln n(\widehat{\gamma} - \gamma_0)$ in Example 2.2.

similar, which confirms that the asymptotic distributions of $\widehat{\beta}$ in all examples are the same. Fifth, as expected, the risks of $\widehat{\beta}$ and $\widehat{\beta}_o$ are closer to each other when n and/or β_0 are larger.

Given the speciality of Example 2.2 in the above simulation, we give more discussion on the slow convergence of $\widehat{\gamma}$ here. Figure 3 shows plots of $\ln n(\widehat{\gamma}_{\text{LLSE}} - \gamma_0)$ and $\ln n(\widehat{\gamma}_{\text{MLSE}} - \gamma_0)$

Table 2. RMSE of $\widehat{\gamma}$ and the SATE and SATT in 10^{-2}.

Examples↓ Estimators→	$\widehat{\gamma}_{\text{LLSE}}$	$\widehat{\gamma}_{\text{MLSE}}$	*SATE*	$SATE_o$	*SATT*	$SATTE_o$
$n = 100$						
Example 2.2: $\alpha = 0.5$	0.076	0.059	22.637	22.640	17.866	17.856
Example 2.2: $\alpha = 1$	1.580	1.137	21.392	21.418	17.622	17.621
Example 2.2: $\alpha = 2$	10.699	6.135	22.577	22.581	18.117	18.114
Example 2.2	25.875	13.951	22.110	22.109	18.426	18.398
Example 2.3	1.570	1.998	22.770	22.479	15.140	14.946
$n = 400$						
Example 2.2: $\alpha = 0.5$	0.003	0.004	10.743	10.730	8.716	8.715
Example 2.2: $\alpha = 1$	0.392	0.288	11.001	10.996	9.007	9.006
Example 2.2: $\alpha = 2$	5.218	3.231	11.277	11.275	9.025	9.025
Example 2.2 ($n = 1000$)	15.841	7.750	7.055	7.052	5.595	5.597
Example 2.3	0.405	0.519	11.194	11.129	7.504	7.499

for $n = 100$ and 1000 in our simulations. Even for a sample size as large as 1000, there is little evidence that the limiting distribution is a good approximation to the true distribution which is discrete. This is understandable since $\ln 1000 = 1.5 \ln 100$ in this example, while $\sqrt{1000} = 3.16\sqrt{100}$ in regular cases.

We next check the effect of $\widehat{\gamma}$ on the SATE and SATT when a regressor x is added in the regression, i.e.

$$y = (1\ x)\,\beta_1 \mathbf{1}(q \le \gamma) + (1\ x)\,\beta_2 \mathbf{1}(q > \gamma) + 0.5e, \tag{4.2}$$

where $\beta_{10} = (-1\ -1)'$, $\beta_{20} = (1\ 1)'$, $x \sim N(0, 1)$ and is independent of q and e. It can be shown that $PATE = 2$ and $PATT = 1$ in this model. The simulation results are summarised in Table 2. To save space, we only report the RMSE of all estimators, leaving the results about the MAE in the Supplementary Information. From Table 2, the convergence rates of $\widehat{\gamma}$ match the theoretical analysis. Also, there is no significant evidence that the estimation of γ will affect the efficiency of the SATE and SATT.

4.3. Application

We now apply the LSE to the growth data used in Durlauf and Johnson (1995) and reanalysed in Hansen (2000) and Yu (2008). Only the MLSE is considered here to save space. The growth theory with multiple equilibria motivates the following threshold regression model:

$$\begin{aligned}
&\ln\left(\frac{Y}{L}\right)_{i,1985} - \ln\left(\frac{Y}{L}\right)_{i,1960} \\
&= \begin{cases}
\beta_{10} + \beta_{11} \ln\left(\dfrac{Y}{L}\right)_{i,1960} + \beta_{12} \ln\left(\dfrac{I}{Y}\right)_i + \beta_{13} \ln(n_i + g + \delta) + \beta_{14} \ln S_i + \sigma_1 e_i, \text{ if } q_i \le \gamma; \\
\beta_{20} + \beta_{21} \ln\left(\dfrac{Y}{L}\right)_{i,1960} + \beta_{22} \ln\left(\dfrac{I}{Y}\right)_i + \beta_{23} \ln(n_i + g + \delta) + \beta_{24} \ln S_i + \sigma_2 e_i, \text{ if } q_i > \gamma.
\end{cases}
\end{aligned}$$

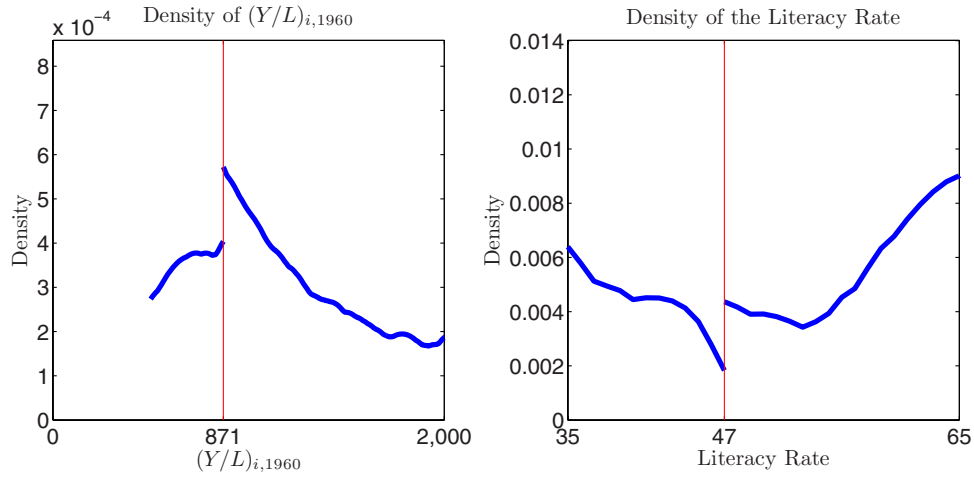

Figure 4. Under-smoothing densities of $(Y/L)_{i,1960}$ and the literacy rate.

For each country i, $(Y/L)_{i,t}$ is the real GDP per member of the population aged 15–64 in year t, $(I/Y)_i$ is the investment to GDP ratio, n_i is the growth rate of the working-age population, and S_i is the fraction of working-age population enrolled in secondary schools. The variables not indexed by t are annual averages over the period 1960–1985. Following Durlauf and Johnson (1995), we set $g + \delta = 0.05$. The sample size is 96, and the data are assumed to be i.i.d. sampled. This assumption is approximately true, since there are not many interactions, such as trade, international capital flows, etc., between any two countries during this period.

Hansen (2000) considers two choices of the threshold variable q. The first choice is the starting GDP $(Y/L)_{i,1960}$, and the corresponding $\widehat{\gamma}_{\text{MLSE}} = 871$. The right regime ($(Y/L)_{i,1960} > 871$) includes 78 countries, and the specification testing suggests that the literacy rate is a suitable threshold variable for this regime. The corresponding $\widehat{\gamma}_{\text{MLSE}} = 47\%$. We now check whether $(Y/L)_{i,1960}$ has a continuous density at 871 and the literacy rate has a continuous density at 47%. The kernel smoother is used to estimate these two densities. To avoid the bias problem in the density estimation at the edge, we use small-than-optimal bandwidths. This under-smoothing can also save the effort to estimate the biases in the testing procedure; see page 703 of McCrary (2008) for further discussions on this point. In our application, we use half of the optimal bandwidths as suggested by Hall (1992). As to the kernel function, we use the rescaled Epanechnikov kernel to adjust the density estimation at the boundary area. The under-smoothing densities in the neighbourhoods of the threshold points are shown in Figure 4. Suppose these densities are between 0 and ∞, then the t-test statistics in testing $f(\gamma_0+) = f(\gamma_0-)$ are -1.45 and -1.41, respectively. The one-side p-values are around 7.5%, so we marginally reject the null at the significance level 5%. This suggests that we at least should take caution in the inference of these threshold regressions.

5. CONCLUSION

This paper examines the sensitivity of the asymptotic distributions of the LSEs of the threshold point to the density of the threshold variable. Specifically, we show that depending on the density

of the threshold variable around the threshold point, the convergence rate can be faster or slower than n; the asymptotic distribution can be discrete, continuous or a mixture of discrete and continuous; the weak limits of the localised objective functions can be non-homogeneous instead of homogeneous compound Poisson processes. These features suggest cautions in the use of inference methods in the classical model. A positive result is that if only the regular parameters or their functionals are of the main interest, then any inference method in the classical model can be applied. Potential applications of the present methodology include the treatment effects models with abnormalities in the treatment assignment variable.

In spite of many theoretical results in this paper, there remain two unsolved problems. The first problem is how to check whether $f(\gamma_0) \in (0, \infty)$ or not. Although Han et al (2011) provide a test statistic in the median regression context by splitting the sample, their method cannot be extended to threshold regression. Nevertheless, since it is easy to estimate $f(q)$ in the left and right neighbourhoods of γ_0 using kernel smoothing or series approximation, it should become a routine to check the shape of $f(q)$ around γ_0 as in the application of Section 4.3 The second problem is the confidence interval construction of γ. In the classical model, Section 4.1 of Yu (2013) summarises five confidence interval construction methods. But all these methods rely on the assumption that $0 < \underline{f} \leq f(q) \leq \overline{f} < \infty$ and cannot be extended to the cases considered in this paper.

ACKNOWLEDGMENTS

This paper is inspired by talks with Han Hong and Peter Phillips. Ping Yu thanks Bruce Hansen, Jack Porter, Andres Aradillas-Lopez, Rustam Ibragimov and seminar participants at UW-Madison and NZESG for helpful comments. The authors also thank the co-editor and two referees for very helpful comments.

REFERENCES

Aldous, D. (1978). Stopping times and tightness. *Annals of Probability 6*, 335–40.

Bai, J. (1997). Estimation of a change point in multiple regression models. *Review of Economics and Statistics 79*, 551–63.

Chan, K.S. (1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *Annals of Statistics 21*, 520–33.

Cheng, M. Y., J. Fan, and J. S. Marron (1997). On automatic boundary corrections. *Annals of Statistics 25*, 1691–708.

Cho, J. S., C. Han, and P. C. B. Phillips (2010). LAD asymptotics under conditional heteroskedasticity with possibly infinite error densities. *Econometric Theory 26*, 953–62.

de Haan, L. (1970). *On Regular Variation and Its Application to the Weak Convergence of Sample Extremes*. Mathematical Centre Tract 32, Amsterdam: Mathematisch Centrum.

Durlauf, S. N. and P. A. Johnson (1995). Multiple regimes and cross-country growth behavior. *Journal of Applied Econometrics 10*, 365–84.

Frölich, M. (2007). Regression Discontinuity Design with Covariates, IZA Discussion Paper No. 3024, Institute for the Study of Labor (IZA).

Gnedenko, B. V. (1943). Sur la distribution limité du terme d' une série aléatoire. *Annals of Mathematics* 44, 423–53. [Translated and reprinted in S. Kotz and N.L. Johnson (Eds.), 1992, *Breakthroughs in Statistics*, *Volume I*, 195–225. Berlin: Springer.]

Hall, P. (1992). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *Annals of Statistics 20*, 675–94.

Han, C., J. S. Cho, and P. C. B. Phillips (2011). Infinite density at the median and the typical shape of stock return distributions. *Journal of Business and Economic Statistics 29*, 282–94.

Hansen, B. E. (2000). Sample splitting and threshold estimation. *Econometrica 68*, 575–603.

Hansen, B. E. (2011). Threshold autoregression in economics. *Statistics and Its Interface 4*, 123–7.

Hoekstra, M. (2009). The effect of attending the flagship state university on earnings: a discontinuity-based approach. *Review of Economics and Statistics 91*, 717–24.

Knight, K. (1997). Some limit theory for L_1-estimators in autoregressive models under general conditions. In Y. Dodge (Ed.), *L_1-Statistical Procedures and Related Topics*, IMS Lecture Notes—Monograph Series, *Volume 31*, 315–28. Hayward, CA: Institute of Mathematical Statistics.

Knight, K. (1998). Limiting distributions for L_1 regression estimators under general conditions. *Annals of Statistics 26*, 755–70.

Knight, K. (1999). Asymptotics for L_1-estimators of regression parameters under heteroscedasticity. *Canadian Journal of Statistics 27*, 497–507.

Lee, S. and M. H. Seo (2008). Semiparametric estimation of a binary response model with a change-point due to a covariate threshold. *Journal of Econometrics 144*, 492–99.

McCrary, J. (2008). Testing for manipulation of the running variable in the regression discontinuity design. *Journal of Econometrics 142*, 698–714.

Newey, W. K. and D. L. McFadden (1994). Large sample estimation and hypothesis testing. In R. F. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, *Volume 4*, 2113–2245. New York: Elsevier Science.

Pakes, A. and D. Pollard (1989). Simulation and the asymptotics of optimization estimators. *Econometrica 57*, 1027–57.

Politis, D. N., J. P. Romano, and M. Wolf (1999). *Subsampling*. New York, NY: Springer.

Pollard, D. (1984). *Convergence of Stochastic Processes*. New York, NY: Springer.

Porter, J. and P. Yu (2012). Regression discontinuity designs with unknown discontinuity points: testing and estimation. Working paper, Department of Economics, University of Auckland.

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

Rogers, A. (2001). Least absolute deviations regression under nonstandard conditions. *Econometric Theory 17*, 820–52.

Seneta, E. (1976). *Regularly varying functions*. *Lecture Notes in Mathematics*, Volume *508*. Berlin: Springer.

Tong, H. (1978). On a threshold model. In C. H. Chen (Ed.), *Pattern Recognition and Signal Processing*, 575–86. Amsterdam: Sijthoff and Noordhoff.

Tong, H. (1983). *Threshold Models in Nonlinear Time Series Analysis*. *Lecture Notes in Statistics,* No. 21. New York, NY: Springer.

Tong, H. (1990). *Non-linear Time Series: A Dynamical System Approach*. Oxford: Clarendon Press.

Tong, H. (2011). Threshold models in time series analysis—30 years on. *Statistics and Its Interface 4*, 107–18.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. New York, NY: Cambridge University Press.

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

Yu, P. (2008). Adaptive estimation of the threshold point in threshold regression. Working paper, Department of Economics, University of Auckland.

Yu, P. (2012). Likelihood estimation and inference in threshold regression. *Journal of Econometrics 167*, 274–94.

Yu, P. (2013). The bootstrap in threshold regression. Forthcoming in *Econometric Theory*.
Zhu, X. (2009). Singularity and regularity of solutions to a certain class of degenerate elliptic equations. *Nonlinear Analysis 71*, 646–53.

APPENDIX A: PROOFS

First, some notations are collected for reference in all lemmas and proofs.

$$\begin{aligned}
\theta_\ell &= \left(\beta_\ell', \sigma_\ell\right)', \\
m\left(w|\theta\right) &= \left(y - x'\beta_1 \mathbf{1}(q \le \gamma) - x'\beta_2 \mathbf{1}(q > \gamma)\right)^2, \\
M_n\left(\theta\right) &= P_n\left(m\left(\cdot|\theta\right)\right), \\
M\left(\theta\right) &= P\left(m\left(\cdot|\theta\right)\right), \\
\mathbb{G}_n m &= \sqrt{n}\left(M_n - M\right). \\
\overline{z}_1\left(w|\theta_2, \widetilde{\theta}_1\right) &= \left(\widetilde{\beta}_1 - \beta_2\right)' xx' \left(\widetilde{\beta}_1 - \beta_2\right) + 2\widetilde{\sigma}_1 \left(\widetilde{\beta}_1 - \beta_2\right) xe, \\
\overline{z}_2\left(w|\theta_1, \widetilde{\theta}_2\right) &= \left(\widetilde{\beta}_2 - \beta_1\right)' xx' \left(\widetilde{\beta}_2 - \beta_1\right) + 2\widetilde{\sigma}_2 \left(\widetilde{\beta}_2 - \beta_1\right) xe.
\end{aligned}$$

Thus $\overline{z}_{1i} = \overline{z}_1\left(w_i|\theta_{20}, \theta_{10}\right)$ and $\overline{z}_{2i} = \overline{z}_2\left(w_i|\theta_{10}, \theta_{20}\right)$.

The following formulae are used repetitively:

$$\begin{aligned}
m\left(w|\theta\right) &= \left(x'\left(\beta_{10} - \beta_1\right) + \sigma_{10}e\right)^2 \mathbf{1}(q_i \le \gamma \wedge \gamma_0) + \left(x'\left(\beta_{20} - \beta_2\right) + \sigma_{20}e\right)^2 \mathbf{1}(q_i > \gamma \vee \gamma_0) \\
&\quad + \left(x'\left(\beta_{10} - \beta_2\right) + \sigma_{10}e\right)^2 \mathbf{1}(\gamma \wedge \gamma_0 < q_i \le \gamma_0) + \left(x'\left(\beta_{20} - \beta_1\right) + \sigma_{20}e\right)^2 \mathbf{1}(\gamma_0 < q \le \gamma \vee \gamma_0).
\end{aligned}$$

So

$$\begin{aligned}
m\left(w|\theta\right) - m\left(w|\theta_0\right) &= \left[\left(\beta_{10} - \beta_1\right)' xx' \left(\beta_{10} - \beta_1\right) + 2\sigma_{10}\left(\beta_{10} - \beta_1\right) xe\right] \mathbf{1}(q \le \gamma \wedge \gamma_0) \\
&\quad + \left[\left(\beta_{20} - \beta_2\right)' xx' \left(\beta_{20} - \beta_2\right) + 2\sigma_{20}\left(\beta_{20} - \beta_2\right) xe\right] \mathbf{1}(q > \gamma \vee \gamma_0) \\
&\quad + \overline{z}_1\left(w|\theta_2, \theta_{10}\right) \mathbf{1}(\gamma \wedge \gamma_0 < q \le \gamma_0) + \overline{z}_2\left(w|\theta_1, \theta_{20}\right) \mathbf{1}(\gamma_0 < q \le \gamma \vee \gamma_0) \\
&\equiv T\left(w|\theta_1, \theta_{10}\right) \mathbf{1}(q \le \gamma \wedge \gamma_0) + T\left(w|\theta_2, \theta_{20}\right) \mathbf{1}(q > \gamma \vee \gamma_0) \\
&\quad + \overline{z}_1\left(w|\theta_2, \theta_{10}\right) \mathbf{1}(\gamma \wedge \gamma_0 < q \le \gamma_0) + \overline{z}_2\left(w|\theta_1, \theta_{20}\right) \mathbf{1}(\gamma_0 < q \le \gamma \vee \gamma_0) \\
&\equiv A\left(w|\theta\right) + B\left(w|\theta\right) + C\left(w|\theta\right) + D\left(w|\theta\right).
\end{aligned}$$

$$D_n\left(v\right) = \sum_{i=1}^n \overline{z}_{1i} \mathbf{1}\left(\gamma_0 + \frac{v}{c_n} < q_i \le \gamma_0\right) + \sum_{i=1}^n \overline{z}_{2i} \mathbf{1}\left(\gamma_0 < q_i \le \gamma_0 + \frac{v}{c_n}\right)$$

$$N_{1n}\left(v\right) = \sum_{i=1}^n \mathbf{1}\left(\gamma_0 + \frac{v}{c_n} < q_i \le \gamma_0\right), N_{2n}\left(v\right) = \sum_{i=1}^n \mathbf{1}\left(\gamma_0 < q_i \le \gamma_0 + \frac{v}{c_n}\right)$$

$$W_{1n} = \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i e_i \mathbf{1}(q_i \le \gamma_0) = \sum_{i=1}^n \frac{S_{1i}}{\sqrt{n}},$$

$$W_{2n} = \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i e_i \mathbf{1}\left(q_i > \gamma_0\right) = \sum_{i=1}^n \frac{S_{2i}}{\sqrt{n}},$$

$$W_i = \left(\frac{S_{1i}'}{\sqrt{n}} \; \frac{S_{2i}'}{\sqrt{n}}\right)' \equiv S_i/\sqrt{n}, S_i = \left(S_{1i}' \; S_{2i}'\right)'.$$

Proof of Theorem 2.1: The proof idea follows from Proposition 0.3 of Resnick (1987), which is borrowed from Gnedenko (1943). Since only the local behaviour of $F(\cdot)$ around γ_0 is relevant, assume $\gamma_0 = 0$, $F(0) = 1$. If we can find all possible limits of $F^n(\frac{x}{a_n})$ (which is the cdf of $a_n \max\{q_i\}$), then the possible $\Lambda_1(\cdot)$ can be recovered from (2.2). For any $t > 0$, we have

$$F^{[nt]}\left(\frac{x}{a_{[nt]}}\right) \to G(x) \text{ with } G(0) = 1$$

and also

$$F^{[nt]}\left(\frac{x}{a_n}\right) = \left(F^n\left(\frac{x}{a_n}\right)\right)^{[nt]/n} \to G^t(x).$$

By the convergence to type theorem (Proposition 0.2) of Resnick (1987),

$$G^t(x) = G(\theta(t)x), \tag{A.1}$$

where $\theta(t) = \lim_{n\to\infty} \frac{a_{[nt]}}{a_n} > 0$ is measurable.[7]

For $t > 0$, $s > 0$ we have on one hand

$$G^{ts}(x) = G\left(\theta(ts)x\right)$$

and on the other

$$G^{ts}(x) = (G^s(x))^t = G\left(\theta(s)x\right)^t = G\left(\theta(t)\theta(s)x\right).$$

If G is non-degenerate,

$$\theta(ts) = \theta(t)\theta(s). \tag{A.2}$$

If G is degenerate, we can always scale a_n such that G is a point mass at -1. In the former case, (A.2) is the famous *Hamel* functional equation. The only finite measurable non-negative solution is of the following form

$$\theta(t) = t^\alpha, \alpha \in \mathbb{R}.$$

We now consider three cases:

- CASE 1. $\alpha = 0$. In this case, $\theta(t) = 1$, so (A.1) becomes $G^t(x) = G(x)$. This is impossible for a non-degenerate G.
- CASE 2. $\alpha < 0$. In this case, $G^t(x) = G(t^\alpha x)$. Since G is non-degenerate, there exists $x_0 < 0$ such that $G(x_0) \in (0, 1)$. Since $G^t(x_0)$ is a decreasing function of t, while $G(t^\alpha x_0)$ is increasing with t, this case cannot happen.
- CASE 3. $\alpha > 0$. In this case, $G^t(t^{-\alpha}x) = G(x)$. It is well known that the only solution for this functional equation with boundary condition $G(0) = 1$ is $\exp\{-(-Ax)^\alpha\}$, where $A > 0$. This distribution is called *reversed Weibull* or *extreme value type III distribution*, and is denoted as $\Psi_\alpha(x)$ in Gnedenko (1943).

The domain of attraction of $\Psi_\alpha(x)$ is stated in Theorem 5 of Gnedenko (1943); see also Proposition 1.13 of Resnick (1987). Actually, we can give a simpler proof by taking advantage of the speciality of the present case. If $1 - F(-x) \in RV_\alpha$ as $x \downarrow 0$, set $a_n = -\frac{1}{(1/(1-F))^{\leftarrow}(n)}$. Then because $1 - F(-\frac{1}{a_n}) \sim n^{-1}$,

[7] Note that the drift term b_n in Resnick (1987) is fixed in our case, so $\beta(t) = 0$. $\theta(t)$ plays the role of $\alpha(t)$ in his proof. We avoid using $\alpha(t)$ here since α is used for the exponent of variation in the definition of regularly varying functions.

we have that for $x > 0$, $n(1 - F(-\frac{x}{a_n})) \sim (1 - F(-\frac{x}{a_n}))/(1 - F(-\frac{1}{a_n})) \to x^\alpha$ as $n \to \infty$ since $a_n \to \infty$. Conversely, if there exists $a_n \to \infty$ such that $n(1 - F(-\frac{x}{a_n})) \to x^\alpha$, then for any $x > 0$,

$$\lim_{t\downarrow 0} \frac{1 - F(-tx)}{1 - F(-t)} = \lim_{n\to\infty} \frac{n\left(1 - F(-\frac{tx}{a_n})\right)}{n\left(1 - F(-\frac{t}{a_n})\right)} = \frac{(tx)^\alpha}{t^\alpha} = x^\alpha.$$

The proof for the domain of attraction of type (b) follows in spirit from Theorem 2 of Gnedenko (1943). The sufficiency part is similar to case (a), so we switch to the necessity part. From the form of Λ_1, for any $\varepsilon > 0$,

$$n\left(1 - F\left(-\frac{1-\varepsilon}{a_n}\right)\right) \to 0 \text{ and } n\left(1 - F\left(-\frac{1+\varepsilon}{a_n}\right)\right) \to \infty. \tag{A.3}$$

Because a_n is non-decreasing, we can find an x such that $\frac{1}{a_n} \le x \le \frac{1}{a_{n-1}}$. Then for all $\varepsilon > 0$ and $\eta > 0$, we have

$$1 - F\left(-\frac{1-\eta}{a_{n-1}}\right) \le 1 - F\left(-x(1-\eta)\right) \le 1 - F\left(-\tfrac{1-\eta}{a_n}\right),$$

$$1 - F\left(-\frac{1+\varepsilon}{a_{n-1}}\right) \le 1 - F\left(-x(1+\varepsilon)\right) \le 1 - F\left(-\tfrac{1+\varepsilon}{a_n}\right),$$

from which it is seen that

$$\frac{1 - F\left(-\frac{1-\eta}{a_{n-1}}\right)}{1 - F\left(-\frac{1+\varepsilon}{a_n}\right)} \le \frac{1 - F\left(-x(1-\eta)\right)}{1 - F\left(-x(1+\varepsilon)\right)} \le \frac{1 - F\left(-\frac{1-\eta}{a_n}\right)}{1 - F\left(-\frac{1+\varepsilon}{a_{n-1}}\right)}.$$

From (A.3), for all $\varepsilon > 0$ and $\eta > 0$, we have

$$\lim_{x\downarrow 0} \frac{1 - F\left(-x(1-\eta)\right)}{1 - F\left(-x(1+\varepsilon)\right)} \to 0,$$

which is essentially the result we want. Similarly, we can prove $\lim_{x\downarrow 0} \frac{1-F(-x(1+\varepsilon))}{1-F(-x(1-\eta))} \to \infty$. So $1 - F(-x) \in RV_\infty$ as $x \downarrow 0$ □.

Proof of Theorem 3.1: Only Case (a) is proved, since the proof for the other two cases is similar. From Lemma B.1,

$$n P_n\left(m\left(\cdot\,\middle|\,\beta_0 + \frac{u}{\sqrt{n}}, \gamma_0 + \frac{v}{c_n}\right) - m\left(\cdot\,|\beta_0, \gamma_0\right)\right)$$

$$= u_1' E\left[x_i x_i' \mathbf{1}(q_i \le \gamma_0)\right] u_1 + u_2' E\left[x_i x_i' \mathbf{1}\left(q_i > \gamma_0\right)\right] u_2 - 2\sigma_{10} u_1' W_{1n} - 2\sigma_{20} u_2' W_{2n} + D_n(v) + o_p(1).$$

The characteristic function is used to find the asymptotic distribution of $D_n(v)$ and prove the asymptotic independence between W_{1n}, W_{2n} and $D_n(v)$. First, define two terms as follows:

$$T_{1i} = \overline{z}_{1i} \mathbf{1}\left(\gamma_0 + \frac{v_1}{c_n} < q_i \le \gamma_0\right),$$

$$T_{2i} = \overline{z}_{2i} \mathbf{1}\left(\gamma_0 < q_i \le \gamma_0 + \frac{v_2}{c_n}\right),$$

where $v_1 < 0$, $v_2 > 0$. Note that

$$\exp\left\{\sqrt{-1} t_1 T_{1i}\right\} = 1 + \mathbf{1}\left(\gamma_0 + \frac{v_1}{c_n} < q_i \le \gamma_0\right)\left[\exp\left\{\sqrt{-1} t_1 \overline{z}_{1i}\right\} - 1\right],$$

$$\exp\left\{\sqrt{-1}t_2T_{2i}\right\} = 1 + \mathbf{1}\left(\gamma_0 < q_i \le \gamma_0 + \frac{v_2}{c_n}\right)\left[\exp\left\{\sqrt{-1}t_2\bar{z}_{2i}\right\} - 1\right].$$

So

$$\begin{aligned}
&E[\exp\{\sqrt{-1}(t_1 \cdot T_{1i} + t_2 \cdot T_{2i} + s'W_i)\}] \\
&= E[\exp\{\sqrt{-1}t_1T_{1i}\}\exp\{\sqrt{-1}t_2T_{2,i}\}\exp\{\sqrt{-1}s'W_i\}] \\
&= E[\exp\{\sqrt{-1}s'W_i\}] + \frac{\Lambda_1(v_1)}{n}E[\exp\{\sqrt{-1}s'W_i\}\{\exp\{\sqrt{-1}t_1\bar{z}_{1i}\} - 1\}|q_i = \gamma_0-] \\
&\quad + \frac{\Lambda_2(v_2)}{n}E[\exp\{\sqrt{-1}s'W_i\}\{\exp\{\sqrt{-1}t_2\bar{z}_{2i}\} - 1\}|q_i = \gamma_0+] + o(\frac{1}{n}) \\
&= 1 + \frac{1}{n}\left[-\frac{1}{2}s'\mathcal{J}s + \Lambda_1(v_1)\left(E\left[\left\{\exp\left\{\sqrt{-1}t_1\bar{z}_{1i}\right\}\right\}|q_i = \gamma_0-\right] - 1\right)\right. \\
&\quad \left. + \Lambda_2(v_2)\left(E\left[\left\{\exp\left\{\sqrt{-1}t_2\bar{z}_{2i}\right\}\right\}|q_i = \gamma_0+\right] - 1\right)\right] + o\left(\frac{1}{n}\right),
\end{aligned}$$

where the last equality is from the Taylor expansion of $\exp\left\{\sqrt{-1}s'W_i\right\}$, and $\mathcal{J} = E\left[S_iS_i'\right]$. From the definition of $\Lambda_1(\cdot)$ and $\Lambda_2(\cdot)$, $o(1)$ in the second equality is a quantity going to zero uniformly over $i = 1, \ldots, n$. It follows that

$$\begin{aligned}
&E\left[\exp\left\{\sqrt{-1}\left(t_1 \cdot \sum_{i=1}^n T_{1i} + t_2 \cdot \sum_{i=1}^n T_{2i} + s'\sum_{i=1}^n W_i\right)\right\}\right] \\
&= \prod_{i=1}^n E\left[\exp\left\{\sqrt{-1}\left(t_1 \cdot T_{1i} + t_2 \cdot T_{2i} + s'W_i\right)\right\}\right] \\
&\to \exp\left\{-\frac{1}{2}s'\mathcal{J}s + \Lambda_1(v_1)\left(E\left[\left\{\exp\left\{\sqrt{-1}t_1\bar{z}_{1i}\right\}\right\}|q_i = \gamma_0-\right] - 1\right)\right. \\
&\quad \left. + \Lambda_2(v_2)\left(E\left[\left\{\exp\left\{\sqrt{-1}t_2\bar{z}_{2i}\right\}\right\}|q_i = \gamma_0+\right] - 1\right)\right\}.
\end{aligned}$$

It is easy to find that the characteristic function of $D(v)$ matches the limit of that of $D_n(v)$, so the finite-dimensional convergence of the objective empirical process is proved.

Now, let's consider the stochastic equicontinuity of $W_n(u)$ and $D_n(v)$. The stochastic equicontinuity of $W_n(u)$ can be trivially proved since they are linear functions of u, so here we concentrate on $D_n(v)$. For this purpose, a condition called Aldous's (1978) condition is sufficient; see Theorem 16 on Page 134 of Pollard (1984). Without loss of generality, we only prove the result for $v > 0$. Suppose $0 < v_1 < v_2$ are stopping times in $[0, K]$ with $K < \infty$, then for any $\varepsilon > 0$,

$$\begin{aligned}
&P\left(\sup_{|v_2 - v_1| < \delta}|D_n(v_2) - D_n(v_1)| > \varepsilon\right) \\
&\le P\left(\sum_{i=1}^n |\bar{z}_{2i}| \cdot \sup_{|v_2 - v_1| < \delta}\mathbf{1}\left(\gamma_0 + \frac{v_1}{a_n} < q_i \le \gamma_0 + \frac{v_2}{a_n}\right) > \varepsilon\right) \\
&\le \sum_{i=1}^n E\left[|\bar{z}_{2i}| \sup_{|v_2 - v_1| < \delta}\mathbf{1}\left(\gamma_0 + \frac{v_1}{a_n} < q_i \le \gamma_0 + \frac{v_2}{a_n}\right)\right] \Big/ \varepsilon \\
&\le \sup_{\gamma_0 < \gamma \le \gamma_0 + \epsilon} E\left[\,|\bar{z}_{2i}|\,|\,q_i = \gamma\right] \cdot E\left[n\sup_{|v_2 - v_1| < \delta}\mathbf{1}\left(\gamma_0 + \frac{v_1}{a_n} < q_i \le \gamma_0 + \frac{v_2}{a_n}\right)\right] \Big/ \varepsilon,
\end{aligned}$$

where the first and third inequalities are obvious, and the second inequality is from Markov's inequality. Since $\sup_{\gamma_0<\gamma\leq\gamma_0+\epsilon} E[|\bar{z}_{2i}||q_i=\gamma]$ is finite from Assumption 3.5(d), we need only prove that $E[n\sup_{|v_2-v_1|<\delta}\mathbf{1}(\gamma_0+\frac{v_1}{a_n}<q_i\leq\gamma_0+\frac{v_2}{a_n})]$ can be made arbitrarily small by choosing n large enough and δ small enough.

$$E\left[n\sup_{|v_2-v_1|<\delta}\mathbf{1}\left(\gamma_0+\tfrac{v_1}{a_n}<q_i\leq\gamma_0+\tfrac{v_2}{a_n}\right)\right]$$

$$\leq \sup_{|v_2-v_1|<\delta}\left|n\left(F\left(\gamma_0+\tfrac{v_2}{a_n}\right)-F\left(\gamma_0+\tfrac{v_1}{a_n}\right)\right)-(\Lambda_2(v_2)-\Lambda_2(v_1))\right|+\sup_{|v_2-v_1|<\delta}|(\Lambda_2(v_2)-\Lambda_2(v_1))|$$

$$\leq \sup_{|v_2-v_1|<\delta}\left|n\left(F\left(\gamma_0+\tfrac{v_2}{a_n}\right)-F(\gamma_0)\right)-\Lambda_2(v_2)\right|+\sup_{|v_2-v_1|<\delta}\left|n\left(F\left(\gamma_0+\tfrac{v_1}{a_n}\right)-F(\gamma_0)\right)-\Lambda_2(v_1)\right|$$
$$+\sup_{|v_2-v_1|<\delta}|\Lambda_2(v_2)-\Lambda_2(v_1)|,$$

and all three terms in the last inequality can be made arbitrarily small when Assumption 3.3 holds.

Proof of Theorem 3.2: The consistency is proved in Lemma B.2, and the convergence rate is proved in Lemma B.3. As to the asymptotic distribution, we follow the proof idea of Theorem 2 in Yu (2012). Basically, a modified version of the argmax continuous mapping theorem (Theorem 3.2.2 in Van der Vaart and Wellner (1996)) is used. The proof of Case (a) is exactly the same as that in Yu (2012). For Case (b), we need to prove $P(a_n(\widehat{\gamma}-\gamma_0)>0)\to 0$, where $\widehat{\gamma}$ can be $\widehat{\gamma}_{\mathrm{LLSE}}$ or $\widehat{\gamma}_{\mathrm{MLSE}}$. If we could prove $P(D_n(v)>0)\to 1$ for any $v>0$, then the proof is complete since $D_n(0)=0$. For any $v>0$, $D_n(v)=\sum_{i=1}^{N_{2n}(v)}\bar{z}_{2i}$. Since $E[N_{2n}(v)]\to\infty$ for any $v>0$ by the definition of a_n and b_n, and $E[\bar{z}_{2i}|q_i=\gamma]>0$ for γ in the right neighbourhood of γ_0 by Assumption 3.5(d), the strong law of large numbers implies $D_n(v)\to\infty$ for any $v>0$ almost surely. Case (c) can be similarly proved. □

Proof of Theorem 3.3: We only prove case (a) in detail, and the other two cases can be proved by combining the proof of Theorem 3.2 and the proof below.

First, we prove $P(c_n(\widehat{\gamma}_{\mathrm{LLSE}}-\gamma_0)\in(-1,1))\to 0$, which is implied by $P(D_n(v)=0, v\in(-1,1))\to 1$. From Assumption 3.4 and Theorem 2.1, $E[N_{1n}(v)]\to 0$ for any $v\in(-1,1)$, so there is no jump in $D_n(v)$ for v on $(-1,1)$ almost surely. Second, a similar analysis as in Theorem 3.2 shows that $P(c_n(\widehat{\gamma}_{\mathrm{LLSE}}-\gamma_0)<-1)\to 0$, and $P(c_n(\widehat{\gamma}_{\mathrm{LLSE}}-\gamma_0)>1)\to 0$. In summary, the asymptotic distribution of $c_n(\widehat{\gamma}_{\mathrm{LLSE}}-\gamma_0)$ must concentrate on two points: -1 and 1. From the definition of weak convergence and the analysis above, we need only prove that $P(c_n(\widehat{\gamma}_{\mathrm{LLSE}}-\gamma_0)\leq v)\to P(D(1)\geq\min\{D(-1),0\})$ for any $v\in(-1,1)$. Note that for any $v\in(-1,1)$,

$$P(c_n(\widehat{\gamma}_{\mathrm{LLSE}}-\gamma_0)\leq v)-P(D_n(1)\geq\min\{D_n(-1),0\})\to 0,$$

and

$$\begin{aligned}&P(D_n(1)\geq\min\{D_n(-1),0\})\\&=P(D_n(-1)<0 \text{ and } D_n(-1)\leq D_n(1)<0)+P(D_n(-1)<0<D_n(1))\\&\quad+P(D_n(-1)<0=D_n(1))+P(D_n(-1)=0=D_n(1))+P(D_n(-1)=0<D_n(1))\\&\quad+P(D_n(-1)>0 \text{ and } D_n(1)>0)+P(D_n(-1)>0=D_n(1)).\end{aligned}\tag{A.4}$$

From Theorem 3.1, $(D_n(-1), D_n(1))$ converge weakly to $(D(-1), D(1))$ with $D(-1)$ and $D(1)$ independent of each other. Because $D_n(-1)$ and $D_n(1)$ are random variables with a point mass only at zero, $P(D_n(\pm1)=0)\to P(D(\pm1)=0)$. As a result, each component in (A.4) converges to the counterpart with $D_n(\pm1)$ substituted by $D(\pm1)$, and the result follows.

The proof for the asymptotic distribution of $\widehat{\gamma}_{\mathrm{MLSE}}$ is similar. Note first that the asymptotic distribution of $c_n(\widehat{\gamma}_{\mathrm{MLSE}}-\gamma_0)$ must concentrate on three points: -1, 0 and 1. So we need only prove that

$$P(c_n(\widehat{\gamma}_{\mathrm{MLSE}}-\gamma_0)\leq v)\to P(D(1)>D(-1), D(-1)<0) \text{ for } v\in(-1,0),$$

$$P(c_n(\widehat{\gamma}_{\mathrm{MLSE}}-\gamma_0)\leq v)\to P(D(1)\geq\min\{D(-1),0\}) \text{ for } v\in(0,1).$$

Since

$$P\left(c_n\left(\widehat{\gamma}_{\mathrm{MLSE}}-\gamma_0\right)\le v\right)-P\left(D_n\left(-1\right)<0 \text{ and } D_n\left(-1\right)\le D_n(1)\right)\to 0 \text{ for } v\in(-1,0),$$

$$P\left(c_n\left(\widehat{\gamma}_{\mathrm{MLSE}}-\gamma_0\right)\le v\right)-P\left(D_n(1)\ge \min\left\{D_n\left(-1\right),0\right\}\right)\ \to 0 \text{ for } v\in(0,1),$$

the result follows from a similar analysis as in $\widehat{\gamma}_{\mathrm{LLSE}}$.

The independence between the asymptotic distributions of $\widehat{\beta}$ and $\widehat{\gamma}$ can be proved in a similar way as in Theorem 3.1. For instance, we can show that W_{1n}, W_{2n}, $D_n(1)$ and $D_n\left(-1\right)$ are asymptotically independent. □

Proof of Theorem 3.4: The asymptotic distribution of $\sqrt{n}\left(\widehat{\beta}_\ell-\beta_{\ell 0}\right)$ can be easily derived by applying the argmax continuous mapping theorem. As to the asymptotic distributions of the SATE and SATT, observe that

$$\begin{aligned}&\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}x_i'\left(\widehat{\beta}_2-\widehat{\beta}_1\right)-E\left[x\right]'\left(\beta_{20}-\beta_{10}\right)\right)\\&=\left(\frac{1}{n}\sum_{i=1}^{n}x_i'\right)\sqrt{n}\left(\widehat{\beta}_2-\widehat{\beta}_1-\left(\beta_{20}-\beta_{10}\right)\right)+\left(\beta_{20}-\beta_{10}\right)'\tfrac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(x_i-E[x]\right),\end{aligned}\tag{A.5}$$

and

$$\begin{aligned}&\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}x_i'\left(\widehat{\beta}_2-\widehat{\beta}_1\right)\mathbf{1}\left(q_i>\widehat{\gamma}\right)-E\left[x\mathbf{1}\left(q>\gamma_0\right)\right]'\left(\beta_{20}-\beta_{10}\right)\right)\\&=\left(\frac{1}{n}\sum_{i=1}^{n}x_i'\mathbf{1}\left(q_i>\widehat{\gamma}\right)\right)\sqrt{n}\left(\widehat{\beta}_2-\widehat{\beta}_1-\left(\beta_{20}-\beta_{10}\right)\right)+\left(\beta_{20}-\beta_{10}\right)'\\&\quad\times\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(x_i\mathbf{1}\left(q_i>\gamma_0\right)-E\left[x\mathbf{1}\left(q>\gamma_0\right)\right]\right)\\&\quad+\left(\beta_{20}-\beta_{10}\right)'\frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_i\mathbf{1}\left(\widehat{\gamma}<q_i\le\gamma_0\right).\end{aligned}$$

So we need only find the asymptotic distributions of

$$\begin{pmatrix}\sqrt{n}\left(\widehat{\beta}_1-\beta_{10}\right)\\ \sqrt{n}\left(\widehat{\beta}_2-\beta_{20}\right)\\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(x_i-E[x]\right)\end{pmatrix}\text{ and }\begin{pmatrix}\sqrt{n}\left(\widehat{\beta}_1-\beta_{10}\right)\\ \sqrt{n}\left(\widehat{\beta}_2-\beta_{20}\right)\\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(x_i\mathbf{1}\left(q_i>\gamma_0\right)-E\left[x\mathbf{1}\left(q>\gamma_0\right)\right]\right)\end{pmatrix},$$

and show that $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_i\mathbf{1}\left(\widehat{\gamma}<q_i\le\gamma_0\right)=o_p(1)$.

The influence function of

$$\begin{pmatrix}\sqrt{n}\left(\widehat{\beta}_1-\beta_{10}\right)\\ \sqrt{n}\left(\widehat{\beta}_2-\beta_{20}\right)\\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(x_i-E[x]\right)\end{pmatrix}$$

is

$$\begin{pmatrix}E\left[xx'\mathbf{1}\left(q\le\gamma_0\right)\right]^{-1}\sigma_{10}xe\mathbf{1}\left(q\le\gamma_0\right)\\ E\left[xx'\mathbf{1}\left(q>\gamma_0\right)\right]^{-1}\sigma_{20}xe\mathbf{1}\left(q>\gamma_0\right)\\ x-E[x]\end{pmatrix},$$

so the three components are asymptotically independent. By Slutsky's theorem and the continuous mapping theorem, it is easy to see that the asymptotic distribution of the SATE is as stated in the theorem. Similarly, the three components of the SATT are asymptotically independent. We now show that $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i \mathbf{1}(\widehat{\gamma} < q_i \leq \gamma_0) = o_p(1)$. For any $\epsilon > 0$, we can choose C large enough such that $P(|\widehat{\gamma}| \leq \frac{C}{c_n}) < \epsilon$, so we need only consider $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i \mathbf{1}(\frac{C}{c_n} < q_i \leq \gamma_0)$. The second moment of $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i \mathbf{1}(\frac{C}{c_n} < q_i \leq \gamma_0)$ is less than $E[\|x\|^2 \mathbf{1}(\frac{C}{c_n} < q_i \gamma_0)]$ which is $O(\frac{1}{n})$ by the definition of c_n. So the asymptotic distribution of the SATT is as stated in the theorem. □

APPENDIX B: LEMMAS

LEMMA B.1. *Suppose Assumptions 3.1–3.3 and 3.5 hold,* $\frac{a_n}{b_n} \to 1$, *then*

$$n P_n \left(m\left(\cdot \middle| \beta_0 + \tfrac{u}{\sqrt{n}}, \gamma_0 + \tfrac{v}{c_n}\right) - m\left(\cdot \middle| \beta_0, \gamma_0\right)\right)$$
$$= u_1' E\left[x_i x_i' \mathbf{1}(q_i \leq \gamma_0)\right] u_1 + u_2' E\left[x_i x_i' \mathbf{1}\left(q_i > \gamma_0\right)\right] u_2 - 2\sigma_{10} u_1' W_{1n} - 2\sigma_{20} u_2' W_{2n} + D_n(v) + o_p(1)$$

where the $o_p(1)$ *is uniform for* h *on any compact set in* $\mathbb{R}^{2k+1}$.

Proof. Note that

$$\begin{aligned}
&n P_n \left(m\left(\cdot \middle| \beta_0 + \frac{u}{\sqrt{n}}, \gamma_0 + \frac{v}{c_n}\right) - m\left(\cdot \middle| \beta_0, \gamma_0\right)\right) \\
&= \sum_{i=1}^{n} \left(u_1' \frac{x_i x_i'}{n} u_1 - u_1' \frac{2\sigma_{10}}{\sqrt{n}} x_i e_i \right) \mathbf{1}\left(q_i \leq \gamma_0 \wedge \gamma_0 + \frac{v}{c_n}\right) + \sum_{i=1}^{n} \left(u_2' \frac{x_i x_i'}{n} u_2 - u_2' \frac{\sigma_{20}}{\sqrt{n}} x_i e_i \right) \\
&\quad \times \mathbf{1}\left(q_i > \gamma_0 \vee \gamma_0 + \frac{v}{c_n}\right) + \sum_{i=1}^{n} \left[\left(\beta_{10} - \beta_{20} - \frac{u_2}{\sqrt{n}}\right)' x_i x_i' \left(\beta_{10} - \beta_{20} - \frac{u_2}{\sqrt{n}}\right) \right. \\
&\quad \left. + 2 x_i' \left(\beta_{10} - \beta_{20} - \frac{u_2}{\sqrt{n}}\right) \sigma_{10} e_i \right] \times \mathbf{1}\left(\gamma_0 + \frac{v}{c_n} < q_i \leq \gamma_0\right) + \sum_{i=1}^{n} \left[\left(\beta_{10} + \frac{u_1}{\sqrt{n}} - \beta_{20}\right)' \right. \\
&\quad \left. x_i x_i' \left(\beta_{10} + \frac{u_1}{\sqrt{n}} - \beta_{20}\right) - 2 x_i' \left(\beta_{10} + \frac{u_1}{\sqrt{n}} - \beta_{20}\right) \sigma_{20} e_i \right] \mathbf{1}\left(\gamma_0 < q_i \leq \gamma_0 + \frac{v}{c_n}\right),
\end{aligned}$$

so $o_p(1)$ includes three parts. The first part is $u_1'(\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \mathbf{1}(q_i \leq \gamma_0 \wedge \gamma_0 + \frac{v}{c_n}) - E[x_i x_i' \mathbf{1}(q_i \leq \gamma_0)])u_1$ and $u_2'(\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \mathbf{1}(q_i > \gamma_0 \vee \gamma_0 + \frac{v}{c_n}) - E[x_i x_i' \mathbf{1}(q_i > \gamma_0)])u_2$, which is $o_p(1)$ from Assumption 3.5(b) and a dominance argument. The second part is $2\sigma_{10} u_1' W_{1n} - u_1'(\frac{2\sigma_{10}}{\sqrt{n}}\sum_{i=1}^{n} x_i e_i \mathbf{1}(q_i \leq \gamma_0 \wedge \gamma_0 + \frac{v}{c_n})) + 2\sigma_{20} u_2' W_{2n} - u_2' \frac{\sigma_{20}}{\sqrt{n}}\sum_{i=1}^{n} x_i e_i \mathbf{1}(q_i > \gamma_0 \vee \gamma_0 + \frac{v}{c_n})$. This part is $o_p(1)$ since

$$\begin{aligned}
&Var\left(u_1' W_{1n} - u_1' \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i e_i \mathbf{1}\left(q_i \leq \gamma_0 \wedge \gamma_0 + \frac{v}{c_n}\right)\right)\right) \\
&= Var\left(u_1' \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i e_i \mathbf{1}\left(\gamma_0 \wedge \gamma_0 + \frac{v}{c_n} < q_i \leq \gamma_0\right)\right)\right) \\
&= u_1' E\left[x x' e^2 \mathbf{1}\left(\gamma_0 \wedge \gamma_0 + \frac{v}{c_n} < q \leq \gamma_0\right)\right] u_1
\end{aligned}$$

$$\leq u_1' \sup_{\gamma_0-\epsilon<\gamma\leq\gamma_0} E\left[xx'e^2|q=\gamma\right]\left(F(\gamma_0)-F\left(\gamma_0\wedge\gamma_0+\frac{v}{c_n}\right)\right)u_1$$

$$\to 0$$

as $n\to\infty$ by the definition of a_n and b_n and Assumption 3.5(d), and

$$Var\left(u_2'W_{2n}-u_2'\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_ie_i\mathbf{1}\left(q_i>\gamma_0\vee\gamma_0+\frac{v}{c_n}\right)\right)\right)\to 0$$

as $n\to\infty$ by a similar argument. The third part is

$$u_2'\sum_{i=1}^{n}\left(\frac{-2}{\sqrt{n}}x_ix_i'(\beta_{10}-\beta_{20})+\frac{x_ix_i'}{n}u_2-\frac{2\sigma_{10}}{\sqrt{n}}x_ie_i\right)\mathbf{1}\left(\gamma_0+\frac{v}{c_n}<q_i\leq\gamma_0\right)$$

$$+u_1'\sum_{i=1}^{n}\left(\frac{2}{\sqrt{n}}x_ix_i'(\beta_{10}-\beta_{20})+\frac{x_ix_i'}{n}u_1-\frac{2\sigma_{20}}{\sqrt{n}}x_ie_i\right)\mathbf{1}\left(\gamma_0<q_i\leq\gamma_0+\frac{v}{c_n}\right),$$

which is $o_p(1)$ by a similar argument as in the second part from Assumption 3.5(d). □

LEMMA B.2. *Under Assumptions 3.1–3.5,* $\widehat{\gamma}_{\mathrm{LLSE}}\longrightarrow^p\gamma_0$, $\widehat{\gamma}_{\mathrm{MLSE}}\longrightarrow^p\gamma_0$, *and* $\widehat{\beta}\longrightarrow^p\beta_0$.

Proof. The proof idea follows from Lemma B.1 of Hansen (2000).

First suppose $\gamma>\gamma_0$. Note that $Y=X\beta_{20}+\sigma_{20}\mathbf{e}+X_{\leq\gamma_0}\delta_{\beta0}+\delta_{\sigma0}\mathbf{e}_{\leq\gamma_0}$, and X lies in the space spanned by $P_\gamma=\overline{X}_\gamma(\overline{X}_\gamma'\overline{X}_\gamma)^{-1}\overline{X}_\gamma'$, where $\mathbf{e}_{\leq\gamma_0}$ is similarly defined as $X_{\leq\gamma_0}$, $\mathbf{e}=(e_1,\ldots,e_n)'$, $\delta_\beta=\beta_1-\beta_2$, $\delta_\sigma=\sigma_1-\sigma_2$, and $\overline{X}_\gamma=[X\ X_{\leq\gamma}]$. So

$$\begin{aligned}\frac{1}{n}Q_n(\gamma)&=\frac{1}{n}Y'\left(I-P_\gamma\right)Y\\&=\frac{1}{n}\left[2\sigma_{20}\delta_{\beta0}'X_{\leq\gamma_0}'\left(I-P_\gamma\right)\mathbf{e}+2\delta_{\sigma0}\delta_{\beta0}'X_{\leq\gamma_0}'\left(I-P_\gamma\right)\mathbf{e}_{\leq\gamma_0}+2\sigma_{20}\delta_{\sigma0}\mathbf{e}\left(I-P_\gamma\right)\mathbf{e}_{\leq\gamma_0}\right.\\&\quad\left.+\delta_{\beta0}'X_{\leq\gamma_0}'\left(I-P_\gamma\right)X_{\leq\gamma_0}\delta_{\beta0}+\sigma_{20}^2\mathbf{e}'\left(I-P_\gamma\right)\mathbf{e}+\delta_{\sigma0}^2\mathbf{e}_{\leq\gamma_0}'\left(I-P_\gamma\right)\mathbf{e}_{\leq\gamma_0}\right]\\&\xrightarrow{p}\delta_{\beta0}'\left(\underline{M}(\gamma_0)-\underline{M}(\gamma_0)\underline{M}(\gamma)^{-1}\underline{M}(\gamma_0)\right)\delta_{\beta0}+E\left[(\sigma_{20}e+\delta_{\sigma0}e\mathbf{1}\left(q\leq\gamma_0\right))^2\right]\equiv Q(\gamma)\end{aligned}$$

uniformly for $\gamma\in(\gamma_0,\overline{\gamma}]$ by a Glivenko–Cantelli theorem, where $\underline{M}(\gamma)=E[xx'\mathbf{1}(q\leq\gamma)]$. The second term of this limit does not depend on γ, so we concentrate on the first term. We need only prove that $\underline{M}(\gamma_0)-\underline{M}(\gamma_0)\underline{M}(\gamma)^{-1}\underline{M}(\gamma_0)>\underline{M}(\gamma_0)-\underline{M}(\gamma_0)\underline{M}(\gamma_0)^{-1}\underline{M}(\gamma_0)=0$ for any $\gamma>\gamma_0$, which reduces to $\underline{M}(\gamma)>\underline{M}(\gamma_0)$. Since $\underline{M}(\gamma_2)\geq\underline{M}(\gamma_1)$ if $\gamma_2>\gamma_1$, we need only prove that for any $\epsilon>0$, $\underline{M}(\gamma_0+\epsilon)-\underline{M}(\gamma_0)>0$. From Assumptions 3.1 and 3.5(d), $\underline{M}(\gamma_0+\epsilon)-\underline{M}(\gamma_0)=E\left[xx'\mathbf{1}(\gamma_0<q\leq\gamma_0+\epsilon)\right]\geq\left(\inf_{\gamma_0<\gamma\leq\gamma_0+\epsilon}E\left[xx'|q=\gamma\right]\right)(F(\gamma_0+\epsilon)-F(\gamma_0))>0$.

Second, suppose $\gamma\leq\gamma_0$. In this case,

$$\begin{aligned}\frac{1}{n}Q_n(\gamma)&\xrightarrow{p}\delta_{\beta0}'(M(\gamma,\gamma_0)-M(\gamma,\gamma_0)\overline{M}(\gamma)^{-1}M(\gamma,\gamma_0))\delta_{\beta0}\\&\quad+E[(\sigma_{20}e+\delta_{\sigma0}e\mathbf{1}(q\leq\gamma_0))^2]\equiv Q(\gamma)\end{aligned}$$

uniformly for $\gamma\in[\underline{\gamma},\gamma_0]$ by a Glivenko–Cantelli theorem, where $M(\gamma,\gamma_0)=E[xx'\mathbf{1}(\gamma<q\leq\gamma_0)]$ and $\overline{M}(\gamma)=E[xx'\mathbf{1}(q>\gamma)]$. As in the case $\gamma>\gamma_0$, we need only prove that $M(\gamma,\gamma_0)-M(\gamma,\gamma_0)\overline{M}(\gamma)^{-1}M(\gamma,\gamma_0)>0$ for $\gamma<\gamma_0$, which reduces to $\overline{M}(\gamma)>M(\gamma,\gamma_0)$ since $M(\gamma,\gamma_0)>0$. By the analysis in the case $\gamma>\gamma_0$, $\overline{M}(\gamma)-M(\gamma,\gamma_0)>E\left[xx'\mathbf{1}(\gamma_0<q\leq\gamma_0+\epsilon)\right]>0$.

If there is no point mass in $F(\cdot)$, $Q(\gamma)$ is continuous, and Theorem 2.1 of Newey and McFadden (1994) can be applied to prove that there is a $\varepsilon>0$ such that $P(\gamma_0-\varepsilon<\widehat{\gamma}<\gamma_0+\varepsilon)\to 0$ as $n\to\infty$. When

there is a point mass at γ_0, $Q(\gamma)$ jumps from a positive number to $Q(\gamma_0) = 0$ when γ approaches γ_0 from the left. So Theorem 2.1 of Newey and McFadden (1994) cannot be used to prove the consistency of $\widehat{\gamma}$. But a modified version of Theorem 2.1 (e.g. Theorem 5.7 of Van der Vaart, 1998) can be used to show that there is a $\varepsilon > 0$ such that $P(\gamma_0 \leq \widehat{\gamma} < \gamma_0 + \varepsilon) \to 0$ as $n \to \infty$.

Next, we show the consistency of $\widehat{\beta}_1$ and $\widehat{\beta}_2$.

$$\begin{aligned}\widehat{\beta}_1 &= \left(X'_{\leq\widehat{\gamma}} X_{\leq\widehat{\gamma}}\right)^{-1} X'_{\leq\widehat{\gamma}} Y \\ &= \left(\frac{1}{n} X'_{\leq\widehat{\gamma}} X_{\leq\widehat{\gamma}}\right)^{-1} \left(\frac{1}{n} X'_{\leq\widehat{\gamma}} X_{\leq\gamma_0}\right) \beta_{10} + \left(\frac{1}{n} X'_{\leq\widehat{\gamma}} X_{\leq\widehat{\gamma}}\right)^{-1} \left(\frac{1}{n} X'_{\leq\widehat{\gamma}} X_{>\gamma_0}\right) \beta_{20} + \left(\frac{1}{n} X'_{\leq\widehat{\gamma}} X_{\leq\widehat{\gamma}}\right)^{-1} \left(\frac{1}{n} X'_{\leq\widehat{\gamma}} \mathbf{e}\right)\end{aligned}$$

First, with probability approaching 1,

$$\begin{aligned}&\left\| \frac{1}{n} X'_{\leq\widehat{\gamma}} X_{\leq\widehat{\gamma}} - E\left[xx\mathbf{1}(q \leq \gamma_0)\right] \right\| \\ &\leq \sup_{\gamma\in(\gamma_0-\epsilon,\gamma_0+\epsilon)} \left\| \frac{1}{n} X'_{\leq\gamma} X_{\leq\gamma} - E\left[xx\mathbf{1}(q \leq \gamma)\right] \right\| + \left\| E\left[xx\mathbf{1}(q \leq \gamma)\right]|_{\gamma=\widehat{\gamma}} - E\left[xx\mathbf{1}(q \leq \gamma_0)\right] \right\|.\end{aligned}$$

The first term converges to zero in probability by a Glivenko–Cantelli theorem. If there is no point mass at γ_0 in $F(\cdot)$, the second term converges to zero in probability since $E[xx\mathbf{1}(q \leq \gamma)]$ is continuous at γ_0. When there is a point mass at γ_0, $E[xx\mathbf{1}(q \leq \gamma)]$ is not left continuous at γ_0, but $P(\widehat{\gamma} < \gamma_0) \to 0$, so the second term still converges to zero in probability. Similarly, we could prove $\frac{1}{n} X'_{\leq\widehat{\gamma}} X_{>\gamma_0} \xrightarrow{p} 0$ and $\frac{1}{n} X'_{\leq\widehat{\gamma}} \mathbf{e} \xrightarrow{p} 0$, so $\widehat{\beta}_1$ is consistent. The consistency of $\widehat{\beta}_2$ can be similarly proved. □

LEMMA B.3. Under Assumptions 3.1–3.5, $c_n(\widehat{\gamma}_{\text{LLSE}} - \theta_0) = O_p(1)$, $c_n(\widehat{\theta}_{\text{MLSE}} - \theta_0) = O_p(1)$ and $\sqrt{n}(\widehat{\beta} - \beta_0) = O_p(1)$.

Proof. Corollary 3.2.6 of Van der Vaart and Wellner (1996) is used in this proof. By checking the proof of Theorem 3.2.5 in Van der Vaart and Wellner (1996), we need only check the two conditions in Corollary 3.2.6 for γ in the right neighbourhood of γ_0 when there is a point mass at γ_0.

First, $M(\theta) - M(\theta_0) \geq C d^2(\theta, \theta_0)$ with $d(\theta, \theta_0) = \|\beta - \beta_0\| + \sqrt{|F(\gamma) - F(\gamma_0)|}$ for θ in a neighbourhood of θ_0.[8] When there is a point mass at γ_0, γ is restricted in the right neighbourhood of γ_0 as mentioned above.

$$\begin{aligned}&M(\theta) - M(\theta_0) \\ &= E\left[T(w|\theta_1, \theta_{10})\mathbf{1}(q \leq \gamma \wedge \gamma_0)\right] + E\left[T(w|\theta_2, \theta_{20})\mathbf{1}(q > \gamma \vee \gamma_0)\right] \\ &\quad + E\left[\overline{z}_1(w|\theta_2, \theta_{10})\mathbf{1}(\gamma \wedge \gamma_0 < q \leq \gamma_0)\right] + E\left[\overline{z}_2(w|\theta_1, \theta_{20})\mathbf{1}(\gamma_0 < q \leq \gamma \vee \gamma_0)\right] \\ &= (\beta_{10} - \beta_1)' E\left[xx'\mathbf{1}(q \leq \gamma \wedge \gamma_0)\right](\beta_{10} - \beta_1) + (\beta_{20} - \beta_2)' E\left[xx'\mathbf{1}(q > \gamma \vee \gamma_0)\right](\beta_{20} - \beta_2) \\ &\quad + (\beta_{10} - \beta_2)' E\left[xx'\mathbf{1}(\gamma \wedge \gamma_0 < q \leq \gamma_0)\right](\beta_{10} - \beta_2) + (\beta_{20} - \beta_1)' E \\ &\quad \times \left[xx'\mathbf{1}(\gamma_0 < q \leq \gamma \vee \gamma_0)\right](\beta_{20} - \beta_1) \\ &\geq C\left(\|\beta_{10} - \beta_1\|^2 + \|\beta_{20} - \beta_2\|^2 + |F(\gamma) - F(\gamma_0)|\right),\end{aligned}$$

where the last inequality is from Assumptions 3.5(a), 3.5(c) and 3.5(d).

Second, $E[\sup_{d(\theta,\theta_0)<\delta} |\mathbb{G}_n(m(w|\theta) - m(w|\theta_0))|] \leq C\delta$. Since $\{T(w|\theta_1, \theta_{10}) : d(\theta, \theta_0) < \delta\}$ is a finite-dimensional vector space of functions, and $\{\mathbf{1}(q \leq \gamma \wedge \gamma_0) : d(\theta, \theta_0) < \delta\}$ is a VC subgraph class of functions by Lemma 2.4 of Pakes and Pollard (1989), $\{A(w|\theta) : d(\theta, \theta_0) < \delta\}$ is VC subgraph by Lemma

[8] When there is a point mass at γ_0, $\sqrt{|F(\gamma) - F(\gamma_0)|} > \sqrt{f(0)} > 0$ for any $\gamma < 0$, so this metric is similar to the discrete metric. Fortunately, we do not need to consider $\gamma < 0$ in this case.

2.14 (ii) of Pakes and Pollard (1989). Similarly, $\{B(w|\theta) : d(\theta, \theta_0) < \delta\}$, $\{C(w|\theta) : d(\theta, \theta_0) < \delta\}$, and $\{D(w|\theta) : d(\theta, \theta_0) < \delta\}$ are VC subgraph. From Theorem 2.14.2 of Van der Vaart and Wellner (1996),

$$E\left[\sup_{d(\theta,\theta_0)<\delta} |\mathbb{G}_n (m(w|\theta) - m(w|\theta_0))|\right] \leq C\sqrt{PF^2},$$

where F is the envelope of $\{m(w|\theta) - m(w|\theta_0) : d(\theta, \theta_0) < \delta\}$. But from the function form of $m(w|\theta) - m(w|\theta_0)$, $\sqrt{PF^2} \leq C\delta$ by Assumption 3.2.2. So $\phi(\delta) = \delta$ in Corollary 3.2.6 of Van der Vaart and Wellner (1996) and $\frac{\delta}{\delta^\alpha}$ is decreasing for all $1 < \alpha < 2$. Since $r_n^2\phi(\frac{1}{r_n}) = r_n$, $\sqrt{n}d(\widehat{\theta} - \theta_0) = O_p(1)$. By the definition of d, $\widehat{\beta}$ is $\sqrt{n}$-consistent, and $n(F(\widehat{\gamma}) - F(\gamma_0))$ is $O_p(1)$. Therefore, for any $\varepsilon > 0$, we can find M_ε such that $P(n|F(\widehat{\gamma}) - F(\gamma_0)| > M_\varepsilon) < \varepsilon$, which is equivalent to $P(n|F(\gamma_0 + \frac{c_n(\widehat{\gamma}-\gamma_0)}{c_n}) - F(\gamma_0)| > M_\varepsilon) < \varepsilon$. From the definition of a_n and b_n, there is M'_ε such that $P(c_n|\widehat{\gamma} - \gamma_0| > M'_\varepsilon) < \varepsilon$. □

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's web site:

Asymptotics in the Simple Example
Algorithms
Extra Simulation Results