Calibrating the Confidence Intervals in Threshold Regression*

Ping Yu†
University of Hong Kong

Peter C. B. Phillips‡
Yale University, University of Auckland
University of Southampton & Singapore Management University

Started: June 2019
First Version: June 2021
This Version: June 2023

Abstract

In this paper, we calibrate the confidence intervals (CIs) in Hansen (2000) to achieve better coverage and/or length. For the threshold point, when the threshold effect is strong, we suggest smaller critical values for the convex hull of LR-CI to obtain shorter confidence intervals with less overcoverage, and when the threshold effect is weak, we suggest a LR-CI based on the least favorable distribution to eliminate or mitigate the undercoverage problem. For slope parameters, when the threshold effect is strong, the usual t-CI works well, and when the threshold effect is weak, we rigorize the Bonferroni-type CI in Hansen (2000). Simulation studies and two empirical applications illustrate the usefulness of our new CIs in practice.

Keywords: threshold regression, confidence interval, convex hull, least favorable distribution
JEL-Classification: C22, C24

*We acknowledge support from the GRF of Hong Kong Government under Grant No. 106200228.

†HKU Business School, The University of Hong Kong, Pokfulam Road, Hong Kong; corresponding author email: pingyu@hku.hk.

‡Cowles Foundation for Research in Economics, Yale University, POBox 208281, New Haven, CT, USA; email: peter.phillips@yale.edu

1 Introduction

Threshold regression (TR), as a parsimonious model of nonlinear relationships between a response and some covariates, is very popular in current practice of econometrics; see Hansen (2011) for an excellent review of applications in time series, cross sections and panel data. The TR model usually assumes

$$\begin{aligned} y_i &= \mathbf{x}_i'\beta_1 1(q_i \le \gamma) + \mathbf{x}_i'\beta_2 1(q_i > \gamma) + \varepsilon_i \\ &= \mathbf{x}_i'\beta_2 + \mathbf{x}_i'\delta_n 1(q_i \le \gamma) + \varepsilon_i, \end{aligned} \tag{1}$$

where y_i is the dependent variable or the response, q_i is the threshold variable which is used to split the sample, $\mathbf{x}_i = (1, x_i', q_i)' \in \mathbb{R}^k$ is the set of covariates and may include q_i as a component, ε_i is the error term and satisfies $E[\varepsilon_i|\mathcal{F}_{i-1}] = 0$ with $\mathcal{F}_{i-1}$ being the sigma field generated by $\{x_{i-j}, q_{i-j}, \varepsilon_{i-1-j}|j \ge 0\}$, and the parameter of interest is $\theta = (\gamma, \beta')'$ with $\beta = (\beta_1', \beta_2')'$, or equivalently, $\theta = (\gamma, \beta_2', \delta_n')'$ with $\delta_n = \beta_1 - \beta_2$ being the threshold effect in conditional mean of y_i. Note here that we use subscript n in δ_n to emphasize the dependence of $\beta_1 - \beta_2$ on n. This model is similar to the linear regression except that the regression coefficients depend on whether the threshold variable q crosses the threshold point γ.

Because $E[\varepsilon_i|\mathcal{F}_{i-1}] = 0$, we can estimate θ based on least squares. Specifically, θ is estimated by minimizing the following objective function,

$$S_n(\theta) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\beta_1 1(q_i \le \gamma) + \mathbf{x}_i'\beta_2 1(q_i > \gamma))^2.$$

Denote the least squares estimator (LSE) of θ as $\widehat{\theta} = \left(\widehat{\gamma}, \widehat{\beta}_1', \widehat{\beta}_2'\right)'$. Often, a two-step procedure is used to obtain $\widehat{\theta}$. First, given γ, run least squares on data with $q_i \le \gamma$ and $q_i > \gamma$ separately to obtain $\widehat{\beta}_1(\gamma)$ and $\widehat{\beta}_2(\gamma)$. Second, minimize the concentrated objective function

$$S_n(\gamma) = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i'\widehat{\beta}_1(\gamma) 1(q_i \le \gamma) + \mathbf{x}_i'\widehat{\beta}_2(\gamma) 1(q_i > \gamma)\right)^2$$

to obtain $\widehat{\gamma}$ and set $\widehat{\beta}_\ell = \widehat{\beta}_\ell(\widehat{\gamma})$, $\ell = 1, 2$. The threshold effect δ_n is estimated by $\widehat{\delta} = \widehat{\beta}_1 - \widehat{\beta}_2$. This paper concerns about the confidence intervals (CIs) for γ and β with good performances in coverage.

Currently, the dominant CI for γ in the literature is the LR-CI of Hansen (2000). Such a CI relies on the shrinking-threshold-effect asymptotics borrowed from the structural change literature such as Picard (1985) and Bai (1997). It is well known that the aim of any asymptotic argument is to provide useful small sample approximations. However, as shown in our simulations in the following Section 4 , the LR-CI undercovers when the threshold effect is small (i.e., the CI is too short) and overcovers when the threshold effect is large (i.e., the CI is too long), so it is desirable to conctruct a CI for γ which has approximate correct coverage irrespective of the magnitude of the threshold effect. This implies that we should use a longer CI when the threshold effect is small and a shorter CI when the threshold effect is large. One purpose of this paper is to provide such a LR-type CI. Usually, when the threshold effect is large, the LR-CI is quite narrow (i.e., quite informative) although it overcovers, so the undercoverage problem when the threshold effect is small seems more urgent.

Hansen's framework involves a shrinking-to-zero threshold effect but the effect is large enough ($n^{1/2}\|\delta_n\| \to \infty$) such that the usual tests for a threshold effect would reject the null of no threshold effect (with the p-value converging to zero); in other words, the threshold effect is large enough such that the threshold model under consideration is out of the contiguous neighborhood ($n^{1/2}\delta_n = b \in \mathbb{R}^k$) of the linear regression model. Such a "large" threshold effect makes the inference for the presence of a threshold trivial although the location of the threshold remains uncertain. Actually, as we will show in Section 5, the contiguous model provides a better approximation to the finite-sample distribution of the LR statistic when the threshold effect is "small". Such a local analysis is standard in the weak IV literature such as those started by Staiger and Stock (1997). In such an analysis, both the presence and the location of the threshold are uncertain.

There is some related literature on the CI construction for the break date in the structural change model. Table 1 in Elliot and Müller (2007, EM hereafter) shows that when the break size is small, the CI for the break date that inverts the t-statistic of Bai (1997) (labelled as t-CI hereafter) also suffers from the undercoverage problem; see also Bai and Perron (2006) and Eo and Morley (2015). EM propose the invariant-to-β CI that is invariant to the magnitude of β_2 and δ under the null and meanwhile maximizes the weighted average power over the break date and δ.[1] It turns out that EM's CI is based on the inversion of a sequence of tests for an additional break given a maintained break date. Their tests are constructed under the assumption that $n^{1/2}\delta_n = b$.[2] Although the EM-CI has accurate coverage even when the threshold effect is small, it is typically much longer than the t-CI (when the t-CI does not undercover or b is large) and LR-CI as observed in EM and Eo and Morley (2015) so is not as informative as the LR-CI in practice.[3]We will employ different CI-construction procedures for small and large b values to avoid this problem. Section 5.2 of Elliot, Müller and Watson (2015, EMW hereafter) improves the power of EM's test in the simple case with $\mathbf{x} = 1$ by neglecting the invariance to δ (but maintaining only the invariance to β_2). They show the limit experiment for testing the location of the break date is a Brownian bridge plus a drift which has a kink at the break date. By searching for the approximate least favorable distribution, they develop a nearly optimal (in the weighted average power) test which has a much higher power than EM's test so induces a shorter CI. However, there are two difficulties in applying EMW's test in the TR context. First, the limit experiment in TR is much more complicated. One may suggest to use the transformation in Lee and Wang (2022) to convert TR to structural change, but their transformation excludes q as a regressor (otherwise, sorting q_i will introduce a time trend in the regression). Second, when k is larger than a small number, say 2, EMW's procedure is computationally intractable.

Hansen (2000) suggests a Bonferroni-type CI for β that takes union of CIs for β when γ falls in a 80% CI. There are two problems associated with this CI. First, the coverage 80% is arbitrary, so may not perform universally well in all applications. Second, as shown in the simulations of Section 4, when δ_n is large, the Bonferroni-type CI tends to overcover, i.e., is too long. Taking $n^{1/2}\delta_n = b$, we suggest to use the Bonferroni-type CI with the coverage for γ appropriately chosen when b is small, and use the usual t-CI when b is large.

There is also some related literature on the CI construction for β in the structural change model. McCloskey (2017) proposes three Bonferroni-based size-corrected critical values which is uniformly valid and possesses desirable power properties by assuming $n^{1/2}\delta_n = b$ with $b \in \mathbb{R}^k_\infty$ and $\mathbb{R}_\infty = \mathbb{R} \cup \{-\infty, \infty\}$, while due to the complexity of the asymptotic distribution of $\widehat{\beta}$ (see Section 5), his size correction procedure is not feasible when k is moderately large. Elliot and Müller (2014, EW2 hereafter) consider the CI for each component of β_ℓ. By developing the limit experiment (which is a Brownian motion plus a kinked drift with the kink location at the break date) under the assumption that $n^{1/2}\delta_n = b$, they follow EMW and construct

[1] Under both H_0 and H_1, the test is invariant to β_2.

[2] Such a small-threshold-effect assumption (i.e., $\sqrt{n}\,\|\delta_n\|$ is small, which can be implied by either small n or small $\|\delta_n\|$) is relevant in practice. For example, in the application of Hansen (2000), the p-value for the threshold model using initial per capital output as q is 0.088 and for the threshold model using the initial literacy rate as q is 0.078, both of which are only marginally significant. In this example, the small $\sqrt{n}\,\|\delta_n\|$ may be due to the small n ($n = 96$ and 78 for these two threshold models).

[3] This is understandable by noticing that the weighting scheme in their test statistic puts null probability on large b values (note that the scale in the covariance matrix of the weighting Gaussian distribution is close to zero).

the CI by inverting an approximately weighted average power maximizing test.[4] However, the two difficulties mentioned above in applying EMW's procedure to constructing a CI for γ still apply here. To circumvent these difficulties, Andrews et al. (2021) consider conditional and unconditional inference on $\beta(\widehat{\gamma})$ by extending Andrews et al. (2019)'s procedure where the nuisance parameter (like γ here) is defined by maximizing the level of an asymptotically normal random variable to the case where the nuisance parameter is defined by maximizing the norm of an asymptotically normal random vector, where $\beta(\widehat{\gamma})$ is the pseudo-true value of β when the true γ is equal to $\widehat{\gamma}$. However, we are interested in β not $\beta(\widehat{\gamma})$. Also, their procedure restricts γ to stay in a finite set, which seems irrelevant in practice, but is problematic in theory. In our setup, γ stays in an expanding (with n) finite set which contains infinite possible values in the limit. Because there are sufficient discussions on the inference of β in the literature, we will focus on the inference of γ in this paper.

This paper is organized as follows. Section 2 reviews the CI construction in Hansen (2000). Section 3 shows what the genuine critical value should be when we use the convex hull of the set that is created by inverting the LR statistic for γ (i.e., a genuine interval). Section 4 motivates our new CIs using the same simulation studies in Hansen (2000). Section 5 develops the local asymptotics for the LSE of γ, β and the LR statistic when $n^{1/2}\delta_n = b \in \mathbb{R}^k$. Section 6 constructs our new CIs for γ and β. Section 6 includes some simulation results, Section 7 includes two empirical applications, and Section 8 concludes.

A word on notation: the symbol ℓ is used to indicate the two regimes in (1) and, to simplify notation in what follows, the explicit values "$\ell = 1, 2$" are often omitted. For $a, b \in \mathbb{R}$, $a \wedge b = \min(a, b)$, and $a \vee b = \max(a, b)$. For two random vectors x, y, $x \perp y$ means x is independent of y. $1_{(\cdot)}$ denotes the indicator function.

2 Review of Hansen (2000)

To review the asymptotic distribution of $\widehat{\theta}$ in the framework of Hansen (2000), we first replicate his Assumption 1 as Assumption D below. First, we define some notations. Let $f(q)$ be the density function of q, and γ_0 be the true value of γ,

$$\begin{aligned} M(\gamma) &= E\left[\mathbf{x}_i\mathbf{x}_i' 1(q_i \leq \gamma)\right], M = E\left[\mathbf{x}_i\mathbf{x}_i'\right], \\ D(\gamma) &= E\left[\mathbf{x}_i\mathbf{x}_i' | q_i = \gamma\right], D = D(\gamma_0), \\ V(\gamma) &= E\left[\mathbf{x}_i\mathbf{x}_i'\varepsilon_i^2 | q_i = \gamma\right], V = V(\gamma_0). \end{aligned}$$

Assumption D:

1. $(x_i, q_i, \varepsilon_i)$ is strictly stationary, ergodic and ρ-mixing, with ρ-mixing coefficients satisfying $\sum_{m=1}^{\infty} \rho_m^{1/2} < \infty$.
2. $E\left[\varepsilon_i | \mathcal{F}_{i-1}\right] = 0$.
3. $E\left[|\mathbf{x}_i|^4\right] < \infty$ and $E\left[|\mathbf{x}_i\varepsilon_i|^4\right] < \infty$.
4. For all $\gamma \in \Gamma$, $E\left[|\mathbf{x}_i|^4 |\varepsilon_i|^4 | q_i = \gamma\right] \leq C$ and $E\left[|\mathbf{x}_i|^4 | q_i = \gamma\right] \leq C$ for some $C < \infty$, and $f(\gamma) \leq \overline{f} < \infty$, where Γ is the parameter space of γ.

[4] When b is large, EW2 employs a switching scheme such that their procedure nearly reduces to the standard inference, and increase their critical values to account for this switch.Specifically, they suggest to use a fixed critical value for the sup$F$ test (such as 90) and then adjust the critical value from 1.96 to 2.01 if the null of no structural change is rejected to ensure overall size control in the t-test for β.

5. $f(\gamma)$, $D(\gamma)$, and $V(\gamma)$ are continuous at $\gamma = \gamma_0$.

6. $\delta_n = cn^{-\varphi}$, with $c \neq 0$ and $\varphi \in (0, 1/2)$.

7. $c'Dc > 0$, $c'Vc > 0$, and $f = f(\gamma_0) > 0$.

8. $M > M(\gamma) > 0$ for all $\gamma \in \Gamma$.

Hansen (2000) provides detailed discussions on these assumption after his Assumption 1, so we only briefly mention some key points. First, Assumption D6 assumes that δ_n shrinks to zero but stays out of the contiguous neighborhood of $\delta_n = 0$ (i.e., $\delta_n = cn^{-1/2}$) so that γ can still be point identifies. Assumption D7 excludes the continuous threshold model discussed in Chan and Tsay (1998) and Hansen (2017). Second, Assumption D8 restricts Γ to be a proper subset of the support of q_i. In practice, we often set $\Gamma = \left[\underline{\gamma}, \overline{\gamma}\right]$, where $\underline{\gamma}$ and $\overline{\gamma}$ are the lower and upper $\epsilon\%$ quantiles of $\{q_i\}_{i=1}^n$. This guarantees that each regime contains at least $\epsilon\%$ of the whole dataset for some $\epsilon > 0$ (typically, 5, 10 or 15). Note that $S_n(\gamma)$ is constant on $[q_{(i)}, q_{(i+1)})$, where $\left\{q_{(i)}\right\}_{i=1}^n$ is the sorted $\{q_i\}_{i=1}^n$. This is why we need only check $\gamma \in \Gamma_n$ to search for $\widehat{\gamma}$ in practice, where $\Gamma_n = \{q_i | q_i \in \Gamma\}$. In other words, $\widehat{\gamma}$ is taken as the left endpoint of the minimizing interval of $S_n(\gamma)$. Yu (2012, 2015) suggests to take the middle point of this interval as $\widehat{\gamma}$ to improve its finite-sample performance, but under Assumption D6, taking any point in this interval as $\widehat{\gamma}$ does not affect its asymptotic properties. Intuitively, this is because the convergence rate of $\widehat{\gamma}$ is slower than n, while the distance between $q_{(i)}$ and $q_{(i+1)}$ is $O\left(n^{-1}\right)$.

Under Assumption D, Theorem 1 of Hansen (2000) shows that

$$n\widehat{f}\frac{\left(\widehat{\delta}'\widehat{D}\widehat{\delta}\right)^2}{\widehat{\delta}'\widehat{V}\widehat{\delta}}\left(\widehat{\gamma} - \gamma_0\right) \xrightarrow{d} \arg\max_v \left[-\frac{|v|}{2} + B(v)\right], \tag{2}$$

where $\widehat{f}$,, $\widehat{D}$ and $\widehat{V}$ are consistent estimators of f, D, and V, respectively. The confidence interval of γ can be constructed by inverting the t statistic in testing $H_0 : \gamma = \gamma_0$ vs. $H_1 : \gamma \neq \gamma_0$. Specifically, a $(1-\alpha)$ CI is

$$\left[\widehat{\gamma} - \frac{c^t_{\alpha/2}}{n\widehat{f}}\frac{\widehat{\delta}'\widehat{V}\widehat{\delta}}{\left(\widehat{\delta}'\widehat{D}\widehat{\delta}\right)^2}, \widehat{\gamma} + \frac{c^t_{\alpha/2}}{n\widehat{f}}\frac{\widehat{\delta}'\widehat{V}\widehat{\delta}}{\left(\widehat{\delta}'\widehat{D}\widehat{\delta}\right)^2}\right],$$

where c^t_α is the upper αth quantile of the distribution of $\arg\max_v \left[-\frac{|v|}{2} + B(v)\right]$ which is developed in Bhattacharya and Brockwell (1976). In the homoskedastic case where $E\left[\varepsilon^2|\mathbf{x}\right] = E\left[\varepsilon^2\right] = \sigma^2$, $\frac{\widehat{\delta}'\widehat{V}\widehat{\delta}}{\left(\widehat{\delta}'\widehat{D}\widehat{\delta}\right)^2}$ can be replaced by $\frac{\widehat{\sigma}^2}{\widehat{\delta}'\widehat{D}\widehat{\delta}}$, where $\widehat{\sigma}^2$ is a consistent estimator of σ^2.

Because γ cannot be identified when $\delta_n = 0$, following Dufour (1997), Hansen (2000) suggests to use the LR-based CI to improve the performance,[5] where the LR statistic is

$$LR_n(\gamma) = n\frac{S_n(\gamma) - S_n(\widehat{\gamma})}{S_n(\widehat{\gamma})}.$$

It can be shown that

$$LR^*_n(\gamma_0) := \frac{LR_n(\gamma_0)}{\widehat{\eta}^2} \xrightarrow{d} \xi,$$

[5] Another advantage of LR-CI over t-CI is that the former is shorter than the latter both asymptotically and in finite samples as shown in Eo and Morley (2015) in the structural change context. Technically, this is because the t-test is less powerful than the LR test.

where $\widehat{\eta}^2 = \frac{\widehat{\delta}'\widehat{V}\widehat{\delta}}{\widehat{\sigma}^2\widehat{\delta}'\widehat{D}\widehat{\delta}}$ is a consistent estimator of

$$\eta^2 = \frac{\delta_n' V \delta_n}{\sigma^2 \delta_n' D \delta_n} = \frac{c' V c}{\sigma^2 c' D c} =: \frac{\lambda}{\sigma^2 \mu},$$

$\widehat{\sigma}^2$ is an estimator of $\sigma^2 = E\left[\varepsilon^2\right]$ such as $\widehat{S}_n\left(\widehat{\gamma}\right)/n$, and $\xi = \max_v \left[-|v| + 2B\left(v\right)\right]$. When $E\left[\varepsilon^2|\mathbf{x}\right] = \sigma^2$, $\widehat{\eta}^2$ can be replaced by 1. As a result, a $(1-\alpha)$ CI is

$$\widehat{\Gamma}\left(1-\alpha\right) = \left\{\gamma : LR_n^*\left(\gamma\right) \leq c_\alpha\right\},$$

where c_α is the upper αth quantile of the distribution of ξ which is $P\left(\xi \leq x\right) = \left(1 - e^{-x/2}\right)^2$. Compared with the t-CI, the LR-CI does not need to estimate f and D in the homoskedastic case.

Quite ofen, $\widehat{\Gamma}\left(1-\alpha\right)$ is a union of segments rather than an interval, so a common practice is to take the convex hull of $\widehat{\Gamma}\left(1-\alpha\right)$ as the CI for γ, denoted as $\text{conv}\left\{\widehat{\Gamma}\left(1-\alpha\right)\right\}$. Because $LR_n\left(\gamma\right)$ is flat on $[q_{(i)}, q_{(i+1)})$, the convex hull of $\widehat{\Gamma}\left(1-\alpha\right)$ takes the form of $[q_{(i)}, q_{(j)})$ for some $i < j$. For comparison, the t-CI is always an interval. The disjointness of LR-CI is also observed in the structural change literature, e.g., Siegmund (1986, 1988), where the convex CI is also considered.[6] Note that $P\left(\gamma_0 \in \text{conv}\left\{\widehat{\Gamma}\left(1-\alpha\right)\right\}\right) \geq P\left(\gamma_0 \in \left\{\widehat{\Gamma}\left(1-\alpha\right)\right\}\right) \to 1-\alpha$, so the critical value c_α is too large for $\text{conv}\left\{\widehat{\Gamma}\left(1-\alpha\right)\right\}$. We label $P\left(\gamma_0 \in \text{conv}\left\{\widehat{\Gamma}\left(1-\alpha\right)\right\}\right)$ as the interval coverage, and $P\left(\gamma_0 \in \widehat{\Gamma}\left(1-\alpha\right)\right)$ as the actual coverage.

Under Assumption D, $\widehat{\beta}_\ell$ is asymptotically normal; also, $\widehat{\beta}_1$, $\widehat{\beta}_2$ and $\widehat{\gamma}$ are asymptotically independent. Hansen (2000) suggests a Bonferroni-type CI for β and δ_n. Specifically, the $(1-\alpha)$ confidence set (CS) for β and δ_n takes the form of

$$\widehat{\Theta}_\rho = \bigcup_{\gamma \in \widehat{\Gamma}(\rho)} \widehat{\Theta}\left(1-\alpha|\gamma\right),$$

where $\widehat{\Theta}\left(1-\alpha|\gamma\right)$ is the $(1-\alpha)$ CI for a component of β or δ_n given that the threshold value were known as γ. More specifically, the $(1-\alpha)$ confidence set (CS) for β_ℓ is

$$\widehat{B}_{\ell,1-\alpha} = \bigcup_{\gamma \in \widehat{\Gamma}(1-\alpha_\gamma)} \widehat{B}_\ell\left(1-\alpha_\ell|\gamma\right)$$

with $\alpha_\ell = \alpha$ and the $(1-\alpha)$ CS for δ_n is

$$\widehat{B}_{1-\alpha} = \bigcup_{\gamma \in \widehat{\Gamma}(1-\alpha_\gamma)} \widehat{B}\left(1-\alpha_\delta|\gamma\right),$$

with $\alpha_\delta = \alpha$, where $\alpha_\gamma = 0.2$, $\widehat{B}_\ell\left(1-\alpha_\ell|\gamma\right)$ is the $(1-\alpha_\ell)$ CS for β_ℓ given γ, and $\widehat{B}\left(1-\alpha_\delta|\gamma\right)$ is the $(1-\alpha_\delta)$ CS for δ_n given γ. Of course, if we are interest in one component of β_ℓ or δ_n, we can replace the CS by a CI.

3 Genuine Critical Values in Hansen's Framework

From the discussion in the last section, we know the critical value c_α is too large if we use the $\text{conv}\left\{\widehat{\Gamma}\left(1-\alpha\right)\right\}$ instead of $\widehat{\Gamma}\left(1-\alpha\right)$ as the CI for γ. A natural question is what the genuine critical value should be if $\text{conv}\left\{\widehat{\Gamma}\left(1-\alpha\right)\right\}$ is employed. The following theorem answers this question.

[6] As commented on page 391 of Siegmund (1986), "In fact, because of the rapid fluctuations of Brownian sample paths, with probability 1 it consists of the union of infinitely many open intervals."

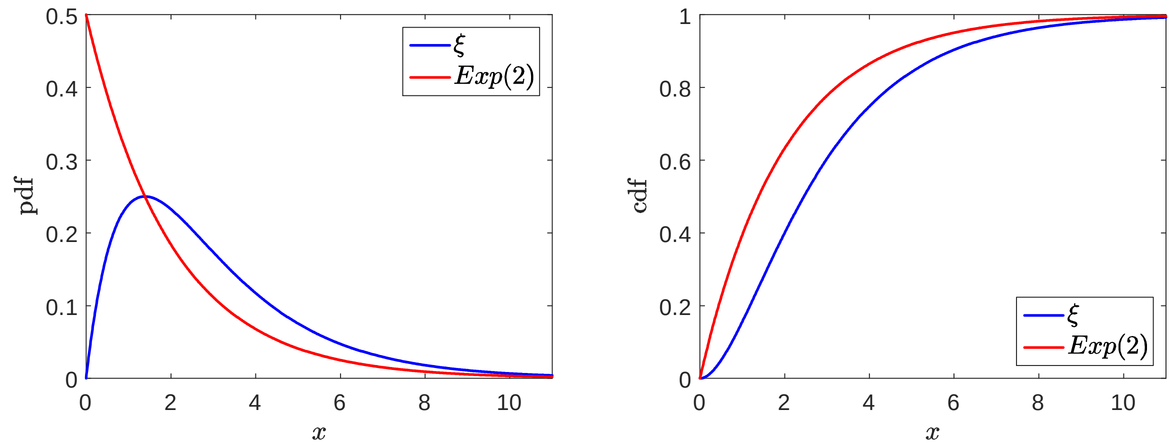

Figure 1: Comparison Between ξ and $Exp\,(2)$

Theorem 1 *Under Assumption D,*

$$P\left(\gamma_0 \in \text{conv}\left\{\gamma : LR_n^*\left(\gamma\right) \leq x\right\}\right) \to p\left(x\right) := P\left(\max\left[\xi_1, \xi_2\right] - \min\left[\xi_1, \xi_2\right] \leq \frac{x}{2}\right) = 1 - e^{-\frac{x}{2}},$$

where ξ_1 and ξ_2 are independent and both follow $Exp\,(1)$.

For comparison, in Hansen (2000),

$$\begin{aligned} P\left(LR_n^*\left(\gamma_0\right) \leq x\right) &\to P\left(\max\left[\xi_1, \xi_2\right] \leq \frac{x}{2}\right) = P\left(\xi \leq \frac{x}{2}\right) \\ &\leq P\left(\max\left[\xi_1, \xi_2\right] - \min\left[\xi_1, \xi_2\right] \leq \frac{x}{2}\right). \end{aligned}$$

Note that $1 - e^{-\frac{x}{2}}$ is the cdf of exponential distribution with mean 2, denoted as $Exp\,(2)$. Now, $p^{-1}\left(x\right) = -2\log\left(1 - x\right)$ while $F_\xi^{-1}\left(x\right) = -2\log\left(1 - \sqrt{x}\right)$. Figure 1 shows the difference between these two distributions. Because $Exp\,(2)$ has a much thinner right tail than ξ, its critical values are significantly smaller and $p\left(F_\xi^{-1}\left(1 - \alpha\right)\right) > 1 - \alpha$ as shown in Table 1.

$1-\alpha$	.80	.85	.90	.925	.95	.975	.99
$F_\xi^{-1}\left(1-\alpha\right)$	4.497	5.101	5.939	6.528	7.352	8.751	10.592
$p\left(c_\alpha\right)$	0.894	0.922	0.949	0.962	0.975	0.987	0.995
$p^{-1}\left(1-\alpha\right)$	3.219	3.794	4.605	5.181	5.991	7.378	9.210

Table 1: Calibrating Critical Values

One may wonder why we need not consider the difference between these two types of CIs for regular parameters. To explain the reason, consider the LR-CI for μ, where μ is the mean of a normal distribution with unit variance, i.e., we observe iid $X_i \sim N\left(\mu, 1\right)$, $i = 1, \cdots, n$. The LR statistic is

$$L_n\left(\mu\right) = 2\log\frac{\mathcal{L}_n\left(\widehat{\mu}\right)}{\mathcal{L}_n\left(\mu\right)} = n\left(\overline{X} - \mu\right)^2,$$

where $\widehat{\mu} = \overline{X}$ is the MLE of μ, and $\mathcal{L}_n\left(\mu\right)$ is the likelihood function for μ. As a result, $\left\{\mu : L_n\left(\mu\right) \leq c_\alpha\right\} = \left[\overline{X} - n^{-1/2}c_\alpha^{1/2}, \overline{X} + n^{-1/2}c_\alpha^{1/2}\right]$, exactly the same as the t-CI, where c_α is the upper αth quantile of $\chi^2\left(1\right)$.

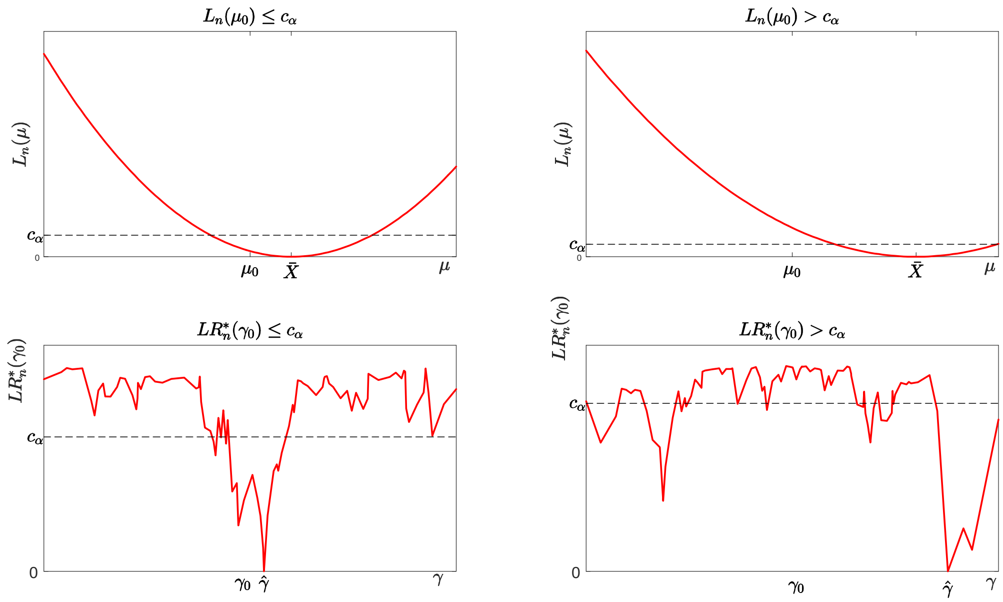

Figure 2: Comparion of the CI Construction for μ and γ

Why we need only check whether $L_n(\mu_0) \le c_\alpha$ to determine whether the CI covers μ_0? This is because $L_n(\mu_0) \le c_\alpha$ is equivalent to $\mu_0 \in \{\mu : L_n(\mu) \le c_\alpha\}$ when $\{\mu : L_n(\mu) \le c_\alpha\}$ is an interval. On the contrary, for the LR-CI of γ in TR, the event $\{LR_n^*(\gamma_0) \le c_\alpha\}$ is smaller than $\{\gamma_0 \in \text{conv}\{\gamma : LR_n^*(\gamma) \le c_\alpha\}\}$. The following Figure 2 illustrates this point. From the two upper graphs, we can see if $L_n(\mu_0) \le c_\alpha$, then the CI covers μ_0, and vice versa. On the contrary, from the two lower graphs, although $LR_n^*(\gamma_0) \le c_\alpha$ implies the CI covers γ_0, $LR_n^*(\gamma_0) > c_\alpha$ does not imply the interval CI excludes γ_0.

4 Motivation of New Confidence Intervals

To motivate our new confidence intervals for γ and β, we re-examine the simulation studies in Hansen (2000), where $\mathbf{x} = (1, z_i)'$ with $z_i = q_i$ or $z_i = x_i$ being either iid $N(0,1)$, $\varepsilon_i \sim N(0,1)$, $q_i \sim N(2,1)$, $\beta_2 = \mathbf{0}$, $\delta_n = (0, \delta_{20})'$, and $\gamma_0 = 2$. We label the case with $z_i = q_i$ as DGP1 and $z_i = x_i$ as DPG2. We set $\delta_{20} = 0, 0.25, 0.5, 1$ and $n = 50, 100, 500$ to check the sensitivity of coverage, and we use 10000 replications to improve the preciseness of simulation.

We first confirm the result in Dufour (1997) in our scenario, i.e., when δ_{20} is close to zero, the length of the t-CI tends to be infinite. To be specific, the normalization factor in (2) converges to

$$nf\frac{(\delta_0' D\delta_0)^2}{\delta_0' V\delta_0} = n\frac{\delta_0' D\delta_0}{\sqrt{2\pi}} = \frac{n\delta_{20}^2 D_{22}}{\sqrt{2\pi}},$$

where $D = E\left[\left(\begin{array}{c}1\\ q\end{array}\right)(1,q)\middle| q=2\right] = \left(\begin{array}{cc}1 & 2\\ 2 & 4\end{array}\right)$ or $D = E\left[\left(\begin{array}{c}1\\ x\end{array}\right)(1,x)\right] = \left(\begin{array}{cc}1 & 0\\ 0 & 1\end{array}\right) =: \mathbf{I}_2$, and $V = D$. So in terms of the normalization factor, $2\delta_{20}$ when $z = x$ is similar to δ_{20} when $z = q$. Obviously, this normalization factor tends to zero as δ_{20} shrinks to zero such that the t-CI diverges to the whole line.

		$z=q$				$z=x$			
$n \searrow$	$\delta_{20} \longrightarrow$	0	0.25	0.5	1	0	0.5	1	2
Coverage	50	0.753	0.783	0.851	0.940	0.783	0.834	0.903	0.957
		0.891	0.893	0.909	0.949	0.894	0.903	0.931	0.962
	100	0.700	0.787	0.887	0.956	0.745	0.848	0.929	0.970
		0.897	0.907	0.925	0.960	0.905	0.917	0.945	0.974
	500	0.571	0.881	0.938	0.967	0.615	0.919	0.946	0.975
		0.893	0.932	0.955	0.970	0.893	0.945	0.958	0.979
Length	50	2.154	2.022	1.597	0.530	2.184	1.872	1.097	0.350
	100	2.744	2.374	1.287	0.168	2.792	1.904	0.578	0.169
	500	3.782	1.277	0.098	0.024	3.844	0.394	0.091	0.033

Table 2: Coverage and Length of 90% LR-CI

Note: for coverage, the upper is actual coverage and the lower is interval coverage;
the length is the average length of conv$\{\gamma : LR_n^*(\gamma) \le c_{0.9}\}$

We then report the coverage and length of 90% LR-CI for γ in Table 2, where we report the coverage for both $\widehat{\Gamma}(1-\alpha)$ and conv$\left\{\widehat{\Gamma}(1-\alpha)\right\}$ with the latter larger. From Table 2, we draw the following conclusions. First, although $P\left(LR_n^*(\gamma_0) \le c_{0.9}\right) < 0.9$ when δ_{20} is small, $P\left(\gamma_0 \in \text{conv}\left\{\gamma : LR_n^*(\gamma) \le c_{0.9}\right\}\right)$ matches the nominal coverage pretty well even when δ_{20} is small. Especially, when δ_{20} is close to zero, $P\left(LR_n^*(\gamma_0) \le c_{0.9}\right)$ gets smaller when n gets larger while $P\left(\gamma_0 \in \text{conv}\left\{\gamma : LR_n^*(\gamma) \le x\right\}\right)$ is quite stable and close to 0.9. Second, When δ_{20} is large, both probabilities exceed 0.9. Third, when δ_{20} is small, the gap between these two probabilities is large, while when δ_{20} gets larger, the gap diminishes. Actually, only if δ_{20} is large, $\widehat{\Gamma}(1-\alpha)$ tends to be an interval such that the two coverages are close. Fourth, the CI becomes shorter when δ_{20} gets larger for each n, but need not become shorter as n gets larger when δ_{20} is small. Fifth, the parameter space Γ_n is the q_i's after deleting the five smallest and five largest q_i's, so the expected length of Γ_n is 2.56, 3.29 and 4.65 when $n = 50, 100$ and 500, respectively. When $z = q$ and $\delta_{20} = 0$, the length ratio of the CI to Γ_n is $2.154/2.56 = 0.841$, $2.744/3.29 = 0.834$ and $3.782/4.65 = 0.813$, when $n = 50, 100$ and 500, all less than 0.9, which implies that the distribution of $\widehat{\gamma}$ is not uniform although its distribution must be very disperse. Similar conclusions apply to the $z = x$ case. Sixth, when $\delta_{20} = 0$, the CI is the longest when $n = 500$. This is because such a data size provides a sharp signal that there is no threshold effect, while when n is small, the noise may provide a fake signal that the threshold effect exists so that the CI is smaller.

The good match of $P\left(\gamma_0 \in \text{conv}\left\{\gamma : LR_n^*(\gamma) \le c_{0.9}\right\}\right)$ with 0.9 is the result of misusage of the critical value $c_{0.9}$. Actually, we should compare the interval coverage with $p\left(c_{0.9}\right) = 0.949$, so there is also serious undercoverage just like the actual coverage. Only because we use a larger (than appropriate) critical value, the interval coverage seems to match the target coverage. Table 3 reports the coverage and length of $P\left(LR_n^*(\gamma_0) \le p^{-1}(0.9)\right)$. Because the critical value is smaller, the coverage is smaller and the length are shorter compared with Table 2. Using the modified critical value, we have undercoverage when δ_{20} is small while suffer less overcoverage when δ_{20} is large. A better strategy is to use a larger critical value to maintain the coverage when δ_{20} is small while use a smaller critical value like $p^{-1}(1-\alpha)$ to alleviate the overcoverage when δ_{20} is large. In the following sections, we will discuss how to determine whether δ_{20} is small or large and how to obtain a larger critical value when δ_{20} is small.

		$z=q$				$z=x$			
$n \searrow$	$\delta_{20} \longrightarrow$	0	0.25	0.5	1	0	0.5	1	2
Coverage	50	0.823	0.825	0.840	0.913	0.827	0.833	0.879	0.938
	100	0.828	0.833	0.868	0.930	0.829	0.851	0.905	0.949
	500	0.815	0.867	0.899	0.942	0.814	0.897	0.928	0.957
Length	50	1.917	1.749	1.301	0.409	1.968	1.626	0.881	0.304
	100	2.439	1.991	0.989	0.140	2.483	1.547	0.465	0.146
	500	3.278	0.966	0.074	0.020	3.324	0.314	0.075	0.029

Table 3: Coverage and Length of 90% LR-CI Using Modified Critical Value $p^{-1}(0.9)$

We next discuss the CI for δ_2. As mentioned in Section 2, Hansen (2000) suggests $\alpha_\gamma = 0.2$ and $\alpha_\ell = \alpha_\delta = 0.05$ when $\alpha = 0.05$.[7] We are only sure that the coverage of his CI is greater than 0.95 in his framework but not sure whether the coverage is greater than 0.95 when δ_{20} is smaller. Table 4 reports the coverage and length of 95% CIs for δ_2. From Table 4, we draw the following conclusions. First, as expected, $\widehat{B}_{0.8}$ always has a larger coverage and is longer than $\widehat{B}_0$. When $\sqrt{n}\delta_{20}$ is small, $\widehat{B}_{0.8}$ is much longer than $\widehat{B}_0$, while when $\sqrt{n}\delta_{20}$ is large, they get close in length.[8] Second, both $\widehat{B}_0$ and $\widehat{B}_{0.8}$ get shorter when either n or δ_{20} gets larger because the uncertainty in $\widehat{\gamma}$ and $\widehat{\delta}_2$ gets smaller. Third, the coverage of $\widehat{B}_0$ increases when either n or δ_{20} gets larger, while the coverage of $\widehat{B}_{0.8}$ first increases and then decreases with δ_{20}. Fourth, $\widehat{B}_0$ has undercoverage when $\sqrt{n}\delta_{20}$ is small while perfect coverage when $\sqrt{n}\delta_{20}$ is large. On the other hand, $\widehat{B}_{0.8}$ has perfect coverage when δ_{20} is small while has overcoverage when $\sqrt{n}\delta_{20}$ is large (so is long). A better CI can combine the advantages of both $\widehat{B}_0$ and $\widehat{B}_{0.8}$ – when $\sqrt{n}\,\|\delta_n\|$ is small, use the Bonferroni-type CS to guarantee coverage, and when $\sqrt{n}\,\|\delta_n\|$ is large, use the standard CS to keep the volume of CS small.

		$z=q$				$z=x$			
$n \searrow$	$\delta_{20} \longrightarrow$	0	0.25	0.5	1	0	0.5	1	2
Coverage	50	0.694	0.715	0.747	0.878	0.615	0.780	0.902	0.933
		0.950	0.953	0.955	0.949	0.925	0.952	0.968	0.949
	100	0.735	0.762	0.832	0.929	0.618	0.868	0.933	0.945
		0.961	0.972	0.970	0.963	0.946	0.976	0.971	0.954
	500	0.757	0.875	0.936	0.949	0.670	0.960	0.961	0.964
		0.973	0.979	0.973	0.963	0.965	0.982	0.966	0.965
Length	50	4.551	4.405	3.269	2.051	2.043	1.771	1.436	1.136
		16.24	14.41	10.238	3.844	5.625	4.302	2.186	1.251
	100	2.880	2.401	1.684	1.323	1.238	0.963	0.799	0.794
		8.349	7.017	4.096	1.571	2.793	1.841	0.975	0.838
	500	1.116	0.650	0.584	0.583	0.527	0.355	0.351	0.351
		2.712	1.265	0.689	0.621	1.205	0.424	0.373	0.359

Table 4: Coverage and Length of 95% CIs for δ_2

Note: the upper is $\widehat{B}_0$ and the lower is $\widehat{B}_{0.8}$

[7] Because he uses a larger critical value as argued in Section 3, his critical value at $\alpha_\gamma = 0.2$ is roughly equivalent to $p^{-1}(0.9)$ from Table 1.

[8] This echos the comment in EW2's Introduction, "A Bonferroni procedure based on uniformly valid confidence sets for the break date developed by EM performs well for large breaks, but has poor power for breaks of moderate magnitude".

5 Local Asymptotics

As shown in the simulations in the last section, we should pay special attention to the case where $\|\delta_n\|$ is small. Following the literature, we assume $\sqrt{n}\delta_n \to b \in \mathbb{R}^k$, and then study the asymptotic behavior of the LSE and the LR statistics.

The following theorem states the asymptotic distribution of $\widehat{\beta}_1$, $\widehat{\beta}_2$ and $\widehat{\gamma}$, where we use subscript n to emphasize the dependence of the true value of β_1 and β_2 on n. We first strengthen Assumption D.5 and D.8 as D.5$'$ and D.8$'$ below.

Assumption D.5$'$: Over $\gamma \in \Gamma$, $f(\gamma)$, $D(\gamma)$, and $V(\gamma)$ are continuous, and $f(\gamma) > \underline{f} > 0$, $D(\gamma) > 0$, and $V(\gamma) > 0$.
Assumption D.8$'$: $M > M(\gamma) > 0$ and $E\left[\mathbf{x}\mathbf{x}'\varepsilon^2\right] > E\left[\mathbf{x}\mathbf{x}'\varepsilon^2 1_{(q\le\gamma)}\right] > 0$ for all $\gamma \in \Gamma$.

Because $E\left[\mathbf{x}\mathbf{x}'1_{(\gamma_1<q\le\gamma_2)}\right] = \int_{\gamma_1}^{\gamma_2} D(\gamma) f(\gamma) d\gamma$, and $E\left[\mathbf{x}\mathbf{x}'\varepsilon^2 1_{(\gamma_1<q\le\gamma_2)}\right] = \int_{\gamma_1}^{\gamma_2} V(\gamma) f(\gamma) d\gamma$, Assumption D.5$'$ implies both terms are positive definite as long as $\gamma_1 < \gamma_2$ and $\gamma_1, \gamma_2 \in \Gamma$. Assumption D.8$'$ implies $E\left[\mathbf{x}\mathbf{x}'\varepsilon^2 1_{(q\le\underline{\gamma})}\right] > 0$, and $E\left[\mathbf{x}\mathbf{x}'\varepsilon^2 1_{(q>\overline{\gamma})}\right] > 0$.

Theorem 2 *Under Assumptions D.1-D.4, D.5$'$ and D.8$'$, if $\sqrt{n}\delta_n \to b \in \mathbb{R}^k$, then*

$$\left(\widehat{\gamma}, \sqrt{n}\left(\widehat{\beta}_1 - \beta_{1n}, \widehat{\beta}_2 - \beta_{2n}\right)\right) \xrightarrow{d} \arg\max_h \Xi\left(h|\gamma_0, b\right),$$

where

$$\Xi\left(h|\gamma_0, b\right) = -\begin{cases} u_1'M(\gamma)u_1 + (b-u_2)'M(\gamma,\gamma_0)(b-u_2) + u_2'\overline{M}(\gamma_0)u_2 \\ -2u_1'W(\gamma) - 2u_2'\overline{W}(\gamma_0) + 2(b-u_2)'W(\gamma,\gamma_0), \text{ if } \gamma \le \gamma_0, \\ u_1'M(\gamma_0)u_1 + (b+u_1)'M(\gamma_0,\gamma)(b+u_1) + u_2'\overline{M}(\gamma)u_2 \\ -2u_1'W(\gamma_0) - 2u_2'\overline{W}(\gamma) - 2(b+u_1)'W(\gamma_0,\gamma), \text{ if } \gamma > \gamma_0, \end{cases}$$

with $h = (\gamma, u_1', u_2')' \in \Gamma \times \mathbb{R}^k \times \mathbb{R}^k$, $M(\gamma_1,\gamma_2) = M(\gamma_2) - M(\gamma_1)$ and $W(\gamma_1,\gamma_2) = W(\gamma_2) - W(\gamma_1)$ for $\gamma_1 \le \gamma_2$, $\overline{M}(\gamma) = M(\gamma,\infty)$ and $\overline{W}(\gamma) = W(\gamma,\infty)$, and $W(\gamma)$ is a Gaussian process with the covariance kernel equal to

$$K(\gamma_1,\gamma_2) = K(\gamma_1 \wedge \gamma_2) := E\left[\mathbf{x}\mathbf{x}'\varepsilon^2 1_{(q\le\gamma_1\wedge\gamma_2)}\right]$$

so that $\overline{W}(\gamma)$ is a Gaussian process with the covariance kernel equal to

$$\overline{K}(\gamma_1,\gamma_2) = \overline{K}(\gamma_1 \vee \gamma_2) := E\left[\mathbf{x}\mathbf{x}'\varepsilon^2 1_{(q\le\gamma_1\vee\gamma_2)}\right].$$

We provide a few comments on Theorem 2 here. First, when δ_n shrinks to zero in the rate of $\sqrt{n}$, γ_0 is not identifiable, but β_n still is. Second, $\widehat{\beta}$ and $\widehat{\gamma}$ are not asymptotically independent, which is dramatically different from the case in Hansen (2000). Third, consider $\Xi(h)$ in some special cases. When $b = \mathbf{0}$, $\Xi(h|\gamma_0, b)$ can be further simplified as

$$\Xi(h) = -u_1'M(\gamma)u_1 - u_2'\overline{M}(\gamma)u_2 + 2u_1'W(\gamma) + 2u_2'\overline{W}(\gamma),$$

which does not depend on γ_0 at all. The form of $\Xi(h)$ also reveals that $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are dependent through $\widehat{\gamma}$. Fourth, when $q \perp (\mathbf{x}, \varepsilon)$, then $W(\gamma) = K^{1/2}B(F(\gamma))$ with $K = E\left[\mathbf{x}\mathbf{x}'\varepsilon^2\right]$ and $B(\cdot)$ being a standard k-dimensional Brownian motion, where $F(\cdot)$ is the cdf of q.

Corollary 1 *Under the assumptions of Theorem 2, if $\sqrt{n}\delta_n \to b \in \mathbb{R}^k$, then*

$$\widehat{\gamma} \xrightarrow{d} \arg\max_{\gamma\in\Gamma} \Xi\left(\gamma|\gamma_0, b\right) =: \widetilde{\gamma},$$

where

$$\Xi(\gamma|\gamma_0,b)=\begin{cases} W(\gamma)' M(\gamma)^{-1} W(\gamma)+\left(\overline{W}(\gamma)+M(\gamma,\gamma_0)b\right)' \overline{M}(\gamma)^{-1}\left(\overline{W}(\gamma)+M(\gamma,\gamma_0)b\right) \\ -b'M(\gamma,\gamma_0)b-2W(\gamma,\gamma_0)'b,\ \textit{if } \gamma\leq\gamma_0, \\ \overline{W}(\gamma)'\overline{M}(\gamma)^{-1}\overline{W}(\gamma)+\left(W(\gamma)-M(\gamma_0,\gamma)b\right)' M(\gamma)^{-1}\left(W(\gamma)-M(\gamma_0,\gamma)b\right) \\ -b'M(\gamma_0,\gamma)b+2W(\gamma_0,\gamma)'b,\ \textit{if } \gamma>\gamma_0. \end{cases}$$

(ii)

$$\begin{aligned} &\sqrt{n}\left(\widehat{\beta}_1-\beta_{1n}\right)\xrightarrow{d} M(\widetilde{\gamma})^{-1}W(\widetilde{\gamma})-1_{(\widetilde{\gamma}>\gamma_0)}M(\widetilde{\gamma})^{-1}M(\gamma_0,\widetilde{\gamma})b, \\ &\sqrt{n}\left(\widehat{\beta}_2-\beta_{2n}\right)\xrightarrow{d} \overline{M}(\widetilde{\gamma})^{-1}\overline{W}(\widetilde{\gamma})+1_{(\widetilde{\gamma}\leq\gamma_0)}\overline{M}(\widetilde{\gamma})^{-1}M(\widetilde{\gamma},\gamma_0)b, \\ &\sqrt{n}\widehat{\delta}-b\xrightarrow{d} M(\widetilde{\gamma})^{-1}W(\widetilde{\gamma})-\overline{M}(\widetilde{\gamma})^{-1}\overline{W}(\widetilde{\gamma})-\left\{1_{(\widetilde{\gamma}>\gamma_0)}M(\widetilde{\gamma})^{-1}M(\gamma_0,\widetilde{\gamma})+1_{(\widetilde{\gamma}\leq\gamma_0)}\overline{M}(\widetilde{\gamma})^{-1}M(\widetilde{\gamma},\gamma_0)\right\}b. \end{aligned}$$

(iii)

$$LR_n(\gamma_0)\xrightarrow{d} LR_\infty(\gamma_0|b):=\frac{\max_{\gamma\in\Gamma}\Xi(\gamma|\gamma_0,b)-\Xi(\gamma_0)}{\sigma^2},$$

where $\frac{\Xi(\gamma_0)}{\sigma^2}=\frac{\Xi(\gamma_0|\gamma_0,b)}{\sigma^2}=\frac{W(\gamma_0)'}{\sigma}M(\gamma_0)^{-1}\frac{W(\gamma_0)}{\sigma}+\frac{\overline{W}(\gamma_0)'}{\sigma}\overline{M}(\gamma_0)^{-1}\frac{\overline{W}(\gamma_0)}{\sigma}$ *does not depend on* b *and follows* χ^2_{2k} *in the homoskedastic case.*

We provide a few comments on Corollary 1 here. First, from (i), the asymptotic distribution of $\widehat{\gamma}$ depends on γ_0, b, Γ, $K(\cdot,\cdot)$ and $M(\cdot,\cdot)$, but not β_2 (which is why we set $\beta_2=\mathbf{0}$ in our simulations of Section 4).[9] Second, from (ii), we cannot treat γ as known in the inference on β because there are some bias terms in the asymptotic distribution of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ beyond $M(\widetilde{\gamma})^{-1}W(\widetilde{\gamma})$ and $\overline{M}(\widetilde{\gamma})^{-1}\overline{W}(\widetilde{\gamma})$ unless $b=\mathbf{0}$; in other words, the asymptotic distribution of $\widehat{\beta}_\ell$ is a mixed normal with a bias. Third, b, as a local parameter, cannot be consistently estimated. Actually, $\widehat{\delta}$ tends to underestimate δ_n in absolute value. For example, if $\mathbf{x}=1$ and $\widehat{\gamma}>\gamma_0$, then $E\left[\sqrt{n}\widehat{\delta}\right]$ converges to $E\left[\frac{F(\gamma_0)}{F(\widetilde{\gamma})}\right]b$ whose absolute value is less than $|b|$. This is intuitively correct since when $\widehat{\gamma}>\gamma_0$, plim$\widehat{\beta}_1$ would be an average of β_1 and β_2 such that the difference between $\widehat{\beta}_1$ and $\widehat{\beta}_2$ would be smaller than δ_n in absolute value; similar analysis applies to $\widehat{\gamma}<\gamma_0$. Fourth, if $\widehat{\gamma}$ were consistent, then the distribution of $\widetilde{\gamma}$ would degenerate to a point mass at γ_0 and $M(\widetilde{\gamma})^{-1}W(\widetilde{\gamma})-1_{(\widetilde{\gamma}>\gamma_0)}M(\widetilde{\gamma})^{-1}M(\gamma_0,\widetilde{\gamma})b=M(\gamma_0)^{-1}W(\gamma_0)$ and $\overline{M}(\widetilde{\gamma})^{-1}\overline{W}(\widetilde{\gamma})+1_{(\widetilde{\gamma}\leq\gamma_0)}\overline{M}(\widetilde{\gamma})^{-1}M(\widetilde{\gamma},\gamma_0)b=\overline{M}(\gamma_0)^{-1}\overline{W}(\gamma_0)$ as usual. Fifth, the proof of (iii) also shows that $LR_n(\gamma_0)$ has a different asymptotic distribution from $LR_{1n}(\gamma_0)=\frac{S_n(\gamma_0,\widehat{\beta})-S_n(\widehat{\gamma},\widehat{\beta})}{S_n(\widehat{\gamma},\widehat{\beta})/n}$ and $LR_{2n}(\gamma_0)=\frac{S_n(\gamma_0,\widehat{\beta}(\gamma_0))-S_n(\widehat{\gamma}(\widehat{\beta}(\gamma_0)),\widehat{\beta}(\gamma_0))}{S_n(\widehat{\gamma},\widehat{\beta})/n}$, where $\widehat{\gamma}\left(\widehat{\beta}(\gamma_0)\right)=\arg\min_{\gamma\in\Gamma}S_n\left(\gamma,\widehat{\beta}(\gamma_0)\right)$. This is dramatically different from the case with γ_0 being identifiable where $LR_n(\gamma_0), LR_{1n}(\gamma_0)$ and $LR_{2n}(\gamma_0)$ have the same distribution. Sixth, $LR_\infty(\gamma_0|b)$ depends on b, which is why the coverage in Tables 2 and 3 varies with the size of δ_{20}. This fact reveals a key difference between the nonregular parameter γ and the usual regular parameter under weak identification, e.g., Staiger and Stock (1997) show that the LR statistic for slope parameters is invariant to the identification strength in the weak IV model. Seventh, in the homoskedastic

[9] Note that due to the symmetricity of the Gaussian distribution, $\Xi(\gamma|\gamma_0,b)$ and $\Xi(\gamma|\gamma_0,-b)$ have the same distribution, which implies the asymptotic distribution of $\widehat{\gamma}$ is symmetric in the sign of b.

case, $\Xi(\gamma|\gamma_0, b)/\sigma^2$ reduces to

$$\Xi_o(\gamma|\gamma_0, b_o) = \begin{cases} W_o(\gamma)' M(\gamma)^{-1} W_o(\gamma) + \left(\overline{W}_o(\gamma) + M(\gamma,\gamma_0) b_o\right)' \overline{M}(\gamma)^{-1} \left(\overline{W}_o(\gamma) + M(\gamma,\gamma_0) b_o\right) \\ -b_o' M(\gamma,\gamma_0) b_o - 2W_o(\gamma,\gamma_0)' b_o, \text{ if } \gamma \le \gamma_0, \\ \overline{W}_o(\gamma)' \overline{M}(\gamma)^{-1} \overline{W}_o(\gamma) + \left(W_o(\gamma) - M(\gamma_0,\gamma) b_o\right)' M(\gamma)^{-1} \left(W_o(\gamma) - M(\gamma_0,\gamma) b_o\right) \\ -b_o' M(\gamma_0,\gamma) b_o + 2W_o(\gamma_0,\gamma)' b_o, \text{ if } \gamma > \gamma_0, \end{cases}$$

where $W_o(\gamma)$ is a Gaussian process with the covariance kernel equal to $K_o(\gamma_1,\gamma_2) = M(\gamma_1 \wedge \gamma_2)$, $\overline{W}_o(\gamma_1,\gamma_2)$ and $\overline{W}_o(\gamma)$ are similarly defined as $\overline{W}(\gamma_1,\gamma_2)$ and $\overline{W}(\gamma)$, and $b_o = b/\sigma$. Eighth, when $b = \mathbf{0}$, $\Xi(\gamma|\gamma_0, b)$ reduces to

$$\Xi(\gamma) = W(\gamma)' M(\gamma)^{-1} W(\gamma) + \overline{W}(\gamma)' \overline{M}(\gamma)^{-1} \overline{W}(\gamma),$$

which does not depend on γ_0. In the homoskedastic case, $\mathbb{B}_-(\gamma) := M(\gamma)^{-1/2} W(\gamma)/\sigma$ (and $\mathbb{B}_+(\gamma) := \overline{M}(\gamma)^{-1/2}\overline{W}(\gamma)/\sigma$) is a Gaussian process with the variance kernel equal to $M(\gamma_1)^{-1/2} M(\gamma_1 \wedge \gamma_2) M(\gamma_2)^{-1/2}$ ($\overline{M}(\gamma_1)^{-1/2}\overline{M}(\gamma_1 \vee \gamma_2)\overline{M}(\gamma_2)^{-1/2}$) which is I_k when $\gamma_1 = \gamma_2$, so $\Xi(\gamma)$ is a χ^2 process $\mathbb{B}_-(\gamma)'\mathbb{B}_-(\gamma) + \mathbb{B}_+(\gamma)'\mathbb{B}_+(\gamma)$ which follows χ^2_{2k} at any γ. Ninth, $\max_{\gamma\in\Gamma}\Xi(\gamma) - \Xi(\gamma_0)$ is different from the null distribution of the LR test in testing $H_0 : \beta_1 = \beta_2$ vs. $H_1 : \beta_1 \neq \beta_2$. Suppose the test statistic is $n\left(\widetilde{\sigma}^2 - \widehat{\sigma}^2\right)$, where

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \widetilde{\varepsilon}_i^2 \text{ and } \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \widehat{\varepsilon}_i^2 = \frac{S_n(\widehat{\gamma})}{n}$$

with $\widetilde{\varepsilon}_i$ being the residual in regressing y_i on $\mathbf{x}_i$ and $\widehat{\varepsilon}_i$ being the residual in threshold regression; then from Chan and Tong (1990), the asymptotic null distribution is

$$\max_{\gamma\in\Gamma} \zeta(\gamma)' \left[M(\gamma) - M(\gamma) M^{-1} M(\gamma)\right]^{-1} \zeta(\gamma),$$

where $\zeta(\gamma)$ is a Gaussian process with the covariance kernel equal to $K(\gamma_1,\gamma_2) - M(\gamma_1) M^{-1} K M^{-1} M(\gamma_2)$. In the homoskedastic case, the null distribution of $n\left(\widetilde{\sigma}^2 - \widehat{\sigma}^2\right)/\widehat{\sigma}^2$ takes the same form $\sup_\gamma \zeta(\gamma)' \left[M(\gamma) - M(\gamma) M^{-1} M(\gamma)\right]$ but the covariance kernel of $\zeta(\gamma)$ reduces to $M(\gamma_1 \wedge \gamma_2) - M(\gamma_1) M^{-1} M(\gamma_2)$, so the null distribution is a χ^2 process. The difference is due to $\widetilde{\sigma}^2 \neq \frac{S_n(\gamma_0)}{n}$, where $n\widetilde{\sigma}^2 - \sum_{i=1}^n \varepsilon_i^2$ converges to $-W'M^{-1}W$, while $S_n(\gamma_0) - \sum_{i=1}^n \varepsilon_i^2$ converges to $-\Xi(\gamma_0)$, and $\Xi(\gamma_0) \neq W'M^{-1}W$, where $W = W(\infty) \sim N(0, K)$. Tenth, when β is known, then

$$\Xi(\gamma|\gamma_0, b) = -\begin{cases} b'M(\gamma,\gamma_0) b + 2b'W(\gamma,\gamma_0), \text{ if } \gamma \le \gamma_0, \\ b'M(\gamma_0,\gamma) b - 2b'W(\gamma_0,\gamma), \text{ if } \gamma > \gamma_0; \end{cases}$$

when assume further that $b = \mathbf{0}$, then $\Xi(\gamma|\gamma_0, b) = 0$ as expected.

In the following two examples, we show that the asymptotic distributions of $\widehat{\gamma}$ appear in the structural change literature can be expressed as special cases of our Theorem 2 and Corollary 1.

Example 1 *In the structural change case, $q \perp (\mathbf{x}, \varepsilon)$ and follows $U[0,1]$, so that $\Xi(\gamma|\gamma_0, b)$ can be much simplified. If $q \sim U[0,1]$, then $F(\gamma) = \gamma$, $M(\gamma_1,\gamma_2) = (\gamma_2 - \gamma_1) M$ with $\gamma_1 < \gamma_2$, and $W(\gamma) = K^{1/2} B(\gamma)$.*

Now,

$$\Xi\left(\gamma|\gamma_0,b\right) = \begin{cases} \frac{W(\gamma)'M^{-1}W(\gamma)}{\gamma} + \left(W\left(\gamma,1\right)+\left(\gamma_0-\gamma\right)Mb\right)'\frac{M^{-1}}{1-\gamma}\left(W\left(\gamma,1\right)+\left(\gamma_0-\gamma\right)Mb\right) \\ -\left(\gamma_0-\gamma\right)b'Mb - 2W\left(\gamma,\gamma_0\right)'b, \text{ if } \gamma\le\gamma_0, \\ \frac{W(\gamma,1)'M^{-1}W(\gamma,1)}{1-\gamma} + \left(W\left(\gamma\right)-\left(\gamma-\gamma_0\right)Mb\right)'\frac{M^{-1}}{\gamma}\left(W\left(\gamma\right)-\left(\gamma-\gamma_0\right)Mb\right) \\ -\left(\gamma-\gamma_0\right)b'Mb + 2W\left(\gamma_0,\gamma\right)'b, \text{ if } \gamma>\gamma_0, \end{cases}$$

$$\stackrel{d}{=} \frac{\widetilde{W}(\gamma)'M^{-1}\widetilde{W}(\gamma)}{\gamma} + \frac{\left(\widetilde{W}(1)-\widetilde{W}(\gamma)\right)'M^{-1}\left(\widetilde{W}(1)-\widetilde{W}(\gamma)\right)}{1-\gamma} - C,$$

where $\widetilde{W}(\gamma) = W\left(\gamma\right) - 1_{(\gamma>\gamma_0)}\left(\gamma-\gamma_0\right)\Sigma_{\mathbf{x}}b$,[10] $\stackrel{d}{=}$ *means "equal in distribution", and* $C = \left(1-\gamma_0\right)b'Mb + 2b'W\left(\gamma_0,1\right)$ *is a random variable that does not depend on* γ*. Also,*

$$\sqrt{n}\left(\widehat{\beta}_1-\beta_{1n}\right) \xrightarrow{d} \frac{1}{\widetilde{\gamma}}M^{-1}W\left(\widetilde{\gamma}\right) - 1_{(\widetilde{\gamma}>\gamma_0)}\frac{\widetilde{\gamma}-\gamma_0}{\widetilde{\gamma}}b,$$

$$\sqrt{n}\left(\widehat{\beta}_2-\beta_{2n}\right) \xrightarrow{d} \frac{1}{1-\widetilde{\gamma}}M^{-1}W\left(\widetilde{\gamma},1\right) + 1_{(\widetilde{\gamma}\le\gamma_0)}\frac{\gamma_0-\widetilde{\gamma}}{1-\widetilde{\gamma}}b,$$

$$\sqrt{n}\widehat{\delta} - b \xrightarrow{d} M^{-1}\frac{W\left(\widetilde{\gamma}\right)-\widetilde{\gamma}W\left(1\right)}{\widetilde{\gamma}\left(1-\widetilde{\gamma}\right)} - \frac{1_{(\widetilde{\gamma}>\gamma_0)}\left(1-\widetilde{\gamma}\right)\left(\widetilde{\gamma}-\gamma_0\right)+1_{(\widetilde{\gamma}\le\gamma_0)}\widetilde{\gamma}\left(\gamma_0-\widetilde{\gamma}\right)}{\widetilde{\gamma}\left(1-\widetilde{\gamma}\right)}b,$$

which implies

$$\sqrt{n}\widehat{\delta} \xrightarrow{d} M^{-1}\frac{M\left(\widetilde{\gamma}\right)-\widetilde{\gamma}M\left(1\right)}{\widetilde{\gamma}\left(1-\widetilde{\gamma}\right)},$$

where $\widetilde{\gamma} = \arg\max_{\gamma\in\Gamma}\Xi\left(\gamma|\gamma_0,b\right)$*. These asymptotic distributions are the same as claimed in Proposition 1 of EM.*

Example 2 *Jiang et al. (2018) consider the structural change model where* $\mathbf{x}=1$ *and* $E\left[\varepsilon^2\right]=\sigma^2$ *which implies* $M=1$ *and* $W(\gamma)=\sigma B\left(\gamma\right)$*. If* β_n *were known, then*

$$\Xi\left(\gamma|\gamma_0,b\right) \stackrel{d}{=} -\left|\gamma-\gamma_0\right|b^2 - 2\sigma bB\left(\gamma-\gamma_0\right),$$

where $B\left(\cdot\right)$ *is a standard two-sided Brownian motion. As a result,*

$$\frac{b^2}{\sigma^2}\left(\widehat{\gamma}-\gamma_0\right) \xrightarrow{d} \arg\max_{v\in\frac{b^2}{\sigma^2}(\Gamma-\gamma_0)} 2\sigma^2\left[-\frac{|v|}{2}+B\left(v\right)\right] = \arg\max_{v\in\frac{b^2}{\sigma^2}(\Gamma-\gamma_0)}\left[-\frac{|v|}{2}+B\left(v\right)\right],$$

which is the in-fill asymptotic distribution in their Theorem 4.1(a) when $\Gamma=(0,1)$*. If* β_n *is unknown, then*

$$\begin{aligned} & \frac{\widetilde{W}(\gamma)'M^{-1}\widetilde{W}(\gamma)}{\gamma} + \frac{\left(\widetilde{W}(1)-\widetilde{W}(\gamma)\right)'M^{-1}\left(\widetilde{W}(1)-\widetilde{W}(\gamma)\right)}{1-\gamma} \\ = \; & C + \begin{cases} \sigma^2\left(\frac{B(\gamma)-\gamma B(1)}{\sqrt{\gamma(1-\gamma)}} + \frac{(1-\gamma_0)\sqrt{\gamma}}{\sqrt{1-\gamma}}\frac{b}{\sigma}\right)^2, \text{ if } \gamma\le\gamma_0, \\ \sigma^2\left(\frac{B(\gamma)-\gamma B(1)}{\sqrt{\gamma(1-\gamma)}} + \frac{\gamma_0\sqrt{1-\gamma}}{\sqrt{\gamma}}\frac{b}{\sigma}\right)^2, \text{ if } \gamma>\gamma_0, \end{cases} \\ = \; & C + \sigma^2\left(\frac{G\left(\gamma\right)-\gamma G\left(1\right)}{\sqrt{\gamma\left(1-\gamma\right)}}\right)^2, \end{aligned}$$

[10] Note that EM's d is like our $-b$, and $\widetilde{W}(\gamma)$ is their $M\left(\gamma\right)$.

where $G(\gamma) = B(\gamma) + \frac{b}{\sigma}\min(\gamma, \gamma_0)$ *and* $C = \sigma^2\left[B(1) + (1-\gamma_0)\frac{b}{\sigma}\right]^2$ *does not depend on* γ. *So*

$$\widehat{\gamma} \xrightarrow{d} \arg\max_{\gamma\in\Gamma}\begin{cases}\left(\frac{B(\gamma)-\gamma B(1)}{\sqrt{\gamma(1-\gamma)}} + \frac{(1-\gamma_0)\sqrt{\gamma}}{\sqrt{1-\gamma}}\frac{b}{\sigma}\right)^2, \text{ if } \gamma \le \gamma_0,\\ \left(\frac{B(\gamma)-\gamma B(1)}{\sqrt{\gamma(1-\gamma)}} + \frac{\gamma_0\sqrt{1-\gamma}}{\sqrt{\gamma}}\frac{b}{\sigma}\right)^2, \text{ if } \gamma > \gamma_0,\end{cases}$$

which is the in-fill asymptotic distribution in their Theorem 4.2(a) when $\Gamma = (0,1)$. *When* $b = 0$, *the objective function* $\left(\frac{B(\gamma)-\gamma B(1)}{\sqrt{\gamma(1-\gamma)}}\right)^2$ *appears in the null distribution of testing for structural change (see, e.g., Theorem 3 of Andrews (1993)).*

Note that although the in-fill asymptotics of Jiang et al. (2018) are seminal in the structural change literature, they do not seem applicable in threshold regression because q *is usually dependent of* $(\mathbf{x}, \varepsilon_1, \varepsilon_2)$ *(e.g.,* $q \in \mathbf{x}$*) and does not follow* $U[0,1]$.

In the online supplement, we show the distribution of $\widetilde{\gamma}$ and $\widehat{\gamma}$ under the two DGPs of Section 4.

5.1 Interval Coverage

As in Section 3, we can also study the interval coverage when $b \in \mathbb{R}^k$.

Theorem 3 *When* $\delta_n = n^{-1/2}b$,

$$\begin{aligned} P(\gamma_0 \in \text{conv}\{\gamma : L_n(\gamma) \le x\}) \quad &\to \quad P_{[\cdot]}(x) = P\left(\frac{LR_{[\cdot]}(\gamma_0|b)}{\sigma^2} \le x\right)\\ &: \quad = P\left(\frac{\max\left[\xi_1^b(\gamma_0), \xi_2^b(\gamma_0)\right] - \min\left[\xi_1^b(\gamma_0), \xi_2^b(\gamma_0)\right]}{\sigma^2} \le x\right),\end{aligned}$$

where $x \in \mathbb{R}_+ := (0,\infty)$, $\xi_1^b(\gamma_0) = \max_{\underline{\gamma}\le\gamma\le\gamma_0}\{\Xi(\gamma|\gamma_0, b)\}$ *and* $\xi_2^b(\gamma_0) = \max_{\gamma_0\le\gamma\le\overline{\gamma}}\{\Xi(\gamma|\gamma_0, b)\}$.

For comparison,

$$\begin{aligned} P(L_n(\gamma_0) \le x) \quad &\to \quad P_{LR}(x) := P\left(\max_{\gamma\in\Gamma}\Xi(\gamma|\gamma_0, b) - \Xi(\gamma_0) \le \sigma^2 x\right)\\ &= \quad P\left(\max\left[\xi_1^b(\gamma_0), \xi_2^b(\gamma_0)\right] - \Xi(\gamma_0) \le \sigma^2 x\right)\\ &\le \quad P\left(\max\left[\xi_1^b(\gamma_0), \xi_2^b(\gamma_0)\right] - \min\left[\xi_1^b(\gamma_0), \xi_2^b(\gamma_0)\right] \le \sigma^2 x\right) = P_{[\cdot]}(x).\end{aligned}$$

Different from Theorem 1, $\xi_1^b(\gamma_0)$ and $\xi_2^b(\gamma_0)$ are correlated and depend on b and γ_0 (and also Γ). Often, we suppress the dependence on γ_0 and write $\xi_1^b(\gamma_0)$ and $\xi_2^b(\gamma_0)$ as ξ_1^b and ξ_2^b if this would not introduce any confusion.

5.2 Numerical Illustration

In the subsection, we analyze the simulation evidences in Section 4 using the asymptotics developed in Theorem 3 and Corollary 1. Because the asymptotic distributions of $\widehat{\gamma}$ and $LR_n(\gamma_0)$ depend on Γ, we set $\Gamma = [1,3]$. The asymptotics in Hansen (2000) cannot explain why when $\sqrt{n}\delta_{20}$ is small, the CI based on inverting the LR statistics does not match the target coverage. To simplify our discussion, we consider only the $n = 100$ case.

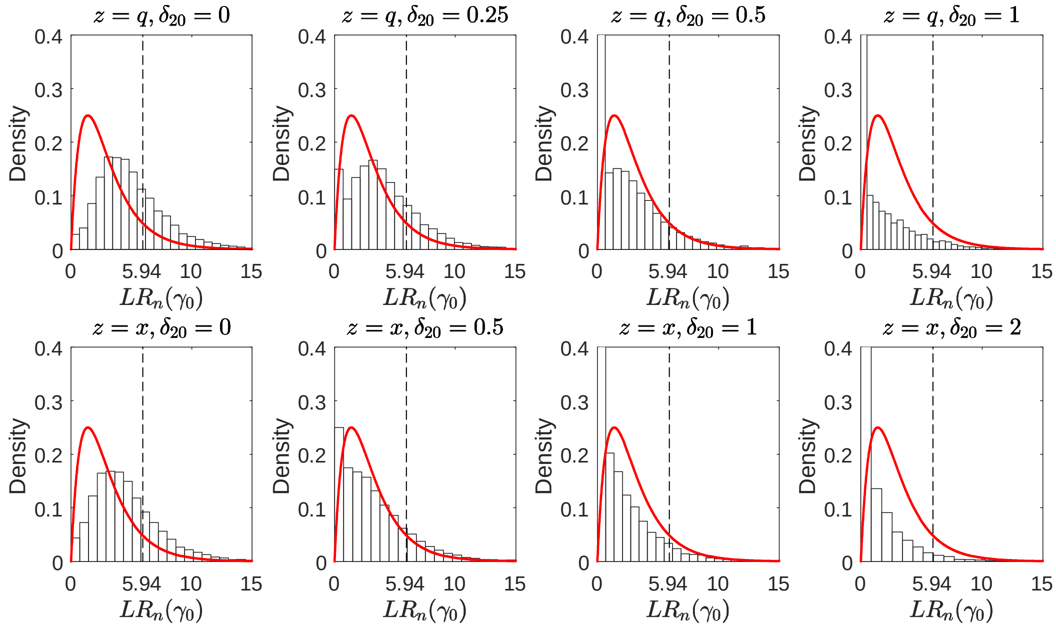

Figure 3: Finite Sample Density of $LR_n(\gamma_0)$ and Its Asymptotic Approximation ξ for $z=q$ or x and Different δ_{20} Values

First, we show intuitively why the actual coverage may be different from 0.9. Figure 3 shows the finite sample distribution of $LR_n(\gamma_0)$ and ξ. When the right tail of the finite sample distribution is thicker than that of the asymptotic distribution, the actual coverage is smaller than 90%, and vice versa. Interestingly, even when δ_{20} is large, the distribution of $LR_n(\gamma_0)$ does not coincide with that of ξ. In other words, the joint asymptotic as n and b diverge to ∞ simultaneously is different from the sequential asymptotic by first letting n diverge to ∞ and then letting b diverge to ∞.

Table 5 reports $P_{LR}(c_{0.9})$, $P_{[\cdot]}(c_{0.9})$, $P_{LR}^{-1}(0.9)$, and $P_{[\cdot]}^{-1}(0.9)$, where $P_{LR}^{-1}(\cdot)$ and $P_{[\cdot]}^{-1}(\cdot)$ are inverse functions of $P_{LR}(\cdot)$ and $P_{[\cdot]}(\cdot)$. To obtain $P_{LR}(\cdot)$ and $P_{[\cdot]}(\cdot)$, we need to get $M(\gamma)$ and $K(\cdot)$. In DGP1,

$$\begin{aligned}
M(\gamma) &= K(\gamma)=E\left[\begin{pmatrix}1\\ q\end{pmatrix}(1,q)\,1_{(q\le\gamma)}\right]\\
&= \begin{pmatrix}\Phi(\gamma-2) & 2\Phi(\gamma-2)-\phi(\gamma-2)\\ 2\Phi(\gamma-2)-\phi(\gamma-2) & 4\Phi(\gamma-2)-4\phi(\gamma-2)+\Phi(\gamma-2)-(\gamma-2)\phi(\gamma-2)\end{pmatrix}\\
&= \begin{pmatrix}\Phi(\gamma-2) & 2\Phi(\gamma-2)-\phi(\gamma-2)\\ 2\Phi(\gamma-2)-\phi(\gamma-2) & 5\Phi(\gamma-2)-(\gamma+2)\phi(\gamma-2)\end{pmatrix},
\end{aligned}$$

and in DGP2,

$$M(\gamma)=K(\gamma)=E\left[\begin{pmatrix}1\\ x\end{pmatrix}(1,x)\,1_{(q\le\gamma)}\right]=\Phi(\gamma-2)\,\mathbf{I}_2,$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cdf and pdf of standard normal, respectively. Compared with Table 2 when $n=100$, $P_{LR}(c_{0.9})$ and $P_{[\cdot]}(c_{0.9})$ match their finite-sample counterparts, say, $P_{n,LR}(c_{0.9})$ and $P_{n,[\cdot]}(c_{0.9})$, quite well. The appropriate critical values depend on b_2 and may be larger or smaller than $c_{0.9}$. Correspondingly,

$P_{LR}(c_{0.9})$ and $P_{[\cdot]}(c_{0.9})$ may be larger or smaller than 0.9 and $p(c_{0.9}) = 0.949$.

	$z = q$				$z = x$				
$b =$	0	2.5	5	10	0	5	10	20	∞
$P_{LR}(c_{0.9})$	0.703	0.781	0.889	0.948	0.724	0.832	0.913	0.944	0.900
$P_{[\cdot]}(c_{0.9})$	0.917	0.920	0.938	0.961	0.915	0.916	0.943	0.958	0.949
$P_{LR}^{-1}(0.9)$	8.508	7.723	6.200	4.554	8.267	7.241	5.620	4.712	5.939
$P_{[\cdot]}^{-1}(0.9)$	5.538	5.541	4.992	4.035	5.594	5.553	4.787	4.195	4.605

Table 5: Asymptotic Approximate Coverage and Critical Value
Note: $c_{0.9} = 5.939 = F_{\xi}^{-1}(0.9)$

In the following, we provide more detailed analysis on $P_{LR}(\cdot)$, $P_{[\cdot]}(\cdot)$, $P_{n,LR}(\cdot)$ and $P_{n,[\cdot]}(\cdot)$. First, Figure 4 shows $P_{LR}(\cdot)$ for different b_2 values and $P_{n,LR}(\cdot)$ for the corresponding δ_{20} values. From the figure, we can draw three conclusions. (i) $P_{LR}(\cdot)$ and $P_{n,LR}(\cdot)$ match quite well especially in the right tail, while ξ matches $P_{n,LR}(\cdot)$ closely in the right tail only when $\delta_{20} = 0.5$ when $z = q$ and $\delta_{20} = 1$ when $z = x$. (ii), the ranking of the distributions of $P_{LR}(\cdot)$ (or $P_{n,LR}(\cdot)$) and ξ in the right tail depends on the value of b_2. In other words, the actual coverage based on the critical value $c_{0.9}$ depends on b_2, which matches the simulation results in Table 2. (iii) When $b_2 \neq 0$, there is a point mass at zero in the distributions of $P_{LR}(\cdot)$, which matches the finite sample distribution in Figure 3. Actually, when $|b_2| \to \infty$, $P_{LR}(\cdot)$ converges to a point mass at zero. This can be seen from the form of $LR_{\infty}(\gamma_0|b)$. When $|b| \to \infty$, the term $-b'M(\gamma \wedge \gamma_0, \gamma \vee \gamma_0)b$ dominates, and its maximum is achieved at $\gamma = \gamma_0$ so that $LR_{\infty}(\gamma_0|b) = 0$.[11] Second, Figure 5 shows $P_{[\cdot]}(\cdot)$ and $P_{n,[\cdot]}(\cdot)$ corresponding to Figure 4. The three conclusions can be parallelly applied here.

In summary, different from the regular case as discussed in Sections 4.4-4.7 of Andrews and Cheng (2012), the distribution of $LR_{\infty}(\gamma_0|b)$ (and $LR_{[\cdot]}(\gamma_0|b)$) when $b \in \mathbb{R}^k$ is very different from that when $\max_j |b_j| \to \infty$ (as $n \to \infty$) such that the LR-CI based on ξ is not reliable.

Finally, Figure 6 shows the distribution of $\left|\sqrt{n}\left(\widehat{\delta}_2 - \delta_{20}\right)\right|$, its asymptotic approximation in Corollary 1, and the usual normal approximation. Although the approximation in Corollary 1 has a thinner tail then the finite-sample distribution, it is still better than the normal approximation in most cases, while the normal approximation is good only when b_2 is large (which is much expected).

5.3 Least Favorable Distribution

Because $LR_{\infty}(\gamma_0|b)$ depends on b, and b cannot be consistently estimated, we take the least favorable distribution of $LR_{\infty}(\gamma_0|b)$ to avoid undercoverage.[12] We guess $b = \mathbf{0}$ is the least favorable case. Specifically, for $\min_{b \in \mathbb{R}^k} P(\gamma_0 \in \{\gamma | LR_{\infty}(\gamma|b) \leq \bar{c}_{1-\alpha}(\gamma)\}) \geq 1 - \alpha$, we need show that

$$P\left(\max_{\gamma \in \Gamma} \Xi(\gamma|\gamma_0, b) \leq x\right) \geq P\left(\max_{\gamma \in \Gamma} \Xi(\gamma) \leq x\right) \tag{3}$$

[11] This may seem different from the distribution of ξ. However, ξ is achieved by maximizing $S_n(\gamma_0) - S_n(\gamma)$ over the localized parameter space $\left\{\gamma : \gamma = \gamma_0 + a_n^{-1}v, v \in \mathbb{R}, a_n = n^{1-2\alpha}\right\}$ which is dramatically different from Γ. The point is that which approximation would match the finite sample distribution better.

[12] As in McCloskey (2017), one may suggest to consider the least favorable distribution over b in a confidence set instead $\mathbb{R}_{\infty}^k$, but this is infeasible in practice as discussed in the Introduction although such a least favorable distribution is more informative.

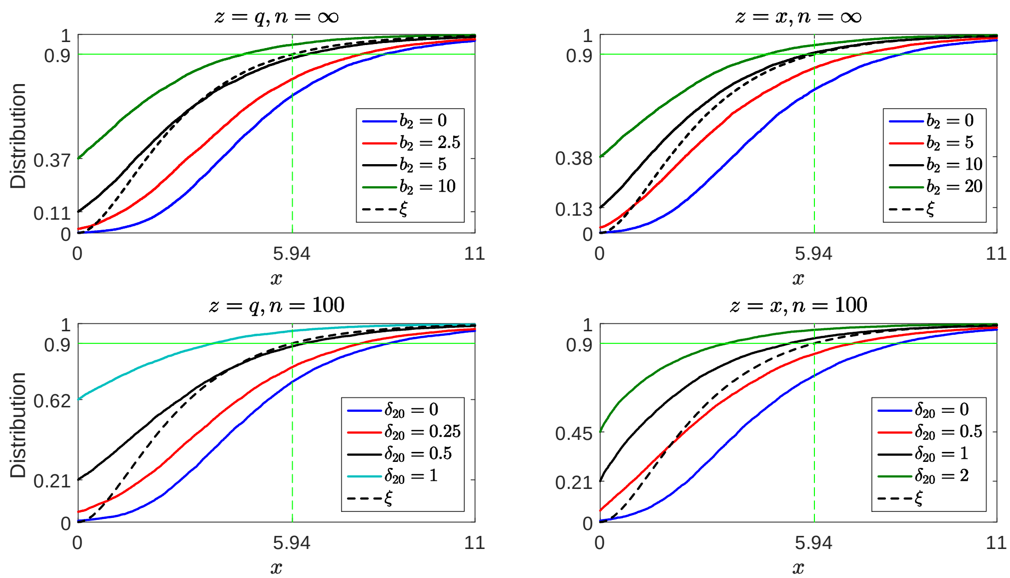

Figure 4: $P_{LR}(\cdot)$ and $P_{n,LR}(\cdot)$

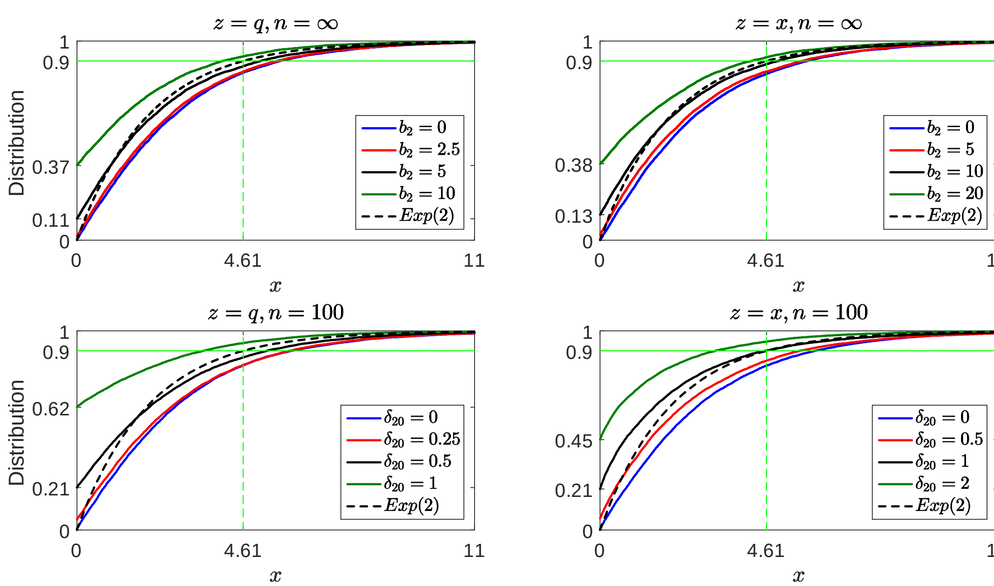

Figure 5: $P_{[\cdot]}(\cdot)$ and $P_{n,[\cdot]}(\cdot)$

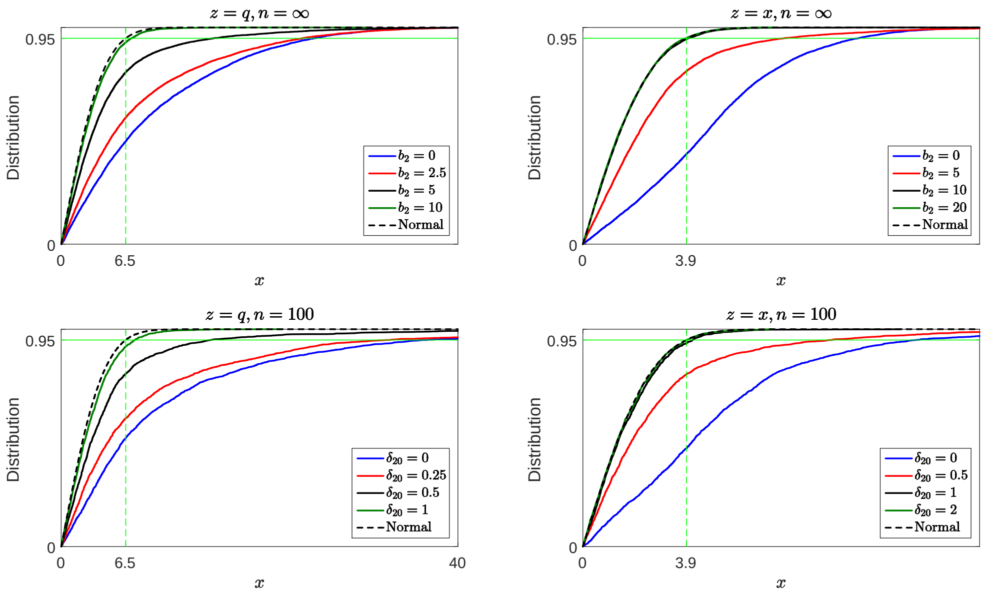

Figure 6: Distribution of $\left|\sqrt{n}\left(\widehat{\delta}_2 - \delta_{20}\right)\right|$ and Its Asymptotic Approximation

for any $b \neq \mathbf{0}$ and $x \in \mathbb{R}_+$, and for $\min_{b\in\mathbb{R}^k} P\left(\gamma_0 \in \text{conv}\left\{\gamma | LR_\infty\left(\gamma|b\right) \leq \underline{c}_{1-\alpha}\left(\gamma\right)\right\}\right) \geq 1-\alpha$, we need show that

$$P\left(\max\left[\xi_1^b, \xi_2^b\right] - \min\left[\xi_1^b, \xi_2^b\right] \leq x\right) \geq P\left(\max\left[\xi_1^0, \xi_2^0\right] - \min\left[\xi_1^0, \xi_2^0\right] \leq x\right) \tag{4}$$

for any $b \neq \mathbf{0}$ and $x \in \mathbb{R}_+$, where $P\left(LR_\infty\left(\gamma_0|\mathbf{0}\right) \leq \overline{c}_{1-\alpha}\left(\gamma_0\right)\right) = 1-\alpha$, and $P\left(LR_{[\cdot]}\left(\gamma_0|\mathbf{0}\right) \leq \underline{c}_{1-\alpha}\left(\gamma_0\right)\right) = 1-\alpha$. Note that because $LR_\infty\left(\gamma_0|\mathbf{0}\right)$ and $LR_{[\cdot]}\left(\gamma_0|\mathbf{0}\right)$ are not pivotal to γ_0, the critical values $\overline{c}_{1-\alpha}\left(\gamma_0\right)$ and $\underline{c}_{1-\alpha}\left(\gamma_0\right)$ depend on γ_0. When $\max_j |b_j| = \infty$, we need further show

$$P\left(\eta^2\xi \leq x\right) \geq P\left(\max_{\gamma\in\Gamma} \Xi\left(\gamma\right) - \Xi\left(\gamma_0\right) \leq \sigma^2 x\right) \tag{5}$$

and

$$P\left(2\eta^2\left(\max\left[\xi_1, \xi_2\right] - \min\left[\xi_1, \xi_2\right]\right) \leq x\right) \geq P\left(\max\left[\xi_1^0, \xi_2^0\right] - \min\left[\xi_1^0, \xi_2^0\right] \leq \sigma^2 x\right) \tag{6}$$

for any $x \in \mathbb{R}_+$ to justify that $b = \mathbf{0}$ is the least favorable case, where note that ξ is the asymptotic distribution of $LR_n\left(\gamma_0\right)/\widehat{\eta}^2$ while $\left(\max_{\gamma\in\Gamma} \Xi\left(\gamma\right) - \Xi\left(\gamma_0\right)\right)/\sigma^2$ is the asymptotic distribution of $LR_n\left(\gamma_0\right)$ so that $\eta^2\xi$ is comparable with $\left(\max_{\gamma\in\Gamma} \Xi\left(\gamma\right) - \Xi\left(\gamma_0\right)\right)/\sigma^2$, similarly, $2\eta^2\left(\max\left[\xi_1, \xi_2\right] - \min\left[\xi_1, \xi_2\right]\right)$ is comparable with $\left(\max\left[\xi_1^0, \xi_2^0\right] - \min\left[\xi_1^0, \xi_2^0\right]\right)/\sigma^2$, and ξ_1 and ξ_2 are defined in Theorem 1.

In practice, we always use an interval to cover γ_0, so we need only check (4) and (6). To check these two inequalities in practice, we need to simulate $\Xi\left(\gamma|\gamma_0, b\right)$ for $b \in \mathbb{R}^k$ which requires the estimation of nuisance parameters like $K\left(\cdot,\cdot\right)$ and $M\left(\cdot\right)$; see Section 6.1 below for the details of such estimation.[13] In Figures 4 and 5, these four inequalities all hold under the DGPs of our simulation where $\eta^2 = 1$. We guess they would still hold in the general case but this is not easy to show due to the complicated forms of $LR_\infty\left(\gamma_0|b\right)$ and $LR_{[\cdot]}\left(\gamma_0|b\right)$. Finally, note that we have only the stochastic dominance but not the almost sure dominance in

[13] In practice, we could simulate b uniformly from a big ball centering at $\mathbf{0}$ and check the first two inequalities hold only for these b values in a specific application. Actually, we need only check half of the ball because $\Xi\left(\gamma|\gamma_0, b\right)$ and $\Xi\left(\gamma|\gamma_0, -b\right)$ share the same distribution.

(3) and (4), i.e., the stronger results $\max_{\gamma\in\Gamma}\Xi(\gamma|\gamma_0,b)\le\max_{\gamma\in\Gamma}\Xi(\gamma)$ and $\left(\max\left[\xi_1^b,\xi_2^b\right]-\min\left[\xi_1^b,\xi_2^b\right]\right)\le\left(\max\left[\xi_1^0,\xi_2^0\right]-\min\left[\xi_1^0,\xi_2^0\right]\right)$ do not hold.

6 New Confidence Intervals

In this section, we develop our CIs for γ and β.

6.1 Confidence Interval for γ

Given $b\in\mathbb{R}^k$, we can show that

$$\begin{aligned}\sqrt{n}\widehat{\delta} &= E\left[\mathbf{x}\mathbf{x}'1_{(q\le\widehat{\gamma})}\right]^{-1}n^{-1/2}\sum_{i=1}^n\mathbf{x}_i\varepsilon_i1_{(q_i\le\widehat{\gamma})}-E\left[\mathbf{x}\mathbf{x}'1_{(q>\widehat{\gamma})}\right]^{-1}n^{-1/2}\sum_{i=1}^n\mathbf{x}_i\varepsilon_i1_{(q_i>\widehat{\gamma})}\\ &\quad+\left\{1-1_{(\widehat{\gamma}>\gamma_0)}E\left[\mathbf{x}\mathbf{x}'1_{(q\le\widehat{\gamma})}\right]^{-1}E\left[\mathbf{x}\mathbf{x}'1_{(\gamma_0<q\le\widehat{\gamma})}\right]-1_{(\widehat{\gamma}\le\gamma_0)}E\left[\mathbf{x}\mathbf{x}'1_{(q>\widehat{\gamma})}\right]^{-1}E\left[\mathbf{x}\mathbf{x}'1_{(\widehat{\gamma}<q\le\gamma_0)}\right]\right\}b+o_p(1)\\ &=: T_1+T_2+T_3+o_p(1),\end{aligned}$$

where $E\left[\mathbf{x}\mathbf{x}'1_{(q\le\widehat{\gamma})}\right]:=E\left[\mathbf{x}\mathbf{x}'1_{(q\le\gamma)}\right]|_{\gamma=\widehat{\gamma}}$, and other expectation terms involving $\widehat{\gamma}$ are similarly defined. Obviously,

$$T_1+T_2\xrightarrow{d}M(\widetilde{\gamma})^{-1}W(\widetilde{\gamma})-\overline{M}(\widetilde{\gamma})^{-1}\overline{W}(\widetilde{\gamma})$$

in Corollary 1. The following Proposition 1 shows that

$$\widehat{\Sigma}(\widehat{\gamma}):=\widehat{M}(\widehat{\gamma})^{-1}\widehat{K}(\widehat{\gamma})\widehat{M}(\widehat{\gamma})^{-1}+\widehat{\overline{M}}(\widehat{\gamma})^{-1}\widehat{\overline{K}}(\widehat{\gamma})\widehat{\overline{M}}(\widehat{\gamma})^{-1}$$

can be used to estimate the asymptotic variance matrix of T_1+T_2, where

$$\begin{aligned}\widehat{M}(\gamma) &= \frac{1}{n}\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i'1_{(q_i\le\gamma)},\widehat{\overline{M}}(\gamma)=\frac{1}{n}\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i'1_{(q_i>\gamma)},\\ \widehat{K}(\gamma) &= \frac{1}{n}\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i'\widehat{\varepsilon}_i^21_{(q_i\le\gamma)},\widehat{\overline{K}}(\gamma)=\frac{1}{n}\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i'\widehat{\varepsilon}_i^21_{(q_i>\gamma)}.\end{aligned}$$

Proposition 1 *For any $b\in\mathbb{R}^k$, $\widehat{M}(\gamma)\xrightarrow{p}M(\gamma)$, $\widehat{\overline{M}}(\gamma)\xrightarrow{p}\overline{M}(\gamma)$, $\widehat{K}(\gamma)\xrightarrow{p}K(\gamma)$, and $\widehat{\overline{K}}(\gamma)\xrightarrow{p}\overline{K}(\gamma)$ uniformly in $\gamma\in\Gamma$, where $K(\gamma):=E\left[\mathbf{x}\mathbf{x}'\varepsilon^21_{(q\le\gamma)}\right]$, and $\overline{K}(\gamma):=E\left[\mathbf{x}\mathbf{x}'\varepsilon^21_{(q>\gamma)}\right]$.*

Denote the diagonal matrix by extracting the diagonal elements of $\widehat{\Sigma}(\widehat{\gamma})$ as $\mathrm{diag}\left\{\widehat{\sigma}_1^2,\cdots,\widehat{\sigma}_k^2\right\}$. Because $T_3=O_p(1)$, $\sqrt{n}\left|\widehat{\delta}_j\right|/\widehat{\sigma}_j=O_p(1)$. Picking κ_n such that $\kappa_n\to\infty$ as $n\to\infty$, we can guarantee that $\sqrt{n}\left|\widehat{\delta}_j\right|/\widehat{\sigma}_j\le\kappa_n$ with probability approaching one. In other words, if $\max_j\left\{\sqrt{n}\left|\widehat{\delta}_j\right|/\widehat{\sigma}_j\right\}>\kappa_n$, we are sure that $\max_j|b_j|\to\infty$ as $n\to\infty$ and the approximations in Hansen (2000) and Theorem 1 are appropriate. If $\sqrt{n}\left|\widehat{\delta}_j\right|/\widehat{\sigma}_j\le\kappa_n$, $\max_j|b_j|$ is either finite or diverges to infinite, but we can assume $b=\mathbf{0}$ (or find the least favorable b by simulation as in Section 5.3) and simulate the critical value as in the following Algorithm G. Some popular choices of κ_n are as follows. Hansen (2005) suggests $\sqrt{2\log\log n}$ by applying the law of the iterated logarithm (LIL), Andrews and Soares (2010) suggest $\sqrt{\ln n}$ by applying the BIC, and Chen et al. (2007) suggest $\ln n$. In our simulations of Section 7, we find $\sqrt{2\log\log n}$ seems too small while $\ln n$ seems too large, so we suggest $\sqrt{\ln n}$ in practice.[14]

[14] We can calibrate κ_n by bootstrapping from the original data with the true threshold $\widehat{\gamma}$ and choose κ_n to match the target coverage. But this is computationally too intensive, so is not pursued in this paper.

Algorithm G:

Step 1: Check whether $\max_j \left\{\sqrt{n}\left|\widehat{\delta}_j\right|/\widehat{\sigma}_j\right\} > \kappa_n$. If Yes, use the critical value in Table 2 to construct the CI.

Step 2: If $\max_j \left\{\sqrt{n}\left|\widehat{\delta}_j\right|/\widehat{\sigma}_j\right\} \leq \kappa_n$, simulate the least favorable critical value. For each $\gamma_0 \in \Gamma_n$, we simulate $\xi_1(\gamma_0) := \xi_1^0(\gamma_0) = \max_{\underline{\gamma}\leq\gamma\leq\gamma_0}\{\Xi(\gamma)\}$ and $\xi_2(\gamma_0) := \xi_2^0(\gamma_0) = \max_{\gamma_0\leq\gamma\leq\overline{\gamma}}\{\Xi(\gamma)\}$ and calculate

$$P\left(\max\left[\xi_1(\gamma_0),\xi_2(\gamma_0)\right] - \min\left[\xi_1(\gamma_0),\xi_2(\gamma_0)\right] \leq \sigma^2 x\right)$$

to find the critical value $\widehat{c}_{1-\alpha}(\gamma_0)$ at the α significance level. Then check whether $LR_n(\gamma_0) \leq \widehat{c}_{1-\alpha}(\gamma_0)$ to determine whether γ_0 should be included in the CI. The ultimate CI would be the convex hull of all accepted γ_0.

The simulation in Step 2 is similar to the grid bootstrap (see, e.g., Hansen (1999) and Mikusheva (2007) among others). In practice, we can search from $\underline{\gamma}$ and $\overline{\gamma}$; at each side, stop when some γ_0 is accepted. This can save simulation time because the least favorable distribution is employed so that the CI tends to be wide and consequently the algorithm would stop near $\underline{\gamma}$ and $\overline{\gamma}$.[15] This also implies that the undercoverage problem when $\sqrt{n}\delta_{20}$ is small in Table 3 can be alleviated.

The next question is how to simulate $\Xi(\gamma)$. To simulate $\Xi(\gamma)$, we need to simulate $W(\gamma)$ and$\overline{W}(\gamma)$ and estimate $M(\gamma)$ and $\overline{M}(\gamma)$. For this purpose, we need only have uniformly consistent estimators of $K(\gamma)$, $\overline{K}(\gamma)$, $M(\gamma)$ and $\overline{M}(\gamma)$ for all $\gamma \in \Gamma$. This have been done in Proposition 1. Given these estimators, we estimate $\Xi(\gamma)$ by

$$\widehat{\Xi}(\gamma) = \widehat{W}(\gamma)'\widehat{M}^{-1}(\gamma)\widehat{W}(\gamma) + \widehat{\overline{W}}(\gamma)'\widehat{\overline{M}}^{-1}(\gamma)\widehat{\overline{W}}(\gamma),$$

where $\widehat{W}(\gamma)$ is a Gaussian process with the covariance kernel equal to $\widehat{K}(\gamma_1 \wedge \gamma_2)$, $\widehat{\overline{W}}(\gamma)$ is a Gaussian process with the covariance kernel equal to $\widehat{\overline{K}}(\gamma_1 \vee \gamma_2)$.[16] In practice, we need only simulate $\widehat{W}(\gamma)$ and $\widehat{\overline{W}}(\gamma)$ at $\left\{q_{(1)},\cdots,q_{(m)}\right\}$, which is the set of sorted q_i's in the trimming set Γ_n; this is because for γ values between $q_{(i)}$ and $q_{(i+1)}$, $\widehat{K}(\gamma)$ and $\widehat{\overline{K}}(\gamma)$ do not change. Note also that in simulating $\widehat{W}(\gamma)$ and $\widehat{\overline{W}}(\gamma)$, we need $\widehat{K}(\gamma_2) - \widehat{K}(\gamma_1)$ to be positive definite for any $\gamma_1 < \gamma_2$. This is guaranteed only if $\gamma_2 - \gamma_1$ covers at least k $q_{(i)}$'s due to the finite-sample estimation of $K(\gamma)$.

In the homoskedastic case, the covariance kernel of $\widehat{W}(\gamma)$ and $\widehat{\overline{W}}(\gamma)$ can be replaced by $\widehat{M}(\gamma)$ and $\widehat{\overline{M}}(\gamma)$. Then calculate $P\left(\max\left[\xi_1(\gamma_0),\xi_2(\gamma_0)\right] - \min\left[\xi_1(\gamma_0),\xi_2(\gamma_0)\right] \leq x\right)$ to find the critical value $\widehat{c}_{1-\alpha}(\gamma_0)$ and check whether $LR_n(\gamma_0) \leq \widehat{c}_{1-\alpha}(\gamma_0)$ or check whether

$$P\left(\max\left[\xi_1(\gamma_0),\xi_2(\gamma_0)\right] - \min\left[\xi_1(\gamma_0),\xi_2(\gamma_0)\right] \leq LR_n(\gamma_0)\right) \leq 1-\alpha$$

to determine whether γ_0 should be included in the CI.

6.2 Confidence Interval for β

In the following Algorithm B, we discuss the confidence sets for β_ℓ and δ. By replacing the Wald statistic by the t statistic, we can similarly construct CIs for each element of β and δ.

Algorithm B:

[15] Here we assume $\max_j \left\{\sqrt{n}\left|\widehat{\delta}_j\right|/\widehat{\sigma}_j\right\} \leq \kappa_n$, i.e., $\|\delta_0\|$ is small. When $\max_j \left\{\sqrt{n}\left|\widehat{\delta}_j\right|/\widehat{\sigma}_j\right\} > \kappa_n$, a fixed critical value for all γ_0 (rather than a simulated critical value for each γ_0) is used so the CI for γ can be easily constructed.

[16] Note that $\widehat{W}(\gamma)$ cannot be replaced by $\widehat{K}^{1/2}(\gamma)B(\gamma)$ because the covariance kernels are different; similarly for $\widehat{\overline{W}}(\gamma)$.

Step 1: Check whether $\max_j \left\{ \left| \widehat{\delta}_j \right| / \widehat{\sigma}_j \right\} > \kappa_n / \sqrt{n}$. If No, construct a $(1 - \alpha_\gamma)$ CS for γ, say, $\widehat{\Gamma}(1 - \alpha_\gamma)$ as in the last section.

Step 2: For each $\gamma \in \widehat{\Gamma}(1 - \alpha_\gamma)$, construct a $(1 - \alpha_\ell)$ CS for β_ℓ, say, $\widehat{B}_\ell(1 - \alpha_\ell|\gamma)$ in the following way:

$$\left\{ \beta_\ell | n \left(\widehat{\beta}_\ell - \beta_\ell \right)' \widehat{V}_\ell(\gamma) \left(\widehat{\beta}_\ell - \beta_\ell \right) \leq c_{1-\alpha_\ell} \right\},$$

where $\widehat{V}_1(\gamma) = \widehat{M}^{-1}(\gamma) \widehat{K}(\gamma) \widehat{M}^{-1}(\gamma)$, $\widehat{V}_2(\gamma) = \widehat{\overline{M}}^{-1}(\gamma) \widehat{\overline{K}}(\gamma) \widehat{\overline{M}}^{-1}(\gamma)$, and $c_{1-\alpha_\ell}$ is the $(1 - \alpha_\ell)$ quantile of χ_k^2. The ultimate CS for β_ℓ is

$$\widehat{B}_{\ell,1-\alpha_\gamma}(1 - \alpha_\ell) = \bigcup_{\gamma \in \widehat{\Gamma}(1-\alpha_\gamma)} \widehat{B}_\ell(1 - \alpha_\ell|\gamma).$$

Because γ is the true value, $\lim_{n\to\infty} P\left(\beta_0 \in \widehat{B}(1 - \alpha_\ell|\gamma) \right) \geq 1 - \alpha_\ell$. So

$$\begin{aligned} \overline{\lim_{n\to\infty}} P\left(\beta_0 \notin \widehat{B}_{\ell,1-\alpha_\gamma}(1 - \alpha_\ell) \right) &\leq \overline{\lim_{n\to\infty}} P\left(\beta_0 \notin \widehat{B}_{\ell,1-\alpha_\gamma}(1 - \alpha_\ell), \gamma \in \widehat{\Gamma}(1 - \alpha_\gamma) \right) + \overline{\lim_{n\to\infty}} P\left(\gamma \notin \widehat{\Gamma}(1 - \alpha_\gamma) \right) \\ &\leq \alpha_\ell + \alpha_\gamma = \alpha. \end{aligned}$$

For δ, replace $\widehat{B}_\ell(1 - \alpha_\ell|\gamma)$ and $\widehat{B}_{\ell,1-\alpha_\gamma}(1 - \alpha_\ell)$ by

$$\begin{aligned} \widehat{B}(1 - \alpha_\delta|\gamma) &= \left\{ \delta | n \left(\widehat{\delta} - \delta \right)' \widehat{V}(\gamma) \left(\widehat{\delta} - \delta \right) \leq c_{1-\alpha_\delta} \right\}, \\ \widehat{B}_{1-\alpha_\gamma}(1 - \alpha_\delta) &= \bigcup_{\gamma \in \widehat{\Gamma}(1-\alpha_\gamma)} \widehat{B}(1 - \alpha_\delta|\gamma), \end{aligned}$$

where $\alpha_\delta = \alpha_\ell = \alpha - \alpha_\gamma$, and $\widehat{V}(\gamma) = \widehat{V}_1(\gamma) + \widehat{V}_2(\gamma)$.

Step 3: If Yes, then construct the $(1 - \alpha)$ CS for β_ℓ as $\widehat{B}_\ell(1 - \alpha|\widehat{\gamma})$ and the $(1 - \alpha)$ CS for δ as $\widehat{B}(1 - \alpha|\widehat{\gamma})$.

As for the selection of α_γ and α_ℓ, we can set $\alpha_\gamma = \alpha_\ell = \alpha/2$ in practice. Of course, we can also set $\alpha_\gamma \in [\underline{\alpha}, \overline{\alpha}]$ with $\underline{\alpha} > 0$, $\overline{\alpha} < \alpha$ and $\alpha_\ell = \alpha - \alpha_\gamma$, and then select α_γ by minimizing the length of the corresponding CI. As a result, for different elements of β and δ, α_γ need not be the same.

7 Simulations

In this section, we use the two DGPs in Section 4 to illustrate the performance of our new calibrated CIs. The replication time is 1000; in simulating $\widehat{\Xi}(\gamma)$, the simulation time is 999.

Table 5 reports the coverage and length of 90% calibrated LR-CI for γ. From Table 5, we can draw the following conclusions. First, when $\sqrt{n}\delta_{20}$ is small, the new CI has a better coverage (i.e., closer to 90%) than the interval coverage in Table 3, and when $\sqrt{n}\delta_{20}$ is large, the coverage is similar to the interval coverage in Table 3. Second, corresponding to the coverage, the new CI is generally longer the "old" CI when $\sqrt{n}\delta_{20}$ is small, and has a similar length when $\sqrt{n}\delta_{20}$ is large. Third, similar as in Table 3, the CI becomes shorter when δ_{20} gets larger for each n, but need not become shorter as n gets larger when δ_{20} is small. Fourth, we check $P(\kappa_n) := P\left(\max_j \left\{ \left| \widehat{\delta}_j \right| / \widehat{\sigma}_j \right\} \leq \kappa_n / \sqrt{n} \right)$ for $\kappa_n = \sqrt{\ln n}$. When $\sqrt{n}\delta_{20}$ is large, $P(\kappa_n)$ is close to 0. Anyway, when $\delta_{20} = 0$, $P(\kappa_n)$ is not close to 1 especially when n is small. As discussed in Section 4, when

n is small, a fake threshold may emerge in the noise of the data. In summary, we can safely claim that our new CI of γ performs better than the "old" CI.

		$z=q$				$z=x$			
$n \searrow$	$\delta_{20} \longrightarrow$	0	0.25	0.5	1	0	0.5	1	2
	50	0.828	0.849	0.868	0.922	0.829	0.839	0.880	0.933
Coverage	100	0.865	0.873	0.881	0.932	0.836	0.864	0.910	0.954
	500	0.883	0.894	0.911	0.950	0.870	0.890	0.936	0.950
	50	2.298	2.187	1.675	0.470	2.254	1.819	0.972	0.297
Length	100	2.632	2.336	1.289	0.135	2.515	1.554	0.471	0.155
	500	2.812	1.136	0.081	0.022	2.681	0.311	0.076	0.029
	50	0.636	0.630	0.578	0.386	0.365	0.189	0.034	0
$P(\kappa_n)$	100	0.691	0.684	0.612	0.207	0.470	0.122	0	0
	500	0.812	0.689	0.211	0	0.649	0	0	0

Table 5: Coverage and Length of 90% Calibrated LR-CI for γ

Note: $P(\kappa_n)$ is the probability of $\max_j \left\{ \left| \widehat{\delta}_j \right| / \widehat{\sigma}_j \right\} \leq \kappa_n / \sqrt{n}$

Table 6 reports the coverage and length of 95% calibrated CI for δ_2. From Table 6, we can draw the following conclusions. First, when $\sqrt{n}\delta_{20}$ is small, there is a little undercoverage, but the undercoverage is not serious as the usual t-CI. Second, when $\sqrt{n}\delta_{20}$ is large, the coverage is very close to 95%, and does not suffer the overcoverage as $\widehat{B}_{0.8}$. Third, like $\widehat{B}_0$ and $\widehat{B}_{0.8}$, the length of our CI decreases with n and δ_{20}. Fourth, when $\sqrt{n}\delta_{20}$ is small, our CI has a similar length as $\widehat{B}_{0.8}$, but when $\sqrt{n}\delta_{20}$ is large, our CI is shorter.

		$z=q$				$z=x$			
$n \searrow$	$\delta_{20} \longrightarrow$	0	0.25	0.5	1	0	0.5	1	2
	50	0.924	0.934	0.909	0.926	0.897	0.907	0.937	0.952
Coverage	100	0.943	0.945	0.911	0.942	0.926	0.935	0.962	0.948
	500	0.936	0.928	0.960	0.940	0.933	0.964	0.939	0.947
	50	16.92	15.25	13.02	4.394	5.153	4.446	2.751	1.212
Length	100	9.332	8.732	5.966	1.541	2.804	2.086	1.005	0.814
	500	3.176	1.972	0.654	0.588	1.102	0.379	0.354	0.352

Table 6: Coverage and Length of 95% Calibrated CI for δ_2

8 Empirical Applications

In this section, we apply our calibrating method to two datasets in economics to check its performance. The first dataset is the growth data used in Durlauf and Johnson (1995) and reanalyzed in Hansen (2000). The second dataset is the tipping points data used in Pan (2015). As in Hansen (2000), we will concentrate on the CI for γ.

8.1 Growth and Multiple Equilibria

In this application, the concern is whether there is a threshold effect in the GDP growth. The growth theory with multiple equilibria motivates the following threshold regression model:

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

For each country i, $\left(\frac{Y}{L}\right)_{i,t}$ is the real GDP per member of the population aged 15-64 in year t, $\left(\frac{I}{Y}\right)_i$ 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 $g + \delta = 0.05$. As suggested in Hansen (2000), we will check two possible threshold variables; the first one is $\ln\left(\frac{Y}{L}\right)_{i,1960}$ and the second one is the adult literacy rate in 1960, $LR_{i,1960}$. Also following Hansen (2000), we will consider only the heteroskedastic-consistent procedures.

In the first stage, we use $\ln\left(\frac{Y}{L}\right)_{i,1960}$ as q_i. The middle-point LSE of γ is \$871, which generates the same sample splitting as the left-endpoint LSE \$863. Since $\max_j \left\{\sqrt{n}\left|\widehat{\delta}_j\right|/\widehat{\sigma}_j\right\} = 2.563 > 2.136 = \sqrt{\ln n}$, we are employing the critical value in Table 1 to construct the CI for γ. The 95% critical value is 5.991, much smaller than 7.352 as suggested in Hansen (2000). If we use the kernel method to estimate η^2, then both Hansen's critical value and ours generate the same CI for γ, $[\$594, \$1842)$. The left graph of Figure 7 illustrates these two CIs. Obviously, this CI is quite wide since it covers 40 out of the 96 countries in the sample.

In the second stage, we use $LR_{i,1960}$ as q_i and apply threshold regression to the right regime in the first stage defined by $\ln\left(\frac{Y}{L}\right)_{i,1960} > 871$. The middle-point LSE of γ is 47%. Since $\max_j \left\{\sqrt{n}\left|\widehat{\delta}_j\right|/\widehat{\sigma}_j\right\} = 2.570 > 2.087 = \sqrt{\ln n}$, we will use the critical value in Table 1 to construct the CI for γ. It turns out that Hansen's CI is $[19\%, 61\%)$, while ours is $[29\%, 61\%)$, which is much shorter. Hansen's CI covers 19 while ours covers 14 out of the 78 countries in the subsample. The right graph of Figure 7 illustrates these two CIs.

This application shows that our CI is generally shorter than Hansen's CI. As commented in Hansen (2000), his CI is sufficiently large that there is considerable uncertainty regarding the threshold value. This paper shows that his CI can be much shortened by using an appropriate critical value.

8.2 Tipping Points in Dynamic Segregation

Since the seminal work by Schelling (1971), dynamic segregation models have been intensively studied in the literature. For instance, Card et al. (2008) estimate a dynamic segregation model of neighborhood racial composition between 1970 and 2000; Pan (2015) investigates how tipping points impact the dynamics of occupational gender segregation in the labor market between 1940 and 1990.

Pan (2015) specifies the following TR model to study the dynamic segregation,

$$Dm_{isrj,t} = p\left(f_{isrj,t-10}\right) + d1(f_{isrj,t-10} > f^*_{rj,t-10}) + X'_{isrj,t-10}\beta + \varepsilon_{isrj,t}, \tag{7}$$

where $m_{isrj,t}$ and $f_{isrj,t-10}$ are the shares of male and female employment in occupation i, state s, region r, and the group of white-collar or blue-collar occupations j in year t or $t-10$, respectively. The dependent variable $Dm_{isrj,t}$ is the net change in male employment growth, defined as the difference between male and female employment growth rate between year t and $t-10$. The $f^*_{rj,t-10}$ represents the tipping

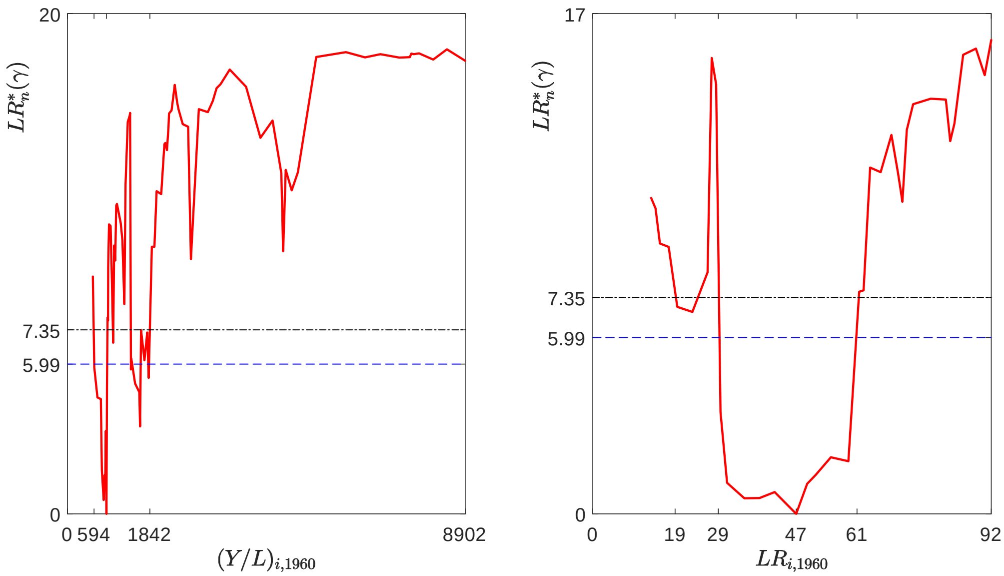

Figure 7: Confidence Interval Construction for γ in Two Stages

point at the region r and the white-collar or blue-collar level j; $p(\cdot)$ is a fourth-order polynomial function; $X_{isrj,t-10}$ includes white-collar region fixed effects, occupation characteristics (average age, education, and log male wages) in the initial period, and one-digit occupation fixed effects; $\varepsilon_{isrj,t}$ is the error term. Pan assumes that there is only threshold effect in the intercept, so many formulas can be simplified, e.g., $\eta^2 = E\left[\varepsilon_i^2 | q_i = \gamma_0\right] / \sigma^2$ because $\delta_n' D \delta_n = d^2$ and $\delta_n' V \delta_n = d^2 E\left[\varepsilon_i^2 | q_i = \gamma_0\right]$. For simplicity, we just assume the model is homoskedastic, i.e., $\eta^2 = 1$; the heteroskedastic case shares quantitatively similar features.

Table 7 contains the γ estimates and the associated 95% CIs for each decadal period from 1940 to 1990. First check the estimate of d. The $\widehat{d}$ values for all periods are negative, which indicates that there is indeed occupational gender segregation between 1940 and 1990. The magnitude and the time trend (first increase and then decrease) of the decline are mostly consistent with the results in Table 3 of Pan (2015). Because $\left|\widehat{d}\right| / se > \kappa_n$ for all periods, we just use the modified critical values in Table 1 to construct the CI for $f_{rj,t-10}^*$. In general, our CIs are shorter than Hansen's CI, and in some period, e.g., 1970-1980, our CI covers much less $f_{isrj,t-10}$'s than Hansen's CI. Figure 8 intuitively illustrates the CI construction for $f_{rj,t-10}^*$ in these five periods.

Time Period	1940-1950	1950-1960	1960-1970	1970-1980	1980-1990
$\widehat{f_{rj,t-10}^*}$	0.275	0.404	0.314	0.201	0.312
Hansen's CI	$[0.255, 0.305)$	$[0.363, 0.421)$	$[0.291, 0.352)$	$[0.140, 0.233)$	$[0.295, 0.331)$
Our CI	$[0.255, 0.300)$	$[0.363, 0.421)$	$[0.291, 0.355)$	$[0.174, 0.232)$	$[0.297, 0.331)$
$\#\|\#$	14|13	17|17	14|13	114|58	97|91
$\widehat{d}$	-0.393	-0.439	-0.497	-0.224	-0.202
se	0.0666	0.0682	0.0653	0.0351	0.0187
$\left\|\widehat{d}\right\| / se$	5.898	6.432	7.607	6.374	10.827
κ_n	2.668	2.680	2.707	2.759	2.916

Table 7: Estimates of $f_{rj,t-10}^*$ and the 95% CI: $\kappa_n = \sqrt{\ln n}$, se is the standard error of $\widehat{d}$, $\#|\#$ is the number of data points covered by Hansen's CI and our CI

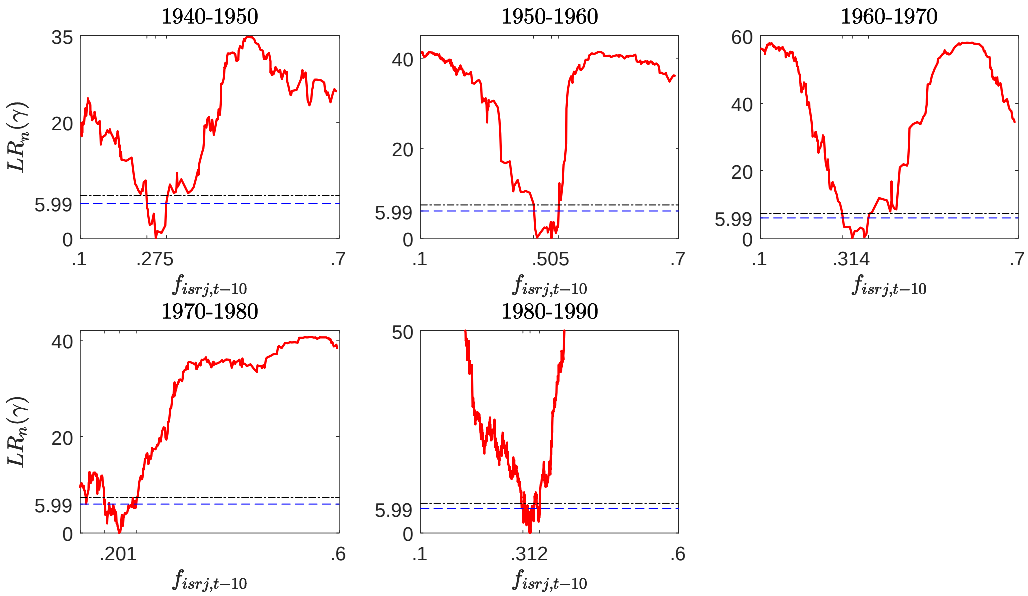

Figure 8: Confidence Interval Construction for $f^*_{rj,t-10}$ in Five Periods

In both applications, $\max_j \left\{ \sqrt{n} \left| \widehat{\delta}_j \right| / \widehat{\sigma}_j \right\} > \kappa_n$; this is because the usual specification test such as Hansen (1996) strongly rejects the null that there is no threshold effect. This is the typical case in practice because we usually consider the inference of γ only after we got a strong rejection in the specification testing of no threshold effects. When the economic theory indicates that there should be threshold effects, but $\max_j \left\{ \sqrt{n} \left| \widehat{\delta}_j \right| / \widehat{\sigma}_j \right\} \leq \kappa_n$, then we may need to employ the simulation method in Algorithm G. The discussion here also indicates that we can determine whether $\max_j |b_j| \to \infty$ as $n \to \infty$ by checking whether the p-value in the specification test shrinks to zero (see, e.g., Proposition 8 of Bai and Perron (1998)), but the rate of shrinking to zero seems hard to calibrate to achieve good performances.

9 Conclusion

In this paper, we calibrate the CIs in Hansen (2000) to achieve better coverage and/or length. Basically, we construct CIs for γ and β based on whether the threshold effect is strong or weak. For the threshold point γ, when the threshold effect is strong, we suggest smaller critical values to obtain shorter LR-CIs with less overcoverage; when the threshold effect is weak, we suggest a least favorable distribution for the LR statistic such that the undercoverage problem of the usual LR-CI can be eliminated or mitigated. For the slope parameter β, when the threshold effect is strong, the usual t-CI works well; when the threshold effect is weak, we rigorize the Bonferroni-type CI in Hansen (2000). Our simulations and empirical applications show that our new CIs indeed have better performances than those in Hansen (2000).

References

Andrews, D.W.K., 1993, Tests for Parameter Instability and Structural Change with Unknown Change Point, *Econometrica*, 61, 821-856.

Andrews, D.W.K. and X. Cheng, 2012, Estimation and Inference with Weak, Semi-Strong and Strong Identification, *Econometrica*, 80, 2153-2211.

Andrews, D.W.K. and G. Soares, 2010, Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection, *Econometrica*, 78, 119-157.

Andrews, I., T. Kitagawa, and A. McCloskey, 2019, Inference on Winners, mimeo.

Andrews, I., T. Kitagawa, and A. McCloskey, 2021, Inference after Estimation of Breaks, *Journal of Econometrics*, 224, 39-59.

Bai, J., 1997, Estimation of a Change Point in Multiple Regression Models, *Review of Economics and Statistics*, 79, 551-563.

Bai, J. and P. Perron, 1998, Estimating and Testing Linear Models with Multiple Structural Changes, *Econometrica*, 66, 47-78.

Bai, J. and P. Perron, 2006, Multiple Structural Change Models: A Simulation Analysis, *Econometric Theory and Practice: Frontiers of Analysis and Applied Research*, P.D. Corbae, S.N. Durlauf and B.E. Hansen, eds., New York: Cambridge University Press.

Bhattacharya, P.K. and P.J. Brockwell, 1976, The Minimum of an Additive Process with Applications to Signal Estimation and Storage Theory, *Z. Wahrschein. Verw. Gebiete*, 37, 51-75.

Card, D., A. Mas and J. Rothstein, 2008, Tipping and The Dynamics of Segregation, *Quarterly Journal of Economics*, 123, 177-218.

Chan, K.S. and H. Tong, 1990, On Likelihood Ratio Tests for Threshold Autoregression, *Journal of the Royal Statistical Society*, *Series B*, 52, 469-476.

Chan, K.S. and R.S. Tsay, 1998, Limiting Properties of the Least Squares Estimator of a Continuous Threshold Autoregressive Model, *Biometrika*, 85, 413-426.

Chen, X., H. Hong and M. Shum, 2007, Nonparametric Likelihood Ratio Model Selection Tests Between Parametric Likelihood and Moment Condition Models, *Journal of Econometrics*, 141, 109-140.

Dufour, J. M., 1997, Some Impossibility Theorems in Econometrics with Applications to Structural Variables and Dynamic Models, *Econometrica*, 65, 1365-1387.

Durlauf, S.N. and P.A. Johnson, 1995, Multiple Regimes and Cross-Country Growth Behavior, Journal of Applied Econometrics, 10, 365-384.

Elliott, G. and U.K. Müller, 2007, Confidence Sets for the Date of a Single Break in Linear Time Series Regressions, *Journal of Econometrics*, 141, 1196-1218.

Elliott, G. and U.K. Müller, 2014, Pre and Post Break Parameter Inference, *Journal of Econometrics*, 180, 141-157.

Elliott, G., U.K. Müller and M.W. Watson, 2015, Nearly Optimal Tests When A Nuisance Parameter Is Present Under the Null Hypothesis, *Econometrica*, 83, 771-811.

Eo, Y. and J. Morley, 2015, Likelihood-Ratio-Based Confidence Sets for the Timing of Structural Breaks, *Quantitative Economics*, 6, 463-497.

Hansen, B.E., 1996, Inference when a Nuisance Parameter is not Identified under the Null Hypothesis, *Econometrica*, 413-430.

Hansen, B.E., 1999, The Grid Bootstrap and the Autoregressive Model, *Review of Economics and Statistics*, 81, 594-607.

Hansen, B.E., 2000, Sample Splitting and Threshold Estimation, *Econometrica*, 575-603.

Hansen, B.E., 2011, Threshold Autoregression in Economics, *Statistics and Its Interface*, 4, 123-127.

Hansen, B.E., 2017, Regression Kink With an Unknown Threshold, *Journal of Business & Economics Statistics*, 35, 228-240.

Hansen, P.R., 2005, A Test for Superior Predictive Ability, *Journal of Business & Economics Statistics*, 23, 365-380.

Kim, J. and D. Pollard, 1990, Cube Root Asymptotics, *Annals of Statistics*, 18, 191-219.

Lee, Y. and Y. Wang, 2022, Testing for Homogeneous Thresholds in Threshold Regression Models, *Econometric Theory*, forthcoming.

Jiang, L., X. Wang and J. Yu, 2018, New Distribution Theory for the Estimation of Structural Break Point in Mean, *Journal of Econometrics*, 205, 156-176.

McCloskey, A., 2017, Bonferroni-Based Size-Correction for Nonstandard Testing Problems, *ournal of Econometrics*, 200, 17-35.

Mikusheva, A., 2007, Uniform Inference in Autoregressive Models, *Econometrica*, 1411-1452.

Pan, J., 2015, Gender Segregation in Occupations: The Role of Tipping and Social Interactions, *Journal of Labor Economics*, 33, 365-408.

Picard, D., 1985, Testing and Estimating Change-Points in Time Series, *Advances in Applied Probability*, 17, 841-867.

Schelling, T.C, 1971, Dynamic Models of Segregation, *Journal of Mathematical Sociology*, 1, 143-186.

Siegmund, D., 1986, Boundary Crossing Probabilities and Statistic Applications, *Annals of Statistics*, 14, 361-404.

Siegmund, D., 1988, Confidence Set in Change-Point Problems, *International Statistical Review*, 56, 31-48.

Staiger, D. and J.H. Stock, 1997, Instrumental Variables Regression with Weak Instruments, *Econometirca*, 65, 557-586.

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

Yu, P., 2012, Likelihood Estimation and Inference in Threshold Regression, *Journal of Econometrics*, 2012, 167, 274-294.

Yu, P., 2015, Adaptive Estimation of the Threshold Point in Threshold Regression, *Journal of Econometrics*, 189, 83-100.

Appendix A: Proofs

We here collect some notations for future reference. $a_n = n^{1-2\varphi}$. $\rightsquigarrow$ signifies weak convergence over a metric space. $u = (u_1', u_2')$ and v are local parameters for β and γ.

Proof of Theorem 1. Note that when $\gamma = \gamma_0 + a_n^{-1} v$, the true γ value is still γ_0. Re-write $LR_n(\gamma)$ as

$$LR_n(\gamma) = \frac{\left[S_n\left(\gamma, \widehat{\beta}(\gamma)\right) - S_n(\gamma_0, \beta_0)\right] - \left[S_n\left(\widehat{\gamma}, \widehat{\beta}\right) - S_n(\gamma_0, \beta_0)\right]}{\left[S_n\left(\widehat{\gamma}, \widehat{\beta}\right) - S_n(\gamma_0, \beta_0)\right]/n + S_n(\gamma_0, \beta_0)/n}.$$

Because $S_n\left(\gamma_0 + a_n^{-1}v, \beta_0 + n^{-1/2}u\right) - S_n(\gamma_0, \beta_0) \rightsquigarrow u_1' E\left[\mathbf{xx}' 1_{(q \le \gamma_0)}\right] u_1 + u_2' E\left[\mathbf{xx}' 1_{(q > \gamma_0)}\right] u_2 - 2u_1' W_1 - 2u_2' W_2 + \mu |v| + 2\sqrt{\lambda} W(v)$, where $\mu = c' D c f$ and $\lambda = c' V c f$, we have

$$S_n\left(\gamma, \widehat{\beta}(\gamma)\right) - S_n(\gamma_0, \beta_0) \rightsquigarrow -W_1' E\left[\mathbf{xx}' 1_{(q \le \gamma_0)}\right]^{-1} W_1 - W_2' E\left[\mathbf{xx}' 1_{(q > \gamma_0)}\right]^{-1} W_2 + \mu |v| + 2\sqrt{\lambda} W(v),$$

by taking minimum with respect to u, and

$$S_n\left(\widehat{\gamma}, \widehat{\beta}\right) - S_n(\gamma_0, \beta_0) \rightsquigarrow -W_1' E\left[\mathbf{xx}' 1_{(q \le \gamma_0)}\right]^{-1} W_1 - W_2' E\left[\mathbf{xx}' 1_{(q > \gamma_0)}\right]^{-1} W_2 + \min_v \left\{\mu |v| + 2\sqrt{\lambda} W(v)\right\}$$

by taking minimum with respect to both u and v. As as result,

$$\begin{aligned} LR_n\left(\gamma_0 + a_n^{-1} v\right) \quad \rightsquigarrow \quad & LR_\infty(v) := \frac{1}{\sigma^2}\left[\mu |v| + 2\sqrt{\lambda} W(v) - \min_v \left\{\mu |v| + 2\sqrt{\lambda} W(v)\right\}\right] \\ & \overset{d}{=} \frac{1}{\sigma^2}\left[\mu |v| + 2\sqrt{\lambda} W(v) + \sup_v \left\{-\mu |v| + 2\sqrt{\lambda} W(v)\right\}\right]. \end{aligned}$$

The interval coverage converges to

$$P\left(\inf_{v \le 0} LR_\infty(v) \le \eta^2 c_\alpha \text{ and } \inf_{v \ge 0} LR_\infty(v) \le \eta^2 c_\alpha\right),$$

where c_α is the upper αth quantile of $\sup_v \{-|v| + 2W(v)\}$. Since

$$\begin{aligned} \inf_{v \le 0} LR_\infty(v) &= \frac{1}{\sigma^2}\left[\inf_{v \le 0}\left\{\mu \left|\frac{\lambda}{\mu^2} v\right| + 2\sqrt{\lambda} W\left(\frac{\lambda}{\mu^2} v\right)\right\} - \inf_v \left\{\mu \left|\frac{\lambda}{\mu^2} v\right| + 2\sqrt{\lambda} W\left(\frac{\lambda}{\mu^2} v\right)\right\}\right] \\ &= \frac{\lambda}{\sigma^2 \mu}\left[\inf_{v \le 0}\{|v| + 2W(v)\} - \inf_v \{|v| + 2W(v)\}\right] \end{aligned}$$

and similarly, $\inf\limits_{v\geq 0} LR_\infty(v) = \eta^2\left[\inf\limits_{v\geq 0}\{|v|+2W(v)\} - \inf_v\{|v|+2W(v)\}\right]$, we have

$$\begin{aligned}
p(c_\alpha) &= P\left(\inf_{v\leq 0}\{|v|+2W(v)\}+\xi\leq c_\alpha \text{ and } \inf_{v\geq 0}\{|v|+2W(v)\}+\xi\leq c_\alpha\right)\\
&= P\left(-2\xi_1+2\max[\xi_1,\xi_2]\leq c_\alpha \text{ and } -2\xi_2+2\max[\xi_1,\xi_2]\leq c_\alpha\right)\\
&= P\left(\max[\xi_1,\xi_2]-\min[\xi_1,\xi_2]\leq\frac{c_\alpha}{2}\right)\\
&= P\left(\xi_1-\xi_2\leq\frac{c_\alpha}{2}\text{ and }\xi_1>\xi_2\right)+P\left(\xi_2-\xi_1\leq\frac{c_\alpha}{2}\text{ and }\xi_2>\xi_1\right)\\
&= 2\int_0^\infty\left[\left(1-e^{-\frac{c_\alpha}{2}-\xi_2}\right)-\left(1-e^{-\xi_2}\right)\right]e^{-\xi_2}d\xi_2\\
&= 2\int_0^\infty\left(e^{-\xi_2}-e^{-\frac{c_\alpha}{2}-\xi_2}\right)e^{-\xi_2}d\xi_2\\
&= 2\left(1-e^{-\frac{c_\alpha}{2}}\right)\int_0^\infty e^{-2\xi_2}d\xi_2 = 1-e^{-\frac{c_\alpha}{2}},
\end{aligned}$$

where $\xi = -\inf_v\{|v|+2W(v)\} \overset{d}{=} \sup_v\{-|v|+2W(v)\}$, $\xi_1 := \sup_{v\leq 0}\left\{-\frac{|v|}{2}+W(v)\right\}\sim Exp(1)$ and $\xi_2 := \sup_{v\geq 0}\left\{-\frac{|v|}{2}+W(v)\right\}\sim Exp(1)$ are independent. ■

Proof of Theorem 2. The $\sqrt{n}$-consistency of $\widehat{\beta}$ is proved in Lemma 1. Given that $\left(\widehat{\gamma},\sqrt{n}\left(\widehat{\beta}-\beta_n\right)\right) = \arg\min_{(\gamma,u)}\left\{S_n\left(\gamma,\beta_n+\frac{u}{\sqrt{n}}\right)-S_n(\beta_n,\gamma_n)\right\}$, we can apply an argmax continuous mapping theorem such as Theorem 3.2.2 in van der Vaart and Wellner (1996) to find the asymptotic distribution of $\left(\widehat{\gamma},\sqrt{n}\left(\widehat{\beta}-\beta_n,\right)\right)$. For this purpose, we need only show $S_n\left(\gamma,\beta_n+\frac{u}{\sqrt{n}}\right)-S_n(\beta_n,\gamma_n)\rightsquigarrow -\Xi(h)$ on every $U\times\Gamma$, where U is any compact subset in $\mathbb{R}^{2k}$, $\Xi(h)$ is continuous and possess a unique maximum at $\widehat{h}$ with $\widehat{h}=O_p(1)$. The weak convergence is proved in Lemma 2, the continuity of $\Xi(h)$ is implied by the continuity of its covariance kernel and Assumption D.5$'$, and the uniqueness of $\arg\max_h\Xi(h)$ is implied by Lemma 2.6 of Kim and Pollard (1990). As to $\widehat{h}=O_p(1)$, note that $\widehat{\gamma}\in\Gamma$ is $O_p(1)$,

$$\widehat{u}_1(\gamma)=\begin{cases} M(\gamma)^{-1}W(\gamma)\text{, if }\gamma\leq\gamma_0,\\ M(\gamma)^{-1}\left(-M(\gamma_0,\gamma)b+W(\gamma_0)+W(\gamma_0,\gamma)\right)\text{, if }\gamma>\gamma_0,\end{cases}$$

and

$$\widehat{u}_2(\gamma)=\begin{cases} \overline{M}(\gamma)^{-1}\left(M(\gamma,\gamma_0)b+\overline{W}(\gamma_0)+W(\gamma,\gamma_0)\right)\text{, if }\gamma\leq\gamma_0,\\ \overline{M}(\gamma)^{-1}\overline{W}(\gamma)\text{, if }\gamma>\gamma_0,\end{cases}$$

both satisfying $\sup_{\gamma\in\Gamma}\widehat{u}_\ell(\gamma)=O_p(1)$. ■

Proof of Corollary 1. For $\gamma\leq\gamma_0$, note that

$$\begin{aligned}
\Xi(h) &= -\left(u_1-M(\gamma)^{-1}W(\gamma)\right)'M(\gamma)\left(u_1-M(\gamma)^{-1}W(\gamma)\right)+W(\gamma)'M(\gamma)^{-1}W(\gamma)\\
&\quad-\left(u_2-\overline{M}(\gamma)^{-1}\left(M(\gamma,\gamma_0)b+\overline{W}(\gamma_0)+W(\gamma,\gamma_0)\right)\right)'\overline{M}(\gamma)(\cdots)\\
&\quad+\left(M(\gamma,\gamma_0)b+\overline{W}(\gamma_0)+W(\gamma,\gamma_0)\right)'\overline{M}(\gamma)^{-1}(\cdots)-b'M(\gamma,\gamma_0)b-2b'W(\gamma,\gamma_0),
\end{aligned}$$

so the concentrated $\Xi(h)$ is equal to $\Xi(\widehat{u}_1(\gamma),\widehat{u}_2(\gamma),\gamma)=\Xi(\gamma|\gamma_0,b)$ in the corollary. Simiar argument applies to the case where $\gamma>\gamma_0$.

Given the distribution of $\arg\max_{\gamma} \Xi(\gamma|\gamma_0, b)$ as $G(\gamma)$, the distribution of $\widehat{u}_1$ is the same as

$$\int_{\underline{\gamma}}^{\gamma_0} M(\gamma)^{-1} W(\gamma)\, dG(\gamma) + \int_{\gamma_0}^{\overline{\gamma}} M(\gamma)^{-1} \left(-M(\gamma_0, \gamma)\, b + W(\gamma)\right) dG(\gamma),$$

and the distribution of $\widehat{u}_2$ is the same as

$$\int_{\underline{\gamma}}^{\gamma_0} \overline{M}(\gamma)^{-1} \left(M(\gamma, \gamma_0)\, b + \overline{W}(\gamma)\right) dG(\gamma) + \int_{\gamma_0}^{\overline{\gamma}} \overline{M}(\gamma)^{-1} \overline{W}(\gamma)\, dG(\gamma).$$

The asymptotic distribution of $\sqrt{n}\widehat{\delta} - b$ is $\widehat{u}_1 - \widehat{u}_2$.

Finally,

$$LR_n(\gamma_n) = \frac{\left[S_n\left(\gamma_n, \widehat{\beta}(\gamma_n)\right) - S_n(\gamma_n, \beta_n)\right] - \left[S_n\left(\widehat{\gamma}, \widehat{\beta}\right) - S_n(\gamma_n, \beta_n)\right]}{\left[S_n\left(\widehat{\gamma}, \widehat{\beta}\right) - S_n(\gamma_n, \beta_n)\right]/n + S_n(\gamma_n, \beta_n)/n}.$$

From the proof in Lemma 2, we can see

$$\begin{aligned}
& S_n\left(\gamma_n, \widehat{\beta}(\gamma_n)\right) - S_n(\gamma_n, \beta_n) \xrightarrow{d} \min_{u_1, u_2} \left\{u_1' M(\gamma_0)\, u_1 + u_2' \overline{M}(\gamma_0)\, u_2 - 2u_1' W(\gamma_0) - 2u_2' \overline{W}(\gamma_0)\right\} \\
= \quad & -W(\gamma_0)' M(\gamma_0)^{-1} W_1(\gamma_0) - \overline{W}(\gamma_0)' \overline{M}(\gamma_0)^{-1} \overline{W}(\gamma_0) =: -\Xi(\gamma_0),
\end{aligned}$$

and

$$\left[S_n\left(\widehat{\gamma}, \widehat{\beta}\right) - S_n(\gamma_n, \beta_n)\right] \xrightarrow{d} \min_{\gamma \in \Gamma} \left\{-\Xi(\gamma|\gamma_0, b)\right\} = O_p(1).$$

So

$$LR_n(\gamma_n) \xrightarrow{d} \frac{-\Xi(\gamma_0) - \min_{\gamma \in \Gamma}\{-\Xi(\gamma|\gamma_0, b)\}}{\sigma^2} = \frac{\max_{\gamma \in \Gamma} \Xi(\gamma|\gamma_0, b) - \Xi(\gamma_0)}{\sigma^2}.$$

This is different from the asymptotic distribution of

$$\begin{aligned}
LR_{1n}(\gamma_n) &= \frac{S_n\left(\gamma_n, \widehat{\beta}\right) - S_n\left(\widehat{\gamma}, \widehat{\beta}\right)}{S_n\left(\widehat{\gamma}, \widehat{\beta}\right)/n} \\
&= \frac{\left[S_n\left(\gamma_n, \widehat{\beta}\right) - S_n(\gamma_n, \beta_n)\right] - \left[S_n\left(\widehat{\gamma}, \widehat{\beta}\right) - S_n(\gamma_n, \beta_n)\right]}{\left[S_n\left(\widehat{\gamma}, \widehat{\beta}\right) - S_n(\gamma_n, \beta_n)\right]/n + S_n(\gamma_n, \beta_n)/n}
\end{aligned}$$

because

$$\begin{aligned}
& S_n\left(\gamma_n, \widehat{\beta}\right) - S_n(\gamma_n, \beta_n) \xrightarrow{d} \widehat{u}_1' M(\gamma_0)\, \widehat{u}_1 + \widehat{u}_2' \overline{M}(\gamma_0)\, \widehat{u}_2 - 2\widehat{u}_1' W(\gamma_0) - 2\widehat{u}_2' \overline{W}(\gamma_0) \\
\neq \quad & -W(\gamma_0)' M(\gamma_0)^{-1} W_1(\gamma_0) - \overline{W}(\gamma_0)' \overline{M}(\gamma_0)^{-1} \overline{W}(\gamma_0) = -\Xi(\gamma_0),
\end{aligned}$$

given that $\widehat{u}_1 \neq M(\gamma_0)^{-1} W_1(\gamma_0) =: \widetilde{u}_1$ and $\widehat{u}_2 \neq \overline{M}(\gamma_0)^{-1} \overline{W}(\gamma_0) =: \widetilde{u}_2$. This is also different from the asymptotic distribution of

$$\begin{aligned} LR_{2n}(\gamma_n) &= \frac{S_n\left(\gamma_n, \widehat{\beta}(\gamma_n)\right) - S_n\left(\widehat{\gamma}\left(\widehat{\beta}(\gamma_n)\right), \widehat{\beta}(\gamma_n)\right)}{S_n\left(\widehat{\gamma}, \widehat{\beta}\right)/n} \\ &= \frac{\left[S_n\left(\gamma_n, \widehat{\beta}(\gamma_n)\right) - S_n(\gamma_n, \beta_n)\right] - \left[S_n\left(\widehat{\gamma}\left(\widehat{\beta}(\gamma_n)\right), \widehat{\beta}(\gamma_n)\right) - S_n(\gamma_n, \beta_n)\right]}{\left[S_n\left(\widehat{\gamma}, \widehat{\beta}\right) - S_n(\gamma_n, \beta_n)\right]/n + S_n(\gamma_n, \beta_n)/n} \end{aligned}$$

because

$$S_n\left(\widehat{\gamma}\left(\widehat{\beta}(\gamma_n)\right), \widehat{\beta}(\gamma_n)\right) - S_n(\gamma_n, \beta_n) \xrightarrow{d} \min_{\gamma\in\Gamma}\left\{-\Xi(\gamma, \widetilde{u}_1, \widetilde{u}_2)\right\} \neq \min_{\gamma\in\Gamma}\left\{-\Xi(\widehat{u}_1(\gamma), \widehat{u}_2(\gamma), \gamma)\right\} = \min_{\gamma\in\Gamma}\left\{-\Xi(\gamma|\gamma_0, b)\right\}.$$

■

Proof of Theorem 3. Note that when $\gamma \in \Gamma$, the true γ value is still γ_0. As in the proof of Theorem 1,

$$LR_n(\gamma) = \frac{\left[S_n\left(\widehat{\beta}(\gamma), \gamma\right) - S_n(\beta_0, \gamma_0)\right] - \left[S_n\left(\widehat{\beta}, \widehat{\gamma}\right) - S_n(\beta_0, \gamma_0)\right]}{\left[S_n\left(\widehat{\beta}, \widehat{\gamma}\right) - S_n(\beta_0, \gamma_0)\right]/n + S_n(\beta_0, \gamma_0)/n}.$$

Because

$$S_n\left(\beta_0 + n^{-1/2}u, \gamma\right) - S_n(\beta_0, \gamma_0) \rightsquigarrow -\Xi(h|\gamma_0, b),$$

we have

$$S_n\left(\widehat{\beta}(\gamma), \gamma\right) - S_n(\beta_0, \gamma_0) \rightsquigarrow -\Xi(\gamma|\gamma_0, b)$$

and

$$S_n\left(\widehat{\beta}, \widehat{\gamma}\right) - S_n(\beta_0, \gamma_0) \rightsquigarrow \min_{\gamma\in\Gamma}\left\{-\Xi(\gamma|\gamma_0, b)\right\} = -\max_{\gamma\in\Gamma}\Xi(\gamma|\gamma_0, b).$$

As a result,

$$LR_n(\gamma) \rightsquigarrow LR_\infty(\gamma|b) = \frac{\max_{\gamma\in\Gamma}\Xi(\gamma|\gamma_0, b) - \Xi(\gamma|\gamma_0, b)}{\sigma^2}.$$

The interval coverage converges to

$$\begin{aligned} & P\left(\inf_{\underline{\gamma}\le\gamma\le\gamma_0} LR_\infty(\gamma|b) \le c_\alpha \text{ and } \inf_{\gamma_0\le\gamma\le\overline{\gamma}} LR_\infty(\gamma|b) \le c_\alpha\right) \\ = \; & P\left(\max_{\underline{\gamma}\le\gamma\le\overline{\gamma}}\left\{\Xi(\gamma|\gamma_0, b)\right\} - \min\left\{\max_{\underline{\gamma}\le\gamma\le\gamma_0}\left\{\Xi(\gamma|\gamma_0, b)\right\}, \max_{\gamma_0\le\gamma\le\overline{\gamma}}\left\{\Xi(\gamma|\gamma_0, b)\right\}\right\} \le \sigma^2 c_\alpha\right) \\ = \; & P\left(\max\left[\xi_1^b, \xi_2^b\right] - \min\left[\xi_1^b, \xi_2^b\right] \le \sigma^2 c_\alpha\right), \end{aligned}$$

where $\xi_1^b = \max_{\underline{\gamma}\le\gamma\le\gamma_0}\left\{\Xi(\gamma|\gamma_0, b)\right\}$ and $\xi_1^b = \max_{\gamma_0\le\gamma\le\overline{\gamma}}\left\{\Xi(\gamma|\gamma_0, b)\right\}$. ■

Proof of Proposition 1. We take $\widehat{K}(\gamma)$ as an example because the proof for $\widehat{\widehat{K}}(\gamma)$ is similar and for $\widehat{M}(\gamma)$ and $\widehat{\widehat{M}}(\gamma)$ is easier.

WLOG, assume $\widehat{\gamma} < \gamma_0$; then

$$\begin{aligned} \widehat{\varepsilon}_i &= y_i - \mathbf{x}_i'\widehat{\beta}_1 1_{(q_i\le\widehat{\gamma})} - \mathbf{x}_i'\widehat{\beta}_2 1_{(q_i>\widehat{\gamma})} \\ &= \varepsilon_i - \mathbf{x}_i'\left(\widehat{\beta}_1 - \beta_{10}\right)1_{(q_i\le\widehat{\gamma})} + \mathbf{x}_i'\left(\beta_{10} - \widehat{\beta}_2\right)1_{(\widehat{\gamma}<q_i\le\gamma_0)} - \mathbf{x}_i'\left(\widehat{\beta}_2 - \beta_{20}\right)1_{(q_i>\gamma_0)} \end{aligned}$$

so

$$\widehat{\varepsilon}_i^2 = \varepsilon_i^2 + \varepsilon_i \mathbf{x}_i' O\left(n^{-1/2}\right) + O\left(n^{-1/2}\right) \mathbf{x}_i \mathbf{x}_i' O\left(n^{-1/2}\right),$$

where $\beta_{10} - \widehat{\beta}_2 = \beta_{10} - \beta_{20} - \left(\widehat{\beta}_2 - \beta_{20}\right) = n^{-1/2}b + O_p\left(n^{-1/2}\right) = O_p\left(n^{-1/2}\right)$. As a result,

$$\begin{aligned}\widehat{K}(\gamma) &= \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\widehat{\varepsilon}_i^2 1_{(q_i\le\gamma)} \\ &= \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\varepsilon_i^2 1_{(q_i\le\gamma)} + \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\varepsilon_i\mathbf{x}_i' O\left(n^{-1/2}\right) + \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i' O\left(n^{-1/2}\right)\mathbf{x}_i\mathbf{x}_i' O\left(n^{-1/2}\right),\end{aligned}$$

where $O\left(n^{-1/2}\right)$ is uniform in γ. Since $\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\varepsilon_i^2 1_{(q_i\le\gamma)} \xrightarrow{p} K_-(\gamma)$ uniformly in γ, we need only show that the other two terms are $o_p(1)$ uniformly in γ. First,

$$\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\varepsilon_i\mathbf{x}_i' O\left(n^{-1/2}\right)\right\| \le O\left(n^{-1/2}\right)\frac{1}{n}\sum_{i=1}^n \|\mathbf{x}_i\|^3 |\varepsilon_i| = O\left(n^{-1/2}\right)O_p(1) = O\left(n^{-1/2}\right),$$

where $O_p(1)$ is because of $E\left[\|\mathbf{x}_i\|^3|\varepsilon_i|\right] < \infty$ by Hölder's inequality and the LLN. Second,

$$\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i' O\left(n^{-1/2}\right)\mathbf{x}_i\mathbf{x}_i' O\left(n^{-1/2}\right)\right\| \le O\left(n^{-1}\right)\frac{1}{n}\sum_{i=1}^n \|\mathbf{x}_i\|^4 = O\left(n^{-1/2}\right)O_p(1) = O\left(n^{-1/2}\right),$$

where $O_p(1)$ is because of $E\left[\|\mathbf{x}_i\|^4\right] < \infty$ and the LLN. ■

Appendix B: Lemmas

Lemma 1 *Under the assumptions of Theorem 2,* $\widehat{\beta} - \beta_n = O_p(n^{-1/2})$.

Proof. Take $\widehat{\beta}_1$ as an example since $\widehat{\beta}_2$ can be similarly analyzed. Note that

$$\sqrt{n}\left(\widehat{\beta}_1 - \beta_{1n}\right) = \left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i' 1_{(q_i\le\widehat{\gamma})}\right)\left(-\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i' 1_{(\gamma_n<q_i\le\widehat{\gamma})}\sqrt{n}\delta_n + \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{x}_i\varepsilon_i 1_{(q_i\le\widehat{\gamma})}\right).$$

From Lemma 1 of Hansen (1996), we have $\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i' 1_{(q_i\le\widehat{\gamma})} = O_p(1)$, and $\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i' 1_{(\gamma_n<q_i\le\widehat{\gamma})} = O_p(1)$. From Lemma A.3 of Hansen (2000), $\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{x}_i\varepsilon_i 1_{(q_i\le\widehat{\gamma})} = O_p(1)$. Given that $\sqrt{n}\delta_n \to b \in \mathbb{R}^k$, we have $\sqrt{n}\left(\widehat{\beta}_1 - \beta_{1n}\right) = O_p(1)$. ■

Lemma 2 *Under the assumptions of Theorem 2,, uniformly for* $u \in U$ *with* U *being any compact set of* $\mathbb{R}^{2k}$ *and* $\gamma \in \Gamma$,

$$\begin{aligned} & S_n\left(\gamma, \beta_n + \frac{u}{\sqrt{n}}\right) - S_n\left(\gamma_n, \beta_n\right) \\ \rightsquigarrow\ & u_1' E\left[\mathbf{x}\mathbf{x}' 1_{(q\le\gamma\wedge\gamma_0)}\right] u_1 + u_2' E\left[\mathbf{x}\mathbf{x}' 1_{(q>\gamma\vee\gamma_0)}\right] u_2 + (b+u_1)' E\left[\mathbf{x}\mathbf{x}' 1_{(\gamma_0<q\le\gamma\vee\gamma_0)}\right](b+u_1) \\ & + (b-u_2)' E\left[\mathbf{x}\mathbf{x}' 1_{(\gamma\wedge\gamma_0<q\le\gamma_0)}\right](b-u_2) - W(u,\gamma) + D(u,\gamma),\end{aligned}$$

where

$$\begin{aligned} W(u,\gamma) &= 2u_1'W(\gamma\wedge\gamma_0)+2u_2'(W(\infty)-W(\gamma\vee\gamma_0)), \\ D(u,\gamma) &= 2(b-u_2)'(W(\gamma_0)-W(\gamma\wedge\gamma_0))-2(b+u_1)'(W(\gamma\vee\gamma_0)-W(\gamma_0)), \end{aligned}$$

with $W(\gamma)$ defined in Theorem 2.

Proof. By Lemma 1 of Hansen (1996), as $\gamma_n\to\gamma_0$, uniformly in $(u,\gamma)\in U\times\Gamma$,

$$\begin{aligned} &S_n\left(\gamma,\beta_n+\tfrac{u}{\sqrt{n}}\right)-S_n(\gamma_n,\beta_n) \\ &=\sum_{i=1}^n\left(u_1'\tfrac{\mathbf{x}_i\mathbf{x}_i'}{n}u_1-\tfrac{2u_1'}{\sqrt{n}}\mathbf{x}_i\varepsilon_i\right)1_{(q_i\le\gamma\wedge\gamma_n)}+\sum_{i=1}^n\left(u_2'\tfrac{\mathbf{x}_i\mathbf{x}_i'}{n}u_2-\tfrac{2u_2'}{\sqrt{n}}x_i\varepsilon_i\right)1_{(q_i>\gamma\vee\gamma_n)} \\ &+\sum_{i=1}^n\left[(\sqrt{n}\delta_n-u_2)'\tfrac{\mathbf{x}_i\mathbf{x}_i'}{n}(\sqrt{n}\delta_n-u_2)+2\left(\tfrac{\sqrt{n}\delta_n-u_2}{\sqrt{n}}\right)'\mathbf{x}_i\varepsilon_i\right]1_{(\gamma\wedge\gamma_n<q_i\le\gamma_n)} \\ &+\sum_{i=1}^n\left[(\sqrt{n}\delta_n+u_1)'\tfrac{\mathbf{x}_i\mathbf{x}_i'}{n}(\sqrt{n}\delta_n+u_1)-2\left(\tfrac{\sqrt{n}\delta_n+u_1}{\sqrt{n}}\right)'\mathbf{x}_i\varepsilon_i\right]1_{(\gamma_n<q_i\le\gamma\vee\gamma_n)} \\ &=u_1'E\left[\mathbf{x}\mathbf{x}'1_{(q\le\gamma\wedge\gamma_0)}\right]u_1+u_2'E\left[\mathbf{x}\mathbf{x}'1_{(q>\gamma\vee\gamma_0)}\right]u_2 \\ &+(b-u_2)'E\left[\mathbf{x}\mathbf{x}'1_{(\gamma\wedge\gamma_0<q\le\gamma_0)}\right](b-u_2)+(b+u_1)'E\left[\mathbf{x}\mathbf{x}'1_{(\gamma_0<q\le\gamma\vee\gamma_0)}\right](b+u_1) \\ &-W_n(u,\gamma)+D_n(u,\gamma)+o_P(1), \end{aligned}$$

where $W_n(u,\gamma)=W_{1n}(u_1,\gamma)+W_{2n}(u_2,\gamma)$ with

$$W_{1n}(u_1,\gamma)=\frac{2u_1'}{\sqrt{n}}\sum_{i=1}^n\mathbf{x}_i\varepsilon_i1_{(q_i\le\gamma\wedge\gamma_n)}\text{ and }W_{1n}(u_1,\gamma)=\frac{2u_2'}{\sqrt{n}}\sum_{i=1}^n\mathbf{x}_i\varepsilon_i1_{(q_i>\gamma\vee\gamma_n)},$$

and $D_n(u,\gamma)=D_{1n}(u_1,\gamma)+D_{2n}(u_2,\gamma)$ with

$$D_{1n}(u_1,\gamma)=2\left(\frac{b-u_2}{\sqrt{n}}\right)'\sum_{i=1}^n\mathbf{x}_i\varepsilon_i1_{(\gamma\wedge\gamma_n<q_i\le\gamma_n)}\text{ and }D_{2n}(u_1,\gamma)=-2\left(\frac{b+u_1}{\sqrt{n}}\right)'\sum_{i=1}^n\mathbf{x}_i\varepsilon_i1_{(\gamma_n<q_i\le\gamma\vee\gamma_n)}.$$

By Lemma A.3 of Hansen (2000),

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n\mathbf{x}_i\varepsilon_i1_{(q_i\le\gamma\wedge\gamma_n)}\rightsquigarrow W(\gamma\wedge\gamma_0).$$

Similarly,

$$\begin{aligned} \frac{1}{\sqrt{n}}\sum_{i=1}^n\mathbf{x}_i\varepsilon_i1_{(q_i>\gamma\vee\gamma_n)} &\rightsquigarrow W(\infty)-W(\gamma\vee\gamma_0), \\ \frac{1}{\sqrt{n}}\sum_{i=1}^n\mathbf{x}_i\varepsilon_i1_{(\gamma\wedge\gamma_n<q_i\le\gamma_n)} &\rightsquigarrow W(\gamma_0)-W(\gamma\wedge\gamma_0), \\ \frac{1}{\sqrt{n}}\sum_{i=1}^n\mathbf{x}_i\varepsilon_i1_{(\gamma_n<q_i\le\gamma\vee\gamma_n)} &\rightsquigarrow W(\gamma\vee\gamma_0)-W(\gamma_0). \end{aligned}$$

The results of this lemma follow. ■

Online Supplement for "Calibrating the Confidence Intervals in Threshold Regression"

More on the Distribution of $\widetilde{\gamma}$

First, we explain why $\arg\max_{\gamma\in\Gamma}\Xi(\gamma)$ is more likely near the boundary of $[\underline{\gamma},\overline{\gamma}]$. Consider the simple structural change model with $\mathbf{x}=1$. In this case, the covariance of $M^{-1/2}(\gamma)W(\gamma)$ ($\overline{M}^{-1/2}(\gamma)\overline{W}(\gamma)$) at s and t with $s<t$ is $s^{-1/2}st^{-1/2}=\sqrt{s/t}$ ($(1-s)^{-1/2}(1-t)(1-t)^{-1/2}=\sqrt{(1-t)/(1-s)}$), which is not independent of s and t. The total covariance is $\sqrt{s/t}+\sqrt{(1-t)/(1-s)}$ is maximized when s and t are at the middle of $[\underline{\gamma},\overline{\gamma}]$ and minimized when s and t are at the boundary of $[\underline{\gamma},\overline{\gamma}]$. In other words, there would be more variation at the boundary than in the middle of $[\underline{\gamma},\overline{\gamma}]$ in $W(\gamma)'M^{-1}(\gamma)W(\gamma)+\overline{W}(\gamma)'\overline{M}^{-1}(\gamma)\overline{W}(\gamma)$. As a result, $\arg\max_{\gamma\in\Gamma}\Xi(\gamma)$ happens more likely near the boundary of $[\underline{\gamma},\overline{\gamma}]$. This is justified in Figure 9, where $\widetilde{\gamma}$ has two modes at the two boundary points in both DGPs when $b=\mathbf{0}$.

As already shown in Table 2 in finite samples, $\arg\max_{\gamma\in\Gamma}\Xi(\gamma)$ will not follow the uniform distribution on Γ. The distribution of $\arg\max_{\gamma\in\Gamma}\Xi(\gamma)$ depends on its covariance kernel. Whether this distribution is symmetric about the center of Γ depends on whether $K(\gamma)$ is symmetric about the center of Γ, say, γ_c; if $K(\gamma)=K(2\gamma_c-\gamma)$, then $\widetilde{\gamma}$ is symmetrically distributed. From Figure 9, even if $K(\gamma)$ is asymmetric about the center of Γ as in DGP1, the distribution of $\widetilde{\gamma}$ is close to be symmetric. The distribution of $\widehat{\gamma}$ does not have modes at the boundary of Γ because in finite samples, Γ is not fixed so that the two modes of $\widetilde{\gamma}$ are dispersed out of Γ. Jiang et al. (2019) show in a simple structural change model that when $b\neq\mathbf{0}$, the distribution of $\widetilde{\gamma}$ (and also $\widehat{\gamma}$ since their Γ is fixed) is asymmetric when $\gamma_0\neq 0.5$ and tri-modal at the two boundaries of Γ and γ_0. Here, we show that trimodality reduces to bimodality and asymmetry becomes symmetry when $b=\mathbf{0}$. This is because γ_0 is not identifiable when $b=\mathbf{0}$ so there is no obvious center for the distribution of $\widetilde{\gamma}$; as a result, the modality at γ_0 disappears and asymmetry when $\gamma_0\neq 0.5$ become symmetry about 0.5. Figure 10 and 11 show the distribution of $\widehat{\gamma}$ (and $\widetilde{\gamma}$) as $\delta_{20}=0.25$ ($b_2=2.5$) in DGP1 and $\delta_{20}=0.5$ ($b_2=5$) in DGP2 when $\gamma_0=2$ and $\gamma_0=1$, respectively. When $\gamma_0=2$, the distributions of $\widehat{\gamma}$ and $\widetilde{\gamma}$ are symmetric in DGP2 and close to be symmetric in DGP1; while when $\gamma_0=1$, the distributions are asymmetric in both DGPs. The distribution of $\widetilde{\gamma}$ are tri-modal in both DGPs for both $\gamma_0=2$ and $\gamma_0=1$, while the modes of $\widehat{\gamma}$ at the two boundaries of Γ are dispersed out as in the $\delta_{20}=0$ ($b_2=0$) case.

Finally, the distributions of $\max_{\gamma\in\Gamma}\Xi(\gamma)-\Xi(\gamma_0)$ and $\max[\xi_1(\gamma_0),\xi_2(\gamma_0)]-\min[\xi_1(\gamma_0),\xi_2(\gamma_0)]$ depends on γ_0. Although $\max_{\gamma\in\Gamma}\Xi(\gamma)=\max[\xi_1(\gamma_0),\xi_2(\gamma_0)]$ does not depend on γ_0, $\Xi(\gamma_0)$ and $\min[\xi_1(\gamma_0),\xi_2(\gamma_0)]$ indeed depend on γ_0. Because $\arg\max_{\gamma\in\Gamma}\Xi(\gamma)$ is more likely near the boundary, $\max_{\gamma\in\Gamma}\Xi(\gamma)-\Xi(\gamma_0)$ tends to be zero (i.e., small) when γ_0 near the boundary. This implies that the critical value tends to be small; however, $L_n(\gamma_0)$ may also tend to be small so that $L_n(\gamma_0)$ is smaller than the critical value. Note further that in (3)-(6), we need to check the inequalities for any $\gamma_0\in\Gamma$; Figures 4 and 5 show these four inequalities hold only for the true γ_0.

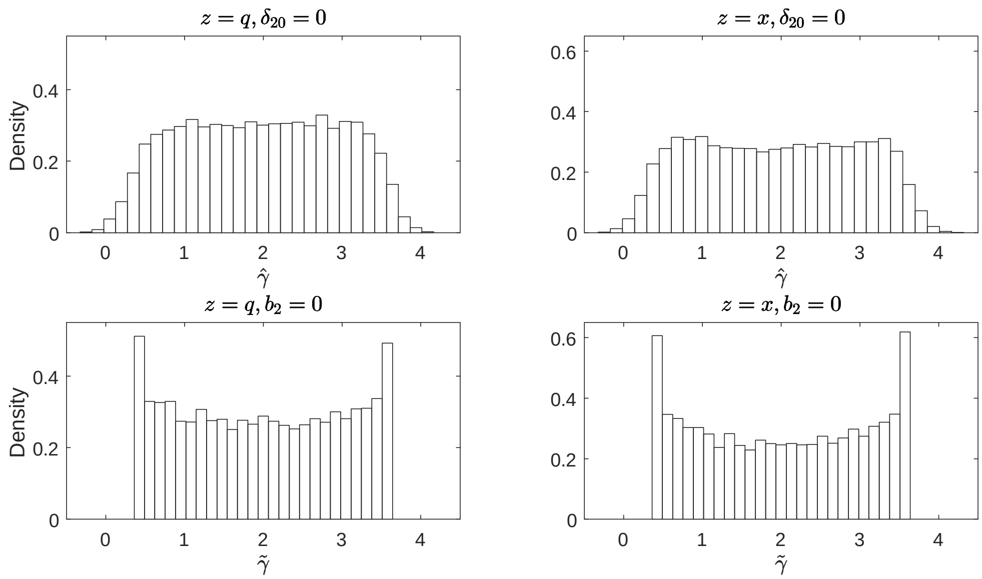

Figure 9: Density of $\widehat{\gamma}$ and $\widetilde{\gamma}$: $n = 100$

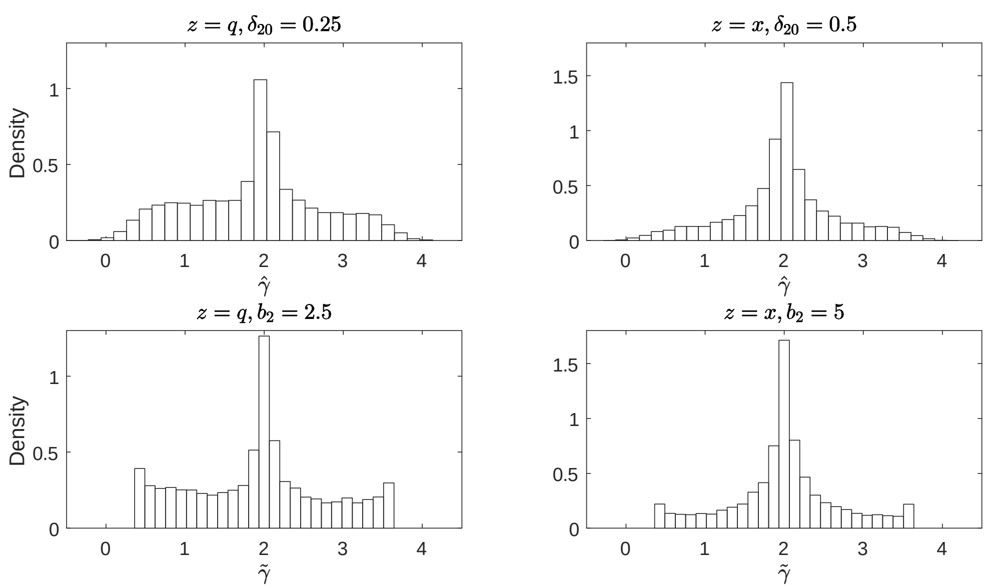

Figure 10: Density of $\widehat{\gamma}$ and $\widetilde{\gamma}$: $n = 100$, $\gamma_0 = 2$

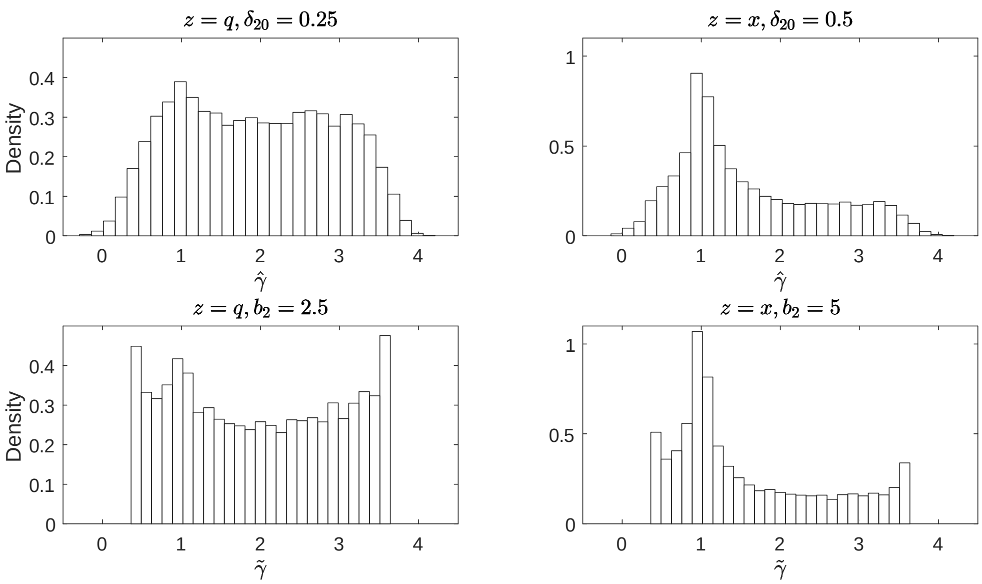

Figure 11: Density of $\widehat{\gamma}$ and $\widetilde{\gamma}$: $n = 100$, $\gamma_0 = 1$