Threshold Regression With A Threshold Boundary*

Ping Yu†

University of Hong Kong

Xiaodong Fan‡

Monash University

*We want to thank the editor, Jianqing Fan, two associate editors and two anonymous referees for constructive comments, the seminar participants at CityUHK, CUHK and SUFE for helpful discussions, and Myung Hwan Seo for helps in coding the MIO algorithms.

†Faculty of Business and Economics, The University of Hong Kong, Pokfulam Road, Hong Kong; Corresponding Author Email: pingyu@hku.hk.

‡Department of Economics, Monash University, Clayton, Victoria 3800, Australia; Email: Xiaodong.Fan@monash.edu.

Corresponding author Ping Yu whistle.yu@gmail.com

Abstract

This paper studies computation, estimation, inference and testing for linearity in threshold regression with a threshold boundary. We first put forward a new algorithm to ease the computation of the threshold boundary, and develop the asymptotics for the least squares estimator in both the fixed-threshold-effect framework and the small-threshold-effect framework. We also show that the inverting-likelihood-ratio method is not suitable to construct confidence sets for the threshold parameters, while the nonparametric posterior interval is still applicable. We then propose a new score-type test to test for the existence of threshold effects. Comparing with the usual Wald-type test, it is computationally less intensive, and its critical values are easier to obtain by the simulation method. Simulation studies corroborate the theoretical results. We finally conduct two empirical applications in labor economics to illustrate the nonconstancy of thresholds.

KEYWORDS: threshold regression, threshold boundary, Poisson point process, compound Poisson field, two-sided Brownian field, epi-convergence, score test, simulation method

JEL-CLASSIFICATION: C21, C24

1 Introduction

In recognition of potential shifts in economic relationships, threshold models have become increasingly popular in recent econometric practice. One typical application of the threshold model in time series is to illustrate asymmetric effects of shocks over the business cycle, see, e.g., Potter (1995). Many other important applications of the threshold model in time series are summarized in Hansen (2011). Threshold models are also useful in cross sectional applications. For example, Hansen (2000a) applies the threshold model to show that depending on the starting point, rich countries and poor countries have different growth patterns. The popularity of threshold models can be explained by two reasons. First, policy makers prefer threshold-type policies. For example, the tax rates depend on a few threshold income levels, and the university scholarships depend on one or a few threshold GPA levels as well. Second, the threshold model is parsimonious and allows for increased flexibility in functional form and at the same time is not as susceptible to curse of dimensionality problems as nonparametric models.

The usual threshold model splits the sample based on an observed threshold variable q:

$$y = \begin{cases} \mathbf{x}'\beta_1 + \varepsilon_1, q \le \gamma, \\ \mathbf{x}'\beta_2 + \varepsilon_2, q > \gamma, \end{cases} \quad (1)$$

where $\mathbf{x} = \left(1, x', q\right)' \in \mathbb{R}^{d+1}$, d is the dimension of nonconstant covariates, x is the nonconstant covariates except q, β_1 and β_2 are slope parameters in the two regimes defined by q exceeding the threshold point γ or not, and all the other

variables have the same definitions as in the linear regression framework. Usual parameters of interest are $\theta=\left(\gamma,\beta_1',\beta_2'\right)':=:\left(\gamma,\beta'\right)'$ or equivalently, $\theta=\left(\gamma,\beta_2',\delta'\right)'$ with $\delta=\beta_1-\beta_2$. Under the mean independence assumption $\mathbb{E}[\varepsilon_\ell \mid x,q]=0,\ell=1,2$, the usual estimator of θ is the least squares estimator (LSE).

A theory of estimation and inference is fairly well developed in this usual setup of threshold regression, see, e.g., Chan (1993), Hansen (2000a) and Yu (2012) among others. However, this setup of threshold regression has a key limitation, namely, the threshold point is the same for all subgroups of population. We use the classical return-to-schooling example to illustrate this point. As in Mincer's (1974) model, suppose y is the log wage, q is the education level, and x includes experience and experience squared. Model (1) states that for different levels of education, the returns to schooling are different. However, it is common to believe that the threshold levels of education for men and women should be different; in other words, it is better to model the threshold as $1(q\le\gamma_1+D\gamma_2)$, where $1(\cdot)$ is the indicator function, and D is a dummy for female. Note that it is impossible (when D is continuous as in our two empirical applications in Section 7) or quite burdensome (when D is discrete) to express such a model as a usual threshold regression with multiple threshold points. For example, in the return-to-schooling example above, the threshold model with a binary D and $\gamma_2>0$ can be re-written as

$$y=\begin{cases}\mathbf{x}'\beta_1+\varepsilon_1, & \text{if } q\le\gamma_1,\\ D\left(\mathbf{x}'\beta_1+\varepsilon_1\right)+\left(1-D\right)\left(\mathbf{x}'\beta_2+\varepsilon_2\right), & \text{if } \gamma_1<q\le\gamma_2,\\ \mathbf{x}'\beta_2+\varepsilon_2, & \text{if } q>\gamma_2.\end{cases}$$

That is, the regressors in each regime are different and there are cross-regime restrictions on the slope parameters. When the sign of γ_2 is unknown, it is impossible to rewrite the original model in this way. When D takes more than two values, the re-expression is even more burdensome. As a result, we will consider threshold regression with a threshold boundary in the following way,

$$y = \begin{cases} \mathbf{x}'\beta_1 + \varepsilon_1, & q \le \mathbf{z}'\gamma, \\ \mathbf{x}'\beta_2 + \varepsilon_2, & q > \mathbf{z}'\gamma, \end{cases} \qquad (2)$$
$$\mathbb{E}[\varepsilon_\ell \mid x, q, z] = 0, \ell = 1, 2,$$

where $\mathbf{z} = (1, z')' \in \mathbb{R}^{k+1}, \gamma = (\gamma_1, \gamma_2')'$, and there may be overlap between x and z. The threshold boundary $1(q \le \mathbf{z}'\gamma)$ can be treated as a normalization of the usual threshold crossing model with a linear index. Specifically, the threshold boundary can be written as $1(\mathbf{Z}'\boldsymbol{\gamma} \le 0)$, where $\mathbf{Z} = (1, q, z')'$ and $\boldsymbol{\gamma} = (-\gamma_1, \gamma_q, -\gamma_2')$ with γ_q normalized as 1. A similar normalization also appears in the maximum score estimation (see, e.g., Abrevaya and Huang (2005)) or the quasi-likelihood estimation (see, e.g., Klein and Spady (1993)) of the binary choice model, but the threshold boundary does not include an error term and all variables in $\mathbf{Z}$ are observable. An alternative normalization of γ is $\|\gamma\| = 1$ as in the maximum score estimation (Example 6.4) of Kim and Pollard (1990). In this normalization, we do not need to know which component of (q, z') has a nonzero coefficient but need to estimate one more unknown parameter. Also, the setup (2) has some computational advantage (e.g., it is hard to embody $\|\gamma\| = 1$ in our algorithm of searching for the estimator of γ) so is the focus of this paper.

Threshold boundary also marks a key difference between threshold regression and the structural change model. It is well known that these two groups of models share many similarities. However, threshold boundary is unique to threshold regression since the threshold variable in the structural change model is the time index so that the form of the threshold can only take the form $1(t \le t_0)$, where t is the time index and t_0 is the true structural change point.

Due to the nonregularity of model (2), asymptotics for the usual estimator, e.g., the LSE, are as yet not developed. Seo and Linton (2007) avoid this technical difficulty by using a ”smoothed” threshold boundary and put forward a new estimator called the smoothed least squares estimator (SLSE). The asymptotic

distribution of the SLSE is normal but its convergence rate is slower than the conventional convergence rate n. We fill this gap of literature by developing the asymptotics for the LSE directly in both the fixed-threshold-effect framework of Chan (1993) and the small-threshold-effect (or shrinking-threshold-effect) framework of Hansen (2000a). Specifically, we show that the asymptotic distribution of the LSE in the former framework is related to a compound Poisson field and in the latter framework related to a two-sided Brownian field. We also extend the model to the nonlinear setup in the conditional mean of y and the threshold boundary. We consider two frameworks rather than concentrate on one of them because we need to further discuss the inference of γ. When $\mathbf{z} = 1$, the small-threshold-effect framework is more convenient since the asymptotic distribution of the likelihood ratio (LR) statistic is accessible such that the confidence set based on inverting the LR statistic in Hansen (2000a) is easy to construct. However, this is not the case when there is a threshold boundary; we show that the current asymptotic distribution is too complicated to be used for inference. On the other hand, the nonparametric posterior interval (NPI) of Yu (2015), which is justified in the fixed-threshold-effect framework, can still be used to construct a confidence interval (CI) for each component of γ.

Besides estimation of the threshold boundary, we also consider the test for a threshold effect (i.e., whether there is a threshold boundary). This test for linearity is special because under the null, the nuisance parameter γ cannot be identified so that the usual three asymptotically equivalent tests are not applicable; see Davies (1977, 1987), Andrews (1993), Andrews and Ploberger (1994) and Hansen (1996) for some classical references when $\mathbf{z} = 1$. Although a Wald-type test can be applied, i.e., check whether $\beta_1 = \beta_2$ for each γ, it is computationally troublesome when the dimension of $\mathbf{z}$ is large. To circumvent this problem, we propose a score-type test which is constructed under the null, so no sample splitting appears in the construction of the test statistic. Given that the asymptotic null distribution is not pivotal, we propose a simulation method to obtain the critical values.

In an independent work by Lee et al. (2018), the authors consider similar problems as in this paper in the time series scenario, but this paper has at least three differences from Lee et al. (2018). First, different from the usual threshold regression, computation of the LSE of γ is not trivial. In Section 2, we suggest to use a simulation method based on the MCMC algorithm to search for the LSE, while Lee et al. (2018) suggest to use the mixed integer optimization (MIO) algorithms. Second, Lee et al. (2018) do not discuss the asymptotics in the fixed-threshold-effect framework, which turns out to be quite challenging. As a corollary, Lee et al. (2018) do not discuss the inference on γ because their asymptotic distribution is too complicated. Third, we suggest to use the score test to test for a threshold effect while Lee et al. (2018) suggest to use the LR test. It seems that our test is more computationally efficient because out test does not need to calculate the likelihood ratio for each γ in the parameter space and needs only to run a single linear regression. On the other hand, Lee et al. (2018) discuss two important problems which are not covered in this paper. We briefly discuss these two issues at the end of our conclusion section. In summary, this paper and Lee et al. (2018) are more complements than substitutes.

The rest of the paper is organized as follows. First, we develop an algorithm to compute the LSE in Section 2. Section 3 includes the asymptotics for the LSE in the two frameworks and Section 4 discusses the inference of γ. Section 5 constructs the new score-type test and simulates the critical values. Section 6 contains some Monte Carlo simulation results. Section 7 presents two empirical applications in labor economics. Section 8 concludes. Mathematical proofs of the theorems are presented in the Appendix. To save space, all supporting lemmas are collected in an online supplement.

Some notations are collected here for future reference. The letter C denotes a generic positive constant whose value may change in each occurrence. The symbols $\wedge$ and $\vee$ are min and max operators, i.e., $a \wedge b = \min\{a,b\}$, $a \vee b = \max\{a,b\}$. $\|\cdot\|$ denotes the Euclidean norm. For a matrix

A, $A<\infty$ means every element of A is finite and $A > 0$ means it is positive definite. Parameters with subscript 0 denote their true values. In the small-threshold-effect framework, the true values of β and δ may depend on n; we still use β_0 to denote the true value of β but use δ_n to denote the true value of δ. Since the problem is trivial if q can be perfectly predicted by z, we define $\epsilon = q - \mathbf{z}'\gamma_0$. $\varepsilon = \varepsilon_1 1(q \le \mathbf{z}'\gamma_0) + \varepsilon_2 1(q > \mathbf{z}'\gamma_0)$ is the error term of the outcome equation. $\rightsquigarrow$ signifies weak convergence over a compact metric space. $U[a,b]$ is the uniform distribution on an interval $[a,b]$. The symbol ℓ is used to indicate the two regimes in (2) and, to simplify notation in what follows, the explicit values "$\ell = 1, 2$" are often omitted.

2 Computation of the Threshold Boundary

Suppose the data $\{w_i\}_{i=1}^n$ are i.i.d. (independent and identically distributed) sampled, where $w_i = (y_i, x_i', z_i', q_i)'$. In (2), the LSE of γ is usually defined by a profiled procedure:

$$\hat{\gamma} = \arg\min_{\gamma \in \Gamma} M_n(\gamma),$$

where Γ is the parameter space for γ, and

$$M_n(\gamma) := \min_{\beta_1, \beta_2} \frac{1}{n}\sum_{i=1}^{n} m(w_i \mid \theta),$$

with

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

When $\mathbf{z} = 1$, $M_n(\gamma)$ reduces to the objective function of LSE in the usual threshold regression model (1). It is well known that there is an interval of γ minimizing $M_n(\gamma)$ in this case. Most literature uses the left-endpoint LSE (LLSE), while Yu

(2012, 2015) shows that the middle-point LSE (MLSE) is more efficient in most cases. It is interesting to define the counterpart of the MLSE in the general case.

The estimation of β is invariant to the estimation of γ as long as the sample splitting is the same. To express the β estimator in matrix notation, define the $n\times 1$ vectors Y by stacking the variables y_i, and the $n\times(d+1)$ matrices $X_{\le\gamma}$ and $X_{>\gamma}$ by stacking the vectors $\mathbf{x}_i'1(q_i\le \mathbf{z}_i'\gamma)$ and $\mathbf{x}_i'1(q_i>\mathbf{z}_i'\gamma)$. Let

$$\begin{pmatrix}\hat\beta_1(\gamma)\\ \hat\beta_2(\gamma)\end{pmatrix}=\arg\min_{\beta_1,\beta_2}\ \frac{1}{n}\sum_{i=1}^{n}m(w_i\mid\theta)=\begin{pmatrix}\left(X_{\le\gamma}'X_{\le\gamma}\right)^{-1}X_{\le\gamma}'Y\\ \left(X_{>\gamma}'X_{>\gamma}\right)^{-1}X_{>\gamma}'Y\end{pmatrix};$$

then the LSE of β is defined as $\hat\beta=\left(\hat\beta_1'(\hat\gamma),\hat\beta_2'(\hat\gamma)\right)'=:\left(\hat\beta_1',\hat\beta_2'\right)'$, and $\hat\theta=\left(\hat\gamma',\hat\beta'\right)'$. Also,

$$M_n(\gamma):=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\mathbf{x}_i'\hat\beta_1(\gamma)1(q_i\le\mathbf{z}_i'\gamma)-\mathbf{x}_i'\hat\beta_2(\gamma)1(q_i>\mathbf{z}_i'\gamma)\right)^2.$$

2.1 Difficulties in the Calculation of $\hat\gamma$

To appreciate the difficulty in calculating the LSE of γ, consider the following simple example,

$$y=1(q\le\gamma_1+\gamma_2 z);\quad (3)$$

in other words, $\mathbf{x}=1,\beta_1=1$ and $\beta_2=0$ are known, and $\varepsilon=0$. A similar example is considered in Section 2 of Yu (2012). Such a simplification can focus attention on the estimation of γ and also avoid the complexities introduced by ε_ℓ. The left panel of Figure 1 shows ten data points sampled from the specification that z follows $U[0,1]$, ϵ follows $U[-0.5,0.5]$ and is independent of z, and $\gamma_{10}=\gamma_{20}=1$. As shown in the right panel of Figure 1, the choice of $\hat\gamma$ that is consistent with the data is not unique. Note here that although γ_0 is identified by a set in finite samples, the set will shrink to a point asymptotically, so the problem here is point

identified rather than partially identified. In Figure 1, we also draw four extreme cases that $\hat{\gamma}$ can take in this dataset. The identified set is defined by the convex hull of the four extreme points. This is generally correct for a linear boundary specification $1(q \le \mathbf{z}'\gamma)$. It is natural to define the counterpart of the MLSE as the center of gravity of the identified set, which is defined as

$$\frac{\int_{\arg\min_\gamma M_n(\gamma)} \gamma d\gamma}{\int_{\arg\min_\gamma M_n(\gamma)} d\gamma}.$$

Since $\arg\min_\gamma M_n(\gamma)$ is a convex set, the centroid reduces to the average of the extreme points. In Figure 1, such an estimate is shown as a blue dot, which is quite close to $\gamma_0 = (1,1)'$, the red open circle in the figure. On the contrary, the counterpart of the LLSE, the two solid boundaries of the convex set, seems less favorable since it may be far from the true value and is not uniquely defined. In what follows, the $\arg\min$ operator always means the centroid of the minimizing set if the set includes more than one points. Finally, note from Yu (2012, 2015), γ can be treated as a boundary of q. However, this boundary is different from the boundary studied in Hirano and Porter (2003), Chernozhukov and Hong (2004) and Knight (2006) – the boundary there is defined by the jump in the conditional density of q "directly", while the boundary in threshold regression is defined " indirectly" by the jump in the conditional expectation of another variable y. This is also why the boundary in Hirano and Porter (2003) and Chernozhukov and Hong (2004) can be identified uniquely in finite samples while the boundary here cannot.

There are two specialities for this simple example, which need not hold in the general case. First, γ_0 must fall in the identified set; second, the identified set can be easily calculated. To appreciate these difficulties in the general case, assume $\beta_{\ell 0}$ is known in (2). Then

$$\hat{\gamma} = \arg\min_{\gamma} \; n\left[M_n(\gamma) - M_n(\gamma_0)\right],$$

where

$$\begin{aligned} & \mathrm{n}\left[\mathrm{M_n}(\gamma)\text{-}\mathrm{M_n}(\gamma_0)\right] \\ &= \sum_{i=1}^{n}\left(y_i - \mathbf{x}_i'\beta_{10}1(q_i \le \mathbf{z}_i'\gamma) - \mathbf{x}_i'\beta_{20}1(q_i > \mathbf{z}_i'\gamma)\right)^2 - \sum_{i=1}^{n}\varepsilon_i^2 \\ &= \sum_{i=1}^{n}\bar{Z}_{1i}1\left(\mathbf{z}_i'\gamma < q_i \le \mathbf{z}_i'\gamma_0\right) + \sum_{i=1}^{n}\bar{Z}_{2i}1\left(\mathbf{z}_i'\gamma_0 < q_i \le \mathbf{z}_i'\gamma\right) \\ &= \sum_{i=1}^{n}\bar{Z}_{1i}1\left(\mathbf{z}_i'\left(\gamma-\gamma_0\right) < \epsilon_i \le 0\right) + \sum_{i=1}^{n}\bar{Z}_{2i}1\left(0 < \epsilon_i \le \mathbf{z}_i'\left(\gamma-\gamma_0\right)\right) \end{aligned} \tag{4}$$

with

$$\begin{aligned} & y_i = \mathbf{x}_i'\beta_{10}1(q_i \le \mathbf{z}_i'\gamma_0) + \mathbf{x}_i'\beta_{20}1(q_i > \mathbf{z}_i'\gamma_0) + \varepsilon_i, \\ & \bar{Z}_{1i} = 2\mathbf{x}_i'\delta_0\varepsilon_{1i} + \delta_0'\mathbf{x}_i\mathbf{x}_i'\delta_0 \text{ and } \bar{Z}_{2i} = -2\mathbf{x}_i'\delta_0\varepsilon_{2i} + \delta_0'\mathbf{x}_i\mathbf{x}_i'\delta_0. \end{aligned} \tag{5}$$

It is obvious that $\hat{\gamma}$ depends on $\left\{\bar{Z}_{\ell i}\right\}_{i=1}^{n}$ and is very complicated to compute. Figure 2 shows one set of possible (random) jumping locations of $n\left[M_n(\gamma) - M_n(\gamma_0)\right]$, where we center the figure at γ_0, the red dashed lines represent the jumping locations for $\sum_{i=1}^{n}\bar{Z}_{1i}1\left(\mathbf{z}_i'\left(\gamma-\gamma_0\right) < \epsilon_i \le 0\right)$, the blue solid lines represent the jumping locations for $\sum_{i=1}^{n}\bar{Z}_{2i}1\left(0 < \epsilon_i \le \mathbf{z}_i'\left(\gamma-\gamma_0\right)\right)$, and the details for the specification of the distributions of ϵ and z are given in Section 3.1. On each convex set defined by these lines, we assign a random jump size defined by partial summation of $\left\{\bar{Z}_{\ell i}\right\}_{i=1}^{n}$. We then determine on which set $n\left[M_n(\gamma) - M_n(\gamma_0)\right]$ is minimized. Obviously, $\hat{\gamma} - \gamma_0$ need not be the convex set covering (0, 0).

2.2 Minimization Using the MCMC Algorithm

Given these difficulties, we suggest to use a simulation method to achieve the minimization. A similar method is employed by Chernozhukov and Hong (2004) in different contexts. Specifically, we use the following algorithm.

Algorithm M:

Step 1: Define

$$p_n(\gamma) = \frac{\exp\{-M_n(\gamma)\} 1(\gamma \in \Gamma)}{\int_\Gamma \exp\{-M_n(\gamma)\}\, d\gamma},$$

which is the (concentrated) quasi-posterior of γ with a uniform prior on Γ.

Step 2: Draw a Markov chain

$$S = \left(\gamma^{(1)}, \cdots, \gamma^{(B)}\right),$$

whose marginal density is approximately given by $p_n(\gamma)$.

Step 3: For each $\gamma^{(b)}, b = 1, \cdots, B$, calculate $M_n(\gamma^{(b)})$. Define $\hat{\gamma}_I = \arg\min_{\gamma \in S} M_n(\gamma)$ as the initial estimation of γ. Note that $\hat{\gamma}_I$ may be a set of γ values.

Step 4: Refine the simulation set in Step 1 from Γ to a neighborhood of $\hat{\gamma}_I$. For example, replace Γ by

$$\tilde{\Gamma} := \bigotimes_{l=1}^{k+1} \left[\max \underline{S}_l - \delta_l, \min \overline{S}_l + \delta_l\right],$$

where $\underline{S}_l$ is the collection of the lth entry of S which is smaller than the corresponding entry of $\hat{\gamma}_I$ (or minimum if $\hat{\gamma}_I$ is a set), $\overline{S}_l$ is the collection of the lth entry of S which is larger than the corresponding entry of $\hat{\gamma}_I$ (or maximum if $\hat{\gamma}_I$ is a set), δ_l is a small number, and $\bigotimes$ defines the product set in each dimension.

Step 5: Repeat Step 2 and 3 to get an updated set of γ estimation, say, $\hat{\gamma}_N$. Then $\hat{\gamma}$ is defined as the average of the points in $\hat{\gamma}_N$, which approximates the center of gravity of the identified set.

The idea of our algorithm is simple – first get a rough idea where the identified set is and then refine the simulation in its neighborhood. We give a few

comments on Algorithm M. First, comparing to the grid search, our simulation method is more efficient since the probability to be simulated for γ values such that $M_n(\gamma)$ is small is high. As a result, more γ values are drawn on (and around) the identified set, which is exactly what we want – depict the shape of the identified set as precisely as we can. Second, in Step 2, the MCMC method such as the Metropolis-Hastings sampler or the slice sampler can be used to simulate S. Because specification of transition probability functions and stopping rules in MCMC algorithms is standard nowadays (see, e.g., Robert and Casella (2004) and Rubinstein and Kroese (2017) for recent reviews), we omit the details here but mention some specifics of our implementation in the simulation section below (i.e., Section 6). Third, for the MCMC algorithm in Step 2, we need to specify a starting point. A natural choice is the LSE or LADE of q on $\mathbf{z}$ which guarantees that each regime contains about half of the observations. In Step 5, we can use $\hat{\gamma}_I$ as the starting point for the MCMC algorithm. Fourth, when B is large enough, we can guarantee that the global minimizer of $M_n(\gamma)$ is achieved in Step 3. This is because when B goes to infinity, the density of $\{\gamma^{(b)}\}_{b=1}^{B}$ from Step 2 would converge to $p_n(\gamma)$ (see, e.g., Section 7.3.2 of Robert and Casella (2004) for the Metropolis-Hastings algorithm and Section 8.3 for the slice sampler which is used in our simulation), while the minimizing set of $M_n(\gamma)$, say $\hat{\Gamma}_n$, is the mode of $p_n(\gamma)$ so that $p_n(\hat{\Gamma}_n) > 0$. In practice, B is finite; how a finite B affects the minimizer of Algorithm M, e.g., $\hat{\gamma}_I$, is a complicated question and beyond the scope of this paper. We leave it as a future research topic. Fifth, Steps 4 and 5 are designed for practical rather than theoretical purpose. In practice, even if B = 1000, some points in $\hat{\Gamma}_n$ were drawn, so we can concentrate on the neighborhood of $\hat{\Gamma}_n$ (rather than the whole Γ) to draw $\gamma^{(b)}$ more efficiently and get a better idea on the shape of $\hat{\Gamma}_n$. To improve the preciseness of the identified set, we can increase B in Step 4 or conduct a further refinement in Step 5. Sixth, in Step 4, δ_l can be set as 0.1 times the range of the lth element of S, which is

used in our simulation. Also, to avoid missing points in the identified set, δ_l can be chosen by the method of trial and error.

2.3 Discussions

First, we discuss the specification of Γ in practice. In the usual threshold regression, Γ is often specified as $[q_{(\alpha_1 n)}, q_{(\alpha_2 n)}]$, where $q_{(l)}$ is the lth order statistic of $\{q_i\}_{i=1}^n$, and α_2 is often specified as $1-\alpha_1$ and $\alpha_1 = 0.05, 0.1, 0.15$ or 0.2. In the general case, it is hard to find all γ values such that each regime includes at least, say, 15%, of all data points. Actually, even if the set of all such γ's can be found, it is not compact. Nevertheless, the specification of Γ does not need to find all possible γ's. To illustrate this point, we consider an extremely simple example. Suppose we have only five data points of (q, z) as shown in the left panel of Figure 3, and we must guarantee that each regime has at least two points to fit a straight line. Then Γ can be the set

$$\gamma_1 \in \begin{cases} [-1,2), & \text{if } \gamma_2 \in (-\infty,-1), \\ [\gamma_2,2), & \text{if } \gamma_2 \in [-1,-0.5), \\ [\gamma_2,-4\gamma_2), & \text{if } \gamma_2 \in [-0.5,0), \\ \varnothing, & \text{if } \gamma_2 = 0, \\ [-4\gamma_2,\gamma_2), & \text{if } \gamma_2 \in (0,0.25], \\ [-1,\gamma_2), & \text{if } \gamma_2 \in (0.25,2], \\ [-1,2), & \text{if } \gamma_2 \in (2,\infty), \end{cases}$$

which is shown in the right panel of Figure 3 as the area encompassed by the blue solid and dashed lines, where $\varnothing$ means the empty set. Obviously, this parameter space is not compact. This result is generic since for any data set, γ_2 can diverge to $\pm\infty$ although γ_1 must be bounded. This choice of Γ is the largest but not necessary. Actually, the number of possible sample splittings determined by Γ is quite limited. In the right panel of Figure 3, we also mark the "typical" combinations of γ_1 and γ_2 for all possible sample splittings by red circles, so totally only four sample splittings are possible (note that in theory, the total

possible number of splitting is $C_5^2 = 10$, much larger than the actual number 4, due to the special sample realization). As a result, choosing Γ as a compact set can generate all possible sample splittings. Specifically, if $\hat{\gamma}^L$ is the LSE or LADE of q on **z**, then first set $\Gamma_- = \bigotimes_{l=2}^{k+1}\left[\hat{\gamma}_l^L - C, \hat{\gamma}_l^L + C\right]$ for a properly large C as the parameter space for γ_-, where γ_- is γ excluding the intercept γ_1, and then choose the parameter space for γ_1 given a specific value γ_- as $\Gamma_1(\gamma_-)$ such that at least α_1 portion of data are contained in each regime with $\alpha_1 = 0.05, 0.1, 0.15$ or 0.2; in other words, the whole parameter space Γ need not take a Cartesian product form like $\Gamma_1 \times \Gamma_-$. When $k = 0$, then $\Gamma = \Gamma_1$ is exactly the specification of Γ in the usual threshold model. We can easily code the MCMC algorithm to let $p_n(\gamma)$ concentrate on Γ (see Section 6 for a concrete example). One advantage of Algorithm M is that it does not require a precise specification of Γ as long as it is a large compact set containing $\hat{\gamma}$ because it can automatically guarantee each regime contains at least α_1 portion of data and seach for $\hat{\gamma}$ efficiently. On the other hand, grid search is generally impractical. First, when $k > 1$, grid search is not applicable due to the curse of dimensionality. Second, in grid searching, we need to discard the points in the initial chosen Γ when one regime contains less than $k+1$ data points, while the MCMC algorithm can adaptively discard such points. As a last comment, note that the problem of specifying the parameter space is not unique to threshold regression; any non-concave maximization method, e.g., the usual maximum likelihood and nonlinear GMM, involves this problem.

Second, Algorithm M provides an alternative to the MIO algorithm suggested by Lee et al. (2018), but a detailed comparison between these two algorithms is beyond the scope of this paper. A comprehensive comparison between Algorithm M and the MIO algorithm (e.g., by simulations) is left as a future research topic.

Third, note that the MCMC algorithm is only auxiliary to the minimization problem. Different from Chernozhukov and Hong (2003a) and Jun et al. (2015)

where the original objective function is smoothed, we did not change our objective function in Algorithm M. The construction of $p_n(\gamma)$ in Step 1 of Algorithm M is only to simulate the possible minimizers in Step 2, and we then check in Step 3 which of these possible minimizers minimizes the original objective function.

Finally, we mention one mistaken algorithm which may be suggested by some practitioners. In the usual threshold regression, we check only q_i in Γ to find $\hat{\gamma}$. Since such q_i's are also sample quantiles of q, one may suggest to check all coefficients in the quantile regression of q on $\mathbf{z}$. It seems reasonable because all possible sample splittings based on the conditional quantiles of q are checked. However, this is not correct because $\mathbf{z}'\gamma_0$ need not be any conditional quantile of q given z. So the quantile interpretation of the threshold point in the usual threshold regression cannot be extended to the general case. For illustration, check the simple data set in Figure 3. The red dot-dashed line shows all possible coefficients in quantile regression of q on $(1, z)'$. But they define only one qualified sample splitting among the four.

3 Asymptotics for the LSE

In this section, we derive the asymptotic distributions of $\hat{\gamma}$ in two frameworks of the threshold effects.

3.1 Asymptotics with Fixed Threshold Effects

Before stating the asymptotic theory for the LSE, we first specify some regularity conditions.

Assumption D:

1. $w_i \in \mathbb{W} := \mathbb{R} \times \mathbb{X} \times \mathbb{Z} \times \mathbb{Q} \subset \mathbb{R}^{d+k+2}$ are i.i.d. sampled. $\mathbb{Z}$ is compact. $\beta_\ell \in B_\ell \subset \mathbb{R}^{d+1}, \gamma \in \Gamma \subset \mathbb{R}^{k+1}, B_\ell$ and Γ are compact, and $\beta_{\ell 0}$ and γ_0 are in the interior of B_ℓ and Γ, respectively.
2. $0 < \mathbb{E}\left[\mathbf{x}\mathbf{x}' 1(q \le \mathbf{z}'\gamma \wedge \mathbf{z}'\gamma_0)\right] < \infty$ and $0 < \mathbb{E}\left[\mathbf{x}\mathbf{x}' 1(q > \mathbf{z}'\gamma \vee \mathbf{z}'\gamma_0)\right] < \infty$ for all $\gamma \in \Gamma$.
3. In the fixed-threshold-effect framework, $0 < \delta_0' \mathbb{E}\left[\mathbf{x}\mathbf{x}' \mid q = \mathbf{z}'\gamma_0, z\right]\delta_0 < \infty$ for $z \in \mathbb{Z}$, and in the small-threshold-effect framework, replace δ_0 by $\delta_n / \| \delta_n \|$.
4. $f_{q|z}(q \mid z)$ is continuous on $\mathbb{Q}$ for each $z \in \mathbb{Z}$, and $0 < \underline{f} \le f_{q|z}(q \mid z) \le \bar{f} < \infty$ for each $z \in \mathbb{Z}$ and $q \mid z \in \{z'\gamma \mid \gamma \in \Gamma\}$.
5. $\mathbb{E}[\varepsilon_\ell^4] < \infty$ and $\mathbb{E}[\| \mathbf{x} \|^4] < \infty$.
6. The conditional distribution of $(\mathbf{x}', \varepsilon_\ell)'$ given $q_i = \mathbf{z}_i'\gamma$ and $z_i = z$ is continuous in γ for γ in a neighborhood of γ_0 and $z \in \mathbb{Z}$.
7. **z** is not multicollinear, i.e., there does not exist a nonzero vector $v \in \mathbb{R}^{k+1}$ such that $P(\mathbf{z}'v = 0) = 1$.
8. $0 < \mathbb{E}\left[\mathbf{x}\mathbf{x}' \varepsilon_1^2 1(\epsilon \le 0)\right] < \infty$, and $0 < \mathbb{E}\left[\mathbf{x}\mathbf{x}' \varepsilon_2^2 1(\epsilon > 0)\right] < \infty$.
9. Z_{1i} and Z_{2i} have absolutely continuous distributions, where $Z_{\ell i}$ follows the conditional distribution of $\bar{Z}_{\ell i}$ given $\epsilon_i = 0$ and z_i, and $\bar{Z}_{\ell i}$ is defined in (5). We denote $Z_{\ell i}$ as $\bar{Z}_{\ell i} \mid (z_i, \epsilon_i = 0)$.

Assumption D1 is standard. We can relax the compactness of B_ℓ given that the objective function is a convex function of β_ℓ. Nevertheless, such an assumption simplifies our proof and seems suitable when $\mathbb{E}[y \mid x, q]$ is extended to be a nonlinear function of β. Assumption D2 guarantees that the sample splitting by any $\gamma \in \Gamma$ would not degenerate. The corresponding assumption in the usual threshold regression is $\mathbb{E}\left[\mathbf{x}\mathbf{x}'\right] > \mathbb{E}\left[\mathbf{x}\mathbf{x}' 1(q \le \gamma)\right] > 0$ for all $\gamma \in \Gamma$, which is not suitable to be extended as $\mathbb{E}\left[\mathbf{x}\mathbf{x}'\right] > \mathbb{E}\left[\mathbf{x}\mathbf{x}' 1(q \le \mathbf{z}'\gamma)\right] > 0$ for all $\gamma \in \Gamma$ because the

collection of events $q \le \mathbf{z}'\gamma$ ($q > \mathbf{z}'\gamma$) and $q \le \mathbf{z}'\gamma \wedge \mathbf{z}'\gamma_0$ ($q > \mathbf{z}'\gamma \vee \mathbf{z}'\gamma_0$), $\gamma \in \Gamma$, need not be the same if $\mathbf{z} \ne 1$ given that $\mathbf{z}'\gamma$ and $\mathbf{z}'\gamma_0$ may cross each other when $\gamma \ne \gamma_0$. The version of Assumption D3 in the usual threshold regression is $\delta_0' \mathbb{E}\left[\mathbf{x}\mathbf{x}' \mid q = \gamma_0\right]\delta_0 > 0$, which excludes the continuous threshold regression (CTR) of Chan and Tsay (1998) (see also Hansen (2017)). Assumption D3 also excludes the CTR in the current setup. For example, if $z \in \mathbb{R}$, $x = z$, $\gamma_0 = (0,-1)'$, $\beta_{10} = (0,0,1)'$ and $\beta_{20} = (0,-1,0)'$, then $\mathbb{E}[y \mid q, z] = q1(q + z \le 0) - z1(q + z > 0)$, which is continuous at the threshold boundary $q = -z$. Such a specification is excluded by Assumption D3 because $\delta_0' \mathbb{E}\left[\mathbf{x}\mathbf{x}' \mid q = \mathbf{z}'\gamma_0, z\right]\delta_0 = (0,1,1)'(1,z,-z)(1,z,-z)'(0,1,1) = 0$. Obviously, Assumption D3 requires more than $\delta_0 \ne \mathbf{0}$. Note also that Assumption D3 requires less than $\mathbb{E}\left[\mathbf{x}\mathbf{x}' \mid q = \mathbf{z}'\gamma_0, z\right] > 0$ because x may be the same as z such that the rank of $\mathbb{E}\left[\mathbf{x}\mathbf{x}' \mid q = \mathbf{z}'\gamma, z\right]$ is d rather than $d + 1$. Assumptions D4 and D5 guarantee that such quantities as $\mathbb{E}\left[\|\mathbf{x}\|^2 \middle| z, \epsilon = 0\right]$ and $\mathbb{E}\left[\|\mathbf{x}\|^2 \varepsilon_\ell^2 \middle| z, \epsilon = 0\right]$ are finite P_z almost surely. Assumption D4 can be relaxed along the line of Yu and Zhao (2013). Assumption D6 states that the distribution of x keeps stable in the neighborhood of $\gamma = \gamma_0$ so the threshold effect is captured only by the changes in the conditional mean of y and the distribution of the error term ε. We can relax this assumption with more complicated notations (e.g., in Assumption D9); see Hansen (2000b) for relaxing this assumption in the structural change testing problem. Assumption D7 guarantees that there is no redundancy in $\mathbf{z}$. This assumption can ensure that the asymptotic distribution of $\hat{\gamma}$ is well defined. Assumption D8 is a regularity assumption guaranteeing the asymptotic distribution of $\hat{\beta}$ nondegenerate. Assumption D9 guarantees that the LSE defined in the last section is asymptotically unique.

Theorem 1. *Under Assumption D,*

$$n\left(\hat{\gamma}-\gamma_0\right)\xrightarrow{d}\arg\min_{v\in\mathbb{R}^{k+1}} D(v)=:Z_\gamma,$$

and

$$\sqrt{n}\left(\hat{\beta}_1-\beta_{10}\right)\xrightarrow{d}N\left(0,\mathbb{E}\left[\mathbf{x}\mathbf{x}'1\left(\epsilon\le 0\right)\right]^{-1}\mathbb{E}\left[\mathbf{x}\mathbf{x}'\varepsilon_1^2 1\left(\epsilon\le 0\right)\right]\mathbb{E}\left[\mathbf{x}\mathbf{x}'1\left(\epsilon\le 0\right)\right]^{-1}\right)=:Z_{\beta_1},$$
$$\sqrt{n}\left(\hat{\beta}_2-\beta_{20}\right)\xrightarrow{d}N\left(0,\mathbb{E}\left[\mathbf{x}\mathbf{x}'1\left(\epsilon>0\right)\right]^{-1}\mathbb{E}\left[\mathbf{x}\mathbf{x}'\varepsilon_2^2 1\left(\epsilon>0\right)\right]\mathbb{E}\left[\mathbf{x}\mathbf{x}'1\left(\epsilon>0\right)\right]^{-1}\right)=:Z_{\beta_2},$$

where

$$D(v)=\sum_{i=1}^{\infty}Z_{1i}1\left(\mathbf{z}_{1i}'v<J_{1i}\le 0\right)+\sum_{i=1}^{\infty}Z_{2i}1\left(0<J_{2i}\le\mathbf{z}_{2i}'v\right),$$
$$J_{1i}=\frac{\mathcal{J}_{1i}}{f_{\epsilon|z}\left(0\,|\,z_{1i}\right)},\ \mathcal{J}_{1i}=-\left(\mathcal{E}_{11}+\cdots+\mathcal{E}_{1i}\right),$$
$$J_{2i}=\frac{\mathcal{J}_{2i}}{f_{\epsilon|z}\left(0\,|\,z_{2i}\right)},\ \mathcal{J}_{2i}=\mathcal{E}_{21}+\cdots+\mathcal{E}_{2i},$$

$\left\{\mathcal{E}_{1i},\mathcal{E}_{2i}\right\}_{i=1}^{\infty}$ *are independent unit exponential variables and independent of* $\left\{z_{1i},z_{2i},Z_{1i},Z_{2i}\right\}_{i=1}^{\infty}$, $Z_{\ell i}=\bar{Z}_{\ell i}\,|\left(z_{\ell i},\epsilon_i=0\right)$ *with* $z_{\ell i}$ *following the same distribution as* z_i, $\mathbf{z}_{\ell i}=(1,z_{\ell i}')'$, *and the i.i.d. copies of* $\left\{z_{1i},Z_{1i}\right\}_{i=1}^{\infty}$ *is independent of the i.i.d copies of* $\left\{z_{2i},Z_{2i}\right\}_{i=1}^{\infty}$. *Furthermore,* Z_γ, Z_{β_1} *and* Z_{β_2} *are independent of each other.*

We provide a few comments on Theorem 1. First, *D*(*v*) is a generalization of the compound Poisson process in the usual threshold regression, and we call $D(\cdot)$ a "compound Poisson field". In the usual threshold regression, $\mathbf{z}_i=1$, so $Z_{\ell i}=\bar{Z}_{\ell i}\,|\left(q_i=\gamma_0\right)$ and $f_{\epsilon|z}\left(0\,|\,z_i\right)=f_q\left(\gamma_0\right)$. As a result, *D*(*v*) can be rewritten as

$$D(v)=\begin{cases}\sum_{i=1}^{N_1(|v|)}Z_{1i}, \text{ if } v\le 0,\\ \sum_{i=1}^{N_2(v)}Z_{2i}, \text{ if } v>0,\end{cases}$$

where $\{Z_{1i}, Z_{2i}\}_{i\geq 1}$, $N_1(\cdot)$ and $N_2(\cdot)$ are independent of each other, and $N_\ell(\cdot)$ is a Poisson process with intensity $f_q(\gamma_0)$. Note that $N_1(\cdot)$ and $N_2(\cdot)$ in the usual case and $1\left(\mathbf{z}_{1i}'v < J_{1i} \leq 0\right)$ and $1\left(0 < J_{2i} \leq \mathbf{z}_{2i}'v\right)$ in the general case determine the jumping locations of $D(\cdot)$. From Assumption D7, **z** is not multicollinear, which implies $\arg\min_v D(v)$ is a bounded set such that the center of gravity of $\arg\min_v D(v)$ is well defined. Figure 2 illustrates a set of jumping locations of $D(\cdot)$ when $k = 1$ and $z_i \sim U[0,1]$. Note that there is no multicollinearity between z and 1; otherwise, if z is a constant, then all lines in Figure 2 are parallel, and $\arg\min_v D(v)$ is not bounded. Although in the usual threshold regression, jumping locations and jump sizes (Z_{1i} and Z_{2i}) are independent, they are correlated (through z_{1i} and z_{2i}) in the general case. Also, for a given sampling path, a fixed v can appear in both summations of $D(v)$, which cannot happen in the usual case. Second, $D(\cdot)$ is more complicated than $\ell_{2\infty}$ in Chernozhukov and Hong (2004) because it cannot be expressed as an integral with respect to the Poisson random measure $\mathbf{N}(\cdot)$ given that Z_{1i} and Z_{2i} include extra randomness than z_{1i} and z_{2i}, where $\mathbf{N}(\cdot) := \sum_{i=1}^n 1\left[\left(J_{1i}, z_{1i}\right) \in \cdot\right] + \sum_{i=1}^n 1\left[\left(J_{2i}, z_{2i}\right) \in \cdot\right]$. This is also why $D(\cdot)$ is not easy to simulate as suggested in Remark 3.1 of Chernozhukov and Hong (2004) in their case. Third, we provide some intuition here on why the asymptotic distribution of $\hat{\gamma}$ involves $D(\cdot)$. This intuition is similar to that in Section 3.2 of Chernozhukov and Hong (2004). From (4),

$$n\left(\hat{\gamma} - \gamma_0\right) = \arg\min_v D_n(v),$$

where

$$D_n(v) = \sum_{i=1}^n \bar{Z}_{1i} 1\left(\mathbf{z}_i'v < n\epsilon_i \leq 0\right) + \sum_{i=1}^n \bar{Z}_{2i} 1\left(0 < n\epsilon_i \leq \mathbf{z}_i'v\right).$$

Suppose ϵ_i follows the double exponential distribution with density $\frac{1}{2}\exp\left(-|\epsilon|\right)$, which corresponds to a "generalized" boundary model with the density at the boundary equal to $f_{\epsilon|z}\left(0\,|\,z\right)=\frac{1}{2}$. For a given v, the behavior of $D_n(v)$ is determined by the order statistics of ϵ_i in the neighborhood of 0 – $n\epsilon^{-}_{(1)}, n\epsilon^{-}_{(2)},\cdots$ in the left neighborhood and $n\epsilon^{+}_{(1)}, n\epsilon^{+}_{(2)},\cdots$ in the right neighborhood. Note that the density of ϵ_i in the neighborhood of 0 is one half of the standard exponential distribution, so the Reny representation allows these rescaled order statistics to be represented almost surely as 2 times

$$-\mathcal{E}_{11},-\mathcal{E}_{11}-\frac{n}{n-1}\mathcal{E}_{12},-\mathcal{E}_{11}-\frac{n}{n-1}\mathcal{E}_{12}-\frac{n}{n-2}\mathcal{E}_{13},\cdots,$$
$$\mathcal{E}_{21},\mathcal{E}_{21}+\frac{n}{n-1}\mathcal{E}_{22},\mathcal{E}_{21}+\frac{n}{n-1}\mathcal{E}_{22}+\frac{n}{n-2}\mathcal{E}_{23},\cdots;$$

see, e.g., Embrechts, et al. (1997, p. 189), where $\mathcal{E}_{\ell i}$'s are defined in Theorem 1. For a given v, essentially only a stochastically bounded number of order statistics matters in $D_n(v)$. Hence J_{1i} and J_{2i} are the limits of these order statistics as $n\to\infty$ and follow gamma distributions with the common scale factor 2, i.e., $J_{\ell i}=2\mathcal{J}_{\ell i}$. In general, the density of ϵ_i's may vary near zero, which changes the hazard rates of the limit gamma variables, resulting in the division of $\mathcal{J}_{1i}$ and $\mathcal{J}_{2i}$ by varying hazard rates $f_{\epsilon|z}\left(0\,|\,z_{1i}\right)$ and $f_{\epsilon|z}\left(0\,|\,z_{2i}\right)$. Since $\hat{\gamma}$ is asymptotically determined by only a small portion of the entire sample near $\epsilon_i=0$, its asymptotic distribution is independent of those of regular parameters β_1 and β_2; see, e.g., Lemma 21.19 of van der Vaart (1998) or Section 4.3 of Resnick (1986). Fourth, there are two difficulties in deriving the asymptotic distribution of $\hat{\gamma}$. (i) in deriving the finite-dimensional limit of $D_n(\cdot)$, it is hard to check the weak convergence of the Poisson point process. Fortunately, the proof method in Theorem 3.1 of Chernozhukov and Hong (2004) can be adapted to solve this problem; most of the technicalities can be found in Chapter 3 of Resnick (1987). (ii) in deriving the

weak limit of $\hat{\gamma}$, the weak convergence of $D_n(\cdot)$ on a compact set is hard to check since the stochastic equicontinuity condition on the *D*-space cannot be easily extended to the multi-dimensional case. In this case, the epi-convergence, which is weaker than the usual weak convergence and is popularized by Geyer (1994), Rockafellar and Wets (1998), and Knight (1999), can be used to apply the argmax theorem. Fifth, the asymptotic results for $\hat{\gamma}$ can be extended to nonlinear models. For example, suppose

$$y = \begin{cases} m(x,q;\beta_1,\lambda)+\varepsilon_1, & q \le g(z;\gamma), \\ m(x,q;\beta_2,\lambda)+\varepsilon_2, & q > g(z;\gamma), \end{cases} \quad (6)$$
$$\mathbb{E}[\varepsilon_\ell \mid x,q,z] = 0, \ell = 1,2,$$

where λ is the parameters that remain the same in the two regimes, m is a smooth function in the parameters, and g specifies a nonlinear threshold boundary. In this case, the corresponding $D(\cdot)$ for the nonlinear LSE of γ is

$$D(v) = \sum_{i=1}^{\infty} Z_{1i} 1\left(\frac{\partial g(z_i;\gamma_0)}{\partial \gamma'} v < J_{1i} \le 0 \right) + \sum_{i=1}^{\infty} Z_{2i} 1\left(0 < J_{2i} \le \frac{\partial g(z_i;\gamma_0)}{\partial \gamma'} v \right),$$

where $Z_{\ell i}$ and $J_{\ell i}$ take the same form as in Theorem 1 but ϵ_i is redefined as $q_i - g(z_i;\gamma_0)$ and $\bar{Z}_{\ell i}$ is redefined as

$$\bar{Z}_{1i} = 2\left[m(x,q;\beta_{10},\lambda_0) - m(x,q;\beta_{20},\lambda_0)\right]\varepsilon_{1i} + \left[m(x,q;\beta_{10},\lambda_0) - m(x,q;\beta_{20},\lambda_0)\right]^2,$$
$$\bar{Z}_{2i} = -2\left[m(x,q;\beta_{10},\lambda_0) - m(x,q;\beta_{20},\lambda_0)\right]\varepsilon_{2i} + \left[m(x,q;\beta_{10},\lambda_0) - m(x,q;\beta_{20},\lambda_0)\right]^2.$$

3.2 Asymptotics with Shrinking Threshold Effects

In this subsection, we assume $\delta_n = \beta_{10} - \beta_{20}$ shrinks to zero. This framework is suitable to the case where the threshold effects are relatively small for the given sample size. Note that different from what happens in the fixed-threshold-effect framework, the asymptotic distribution of $\hat{\gamma}$ does not depend on which point is taken as $\arg\min_\gamma M_n(\gamma)$.

Theorem 2. *Under Assumptions D1-D8 and* $\|\delta_n\| \to 0$, $\sqrt{n}\|\delta_n\| \to \infty$,

$$a_n\left(\hat{\gamma} - \gamma_0\right) \xrightarrow{d} \arg\min_{v \in \mathbb{R}^{k+1}} C(v),$$

where $a_n = n\delta_n'\delta_n$,

$$C(v) = B(v) + \frac{1}{2}I(v).$$

I(v) and B(v) are defined as follows:

$$I(v) = I_1(v) + I_2(v)$$

which is positive when $v \neq 0$*, where*

$$I_1(v) = \lim_{n\to\infty} \frac{\delta_n'\mathbb{E}\left[\mathbb{E}\left[\mathbf{x}\mathbf{x}' \mid z, \epsilon = 0\right] f_{\epsilon|z}(0 \mid z)\left|\mathbf{z}'v\right| 1(\mathbf{z}'v < 0)\right]\delta_n}{\delta_n'\delta_n},$$

$$I_2(v) = \lim_{n\to\infty} \frac{\delta_n'\mathbb{E}\left[\mathbb{E}\left[\mathbf{x}\mathbf{x}' \mid z, \epsilon = 0\right] f_{\epsilon|z}(0 \mid z)\mathbf{z}'v 1(\mathbf{z}'v > 0)\right]\delta_n}{\delta_n'\delta_n};$$

$$B(v) = B_1(v) + B_2(v)$$

is a Gaussian process with a positive variance when $v \neq 0$*, where* $B_1(v)$ *and* $B_2(v)$ *are two independent Gaussian processes with*

$$Var(B_1(v)) = \lim_{n\to\infty} \frac{\delta_n'\mathbb{E}\left[\mathbb{E}\left[\mathbf{x}\mathbf{x}'\varepsilon_1^2 \mid z, \epsilon = 0\right] f_{\epsilon|z}(0 \mid z)\left|\mathbf{z}'v\right| 1(\mathbf{z}'v < 0)\right]\delta_n}{\delta_n'\delta_n} > 0,$$

$$Var(B_2(v)) = \lim_{n\to\infty} \frac{\delta_n'\mathbb{E}\left[\mathbb{E}\left[\mathbf{x}\mathbf{x}'\varepsilon_2^2 \mid z, \epsilon = 0\right] f_{\epsilon|z}(0 \mid z)\mathbf{z}'v 1(\mathbf{z}'v > 0)\right]\delta_n}{\delta_n'\delta_n} > 0,$$

$$Cov\left(B_1(v_1), B_1(v_2)\right) = \lim_{n\to\infty} \frac{\delta_n'\mathbb{E}\left[\mathbb{E}\left[\mathbf{x}\mathbf{x}'\varepsilon_1^2 \mid z, \epsilon = 0\right] f_{\epsilon|z}(0 \mid z)\left|\mathbf{z}'v_1 \vee \mathbf{z}'v_2\right| 1(\mathbf{z}'v_1 \vee \mathbf{z}'v_2 < 0)\right]\delta_n}{\delta_n'\delta_n},$$

$$Cov\left(B_2(v_1), B_2(v_2)\right) = \lim_{n\to\infty} \frac{\delta_n'\mathbb{E}\left[\mathbb{E}\left[\mathbf{x}\mathbf{x}'\varepsilon_2^2 \mid z, \epsilon = 0\right] f_{\epsilon|z}(0 \mid z)\left(\mathbf{z}'v_1 \wedge \mathbf{z}'v_2\right) 1(\mathbf{z}'v_1 \wedge \mathbf{z}'v_2 > 0)\right]\delta_n}{\delta_n'\delta_n}.$$

for any v, v_1 and v_2 in $\mathbb{R}^{k+1}$. Furthermore, $\hat{\beta}_1$ and $\hat{\beta}_2$ have the same asymptotic distributions as in Theorem 1, and $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\gamma}$ are asymptotically independent of each other.

We provide a few comments on Theorem 2. First, if $\mathbf{z} = 1$, $C(v)$ reduces to

$$C(v) = \begin{cases} \sqrt{f_q(\gamma_0)V_1}B_1(-v) + \frac{f_q(\gamma_0)}{2}D|v|, & \text{if } v \le 0, \\ \sqrt{f_q(\gamma_0)V_2}B_2(v) + \frac{f_q(\gamma_0)}{2}Dv, & \text{if } v > 0, \end{cases}$$

where $D = \lim_{n\to\infty} \delta_n' \mathbb{E}\left[\mathbf{x}\mathbf{x}' \mid q = \gamma_0\right]\delta_n / \delta_n'\delta_n, V_\ell = \lim_{n\to\infty} \delta_n' \mathbb{E}\left[\mathbf{x}\mathbf{x}'\varepsilon_\ell^2 \mid q = \gamma_0\right]\delta_n / \delta_n'\delta_n$, and $B_\ell(v), \ell = 1, 2$, are two independent standard Brownian motions defined on $[0, \infty)$. In other words, $C(v)$ is the two-sided Brownian motion in the usual threshold regression. If we assume $\delta_n = cn^{-\alpha}$ as in Hansen (2000a), then the convergence rate is $n^{1-2\alpha}$. We generalize the setup of δ_n in Hansen (2000a) by allowing for each component of δ_n to converge to zero in different rates and the convergence rates are unknown a priori. Now, the asymptotic distribution of $\hat{\gamma}$ is determined by the components of δ_n with the slowest converging rate to zero. In this way, we do not need to know α in Hansen (2000a) a priori and can estimate the convergence rate by $n\hat{\delta}'\hat{\delta}$ with $\hat{\delta} = \hat{\beta}_1 - \hat{\beta}_2$. Second, scrutinizing $C(v)$, we can see that it cannot be expressed in the form of the two-sided Brownian motion. This is because there are nonconstant covariates in **z** such that for some values of v, some $\mathbf{z}$ satisfy $\mathbf{z}'v < 0$ and others satisfy $\mathbf{z}'v > 0$. As a result, for such v's, both $I_1(v)$ ($B_1(v)$) and $I_2(v)$ ($B_2(v)$) are involved. This is parallel to the fact about $D(v)$ in Theorem 1 that both summations of $D(v)$ are involved for some values of v. Similar to $D(\cdot)$, we call $C(\cdot)$ a "two-sided Brownian field". In some sense, $C(v)$ is like a continuous approximation of $D(v)$; see Yu and Phillips (2018b) for a rigorous statement for such an approximation when $\mathbf{z} = 1$. Third, our assumption that $I(v) > 0$ and $Var(B(v)) > 0$ when $v \neq 0$ does not lose generality. For a given

$v \neq 0$, this can be guaranteed by $P\left(\mathbf{z}'v<0\right)\vee P\left(\mathbf{z}'v>0\right)>0$. If there is multicollinearity among **z** such that for some v_o, $P\left(\mathbf{z}'v_o=0\right)=1$, then both $l(v)$ and $Var(B(v))$ (and thus $C(v)$) are zero at v_o. Since $\min_{v\in\mathbb{R}^{k+1}} C(v)$ is quite possible to be zero, $\arg\min_{v\in\mathbb{R}^{k+1}} C(v)$ is not unique given that $C(av_o)=0$ for any $a\in\mathbb{R}$. This is why Assumption D7 excludes multicollinearity among **z**. Fourth, the results in Theorem 2 can be extended to the nonlinear model (6) by replacing $\partial m(x,q;\beta_{10},\lambda_0)/\partial\beta_1$ for **x**, $\partial g(z;\gamma_0)/\partial\gamma$ for **z** and $q-g(z;\gamma_0)$ for ϵ.

4 Inference Methods for γ

Since the inference for β is standard, we concentrate on the inference for γ in this section. In the usual threshold regression, there are two dominating inference methods for γ; see Section 4.1 of Yu (2014) for a thorough summary on the existing inference methods. The first method is through inverting the likelihood ratio (LR) statistic (see, e.g., Hansen (2000a)); the second method is to use the nonparametric posterior interval (NPI) (see, e.g., Yu (2015)). As argued in Section 4.1 of Hansen (2000a), the straightforward Wald-type confidence set by inverting the asymptotic distribution of $\hat{\gamma}$ in Theorem 2 performs unsatisfactorily in finite samples due to the identification failure when $\delta_n=0$. The difficulties in inverting the asymptotic distribution of $\hat{\gamma}$ in Theorem 1 are discussed in Section 4.1 of Yu (2015).

Although the LR statistic in the usual threshold regression is asymptotically pivotal after some transformation, this is not the case in the general setup. This complication is mainly because of the existence of nonconstant covariates in **z** (such that $B(v)$ and $l(v)$ in $C(v)$ cannot be simplified). As a result, we must simulate a Gaussian process with a complicated covariance kernel to find the asymptotic distribution of the LR statistic. For example, we need to estimate $\mathbb{E}\left[\mathbf{x}\mathbf{x}'\varepsilon_\ell^2 \mid z,\epsilon=0\right], \mathbb{E}\left[\mathbf{x}\mathbf{x}' \mid z,\epsilon=0\right]$ and $f_{\epsilon|z}(0\mid z)$ for all $z\in\mathbb{Z}$ to simulate the

Gaussian process. Fortunately, the NPI can still be applied. Specifically, we use the following algorithm to construct the CI for each component of γ.

Algorithm G:

Step 1: Compute the LSE $\left(\hat{\gamma}', \hat{\beta}'\right)'$ and the corresponding residuals $\hat{\varepsilon}_i = y_i - \mathbf{x}_i'\hat{\beta}_1 1\left(q_i \le \mathbf{z}_i'\hat{\gamma}\right) - \mathbf{x}_i'\hat{\beta}_2 1\left(q_i > \mathbf{z}_i'\hat{\gamma}\right), i = 1, \cdots, n$.

Step 2: Obtain a uniformly consistent estimator of the joint density f of $\mathrm{w} := \left(\varepsilon, x', q, z'\right)'$ by kernel smoothing, and denote the estimator as $\hat{f}(\mathrm{w})$.

Step 3: Define the estimated likelihood function as

$$\begin{aligned}
\mathcal{L}_n(\gamma) &= \prod_{i=1}^{n}\left[\hat{f}\left(y_i - \mathbf{x}_i'\hat{\beta}_1, x_i, q_i, z_i\right)1(q_i \le \mathbf{z}_i'\gamma) + \hat{f}\left(y_i - \mathbf{x}_i'\hat{\beta}_2, x_i, q_i, z_i\right)1(q_i > \mathbf{z}_i'\gamma)\right] \\
&= \exp\left\{\sum_{i=1}^{n} 1(q_i \le \mathbf{z}_i'\gamma)\ln\left(\hat{f}\left(y_i - \mathbf{x}_i'\hat{\beta}_1, x_i, q_i, z_i\right)\right) + \sum_{i=1}^{n} 1(q_i > \mathbf{z}_i'\gamma)\ln\left(\hat{f}\left(y_i - \mathbf{x}_i'\hat{\beta}_2, x_i, q_i, z_i\right)\right)\right\} \\
&:= \exp\left\{\hat{L}_n(\gamma)\right\}.
\end{aligned}$$

Step 4: Draw a Markov chain

$$S = \left(\gamma^{(1)}, \cdots, \gamma^{(B)}\right),$$

whose marginal density is approximately given by

$$\hat{p}_n(\gamma) = \frac{\exp\left\{\hat{L}_n(\gamma)\right\}1(\gamma \in \Gamma)}{\int_\Gamma \exp\left\{\hat{L}_n(\gamma)\right\}d\gamma}.$$

Step 5: Take out the lth component of S, denoted as $S_l := \left(\gamma_l^{(1)}, \cdots, \gamma_l^{(B)}\right)$. Then the $(1-\alpha)$ CI for γ_l is constructed as $[\gamma_{l(\alpha/2)}, \gamma_{l(1-\alpha/2)}]$, where $\gamma_{l(\tau)}$ is the τth quantile of S_l.

When x and/or z contain discrete variables, we can extend the method in Racine and Li (2004) from nonparametric regression to density estimation. Another problem is that when the dimension of x and z is large, the estimation of $\hat{f}$ suffers from the curse of dimensionality, so some simplification in the DGP specification is required. A popular simplification is to assume that ε is independent of $(x', q, z')'$. In such a setup, we need only estimate the density of ε_ℓ based on $\hat{\varepsilon}_{\ell i}$, and denote the estimator as $\hat{f}_\ell$, where $\hat{\varepsilon}_{1i}$ is the $\hat{\varepsilon}_i$ such that $q_i \le \mathbf{z}_i'\hat{\gamma}$, and $\hat{\varepsilon}_{2i}$ is the rest of $\hat{\varepsilon}_i$. $\hat{f}\left(y_i - \mathbf{x}_i'\hat{\beta}_\ell, x_i, q_i, z_i\right)$ in Step 3 is replaced by $\hat{f}_\ell\left(y_i - \mathbf{x}_i'\hat{\beta}_\ell\right)$ in each regime. If the number of data points in either regime is not large enough for the density estimation, we can further assume $\varepsilon_\ell = \sigma_\ell e$, where e is independent of $(x', q, z')'$. In other words, the difference between f_1 and f_2 is completely controlled by σ_1 and σ_2. Now, we need only estimate the density of e based on $\hat{e}_i$, and the estimator is denoted as $\hat{f}_e$, where

$$\hat{e}_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_1} 1\left(q_i \le \mathbf{z}_i'\hat{\gamma}\right) + \frac{\hat{\varepsilon}_i}{\hat{\sigma}_2} 1\left(q_i > \mathbf{z}_i'\hat{\gamma}\right)$$

with $\hat{\sigma}_1 = \sum_{i=1}^n \hat{\varepsilon}_i^2 1\left(q_i \le \mathbf{z}_i'\hat{\gamma}\right) / \sum_{i=1}^n 1\left(q_i \le \mathbf{z}_i'\hat{\gamma}\right)$ and $\hat{\sigma}_2 = \sum_{i=1}^n \hat{\varepsilon}_i^2 1\left(q_i > \mathbf{z}_i'\hat{\gamma}\right) / \sum_{i=1}^n 1\left(q_i > \mathbf{z}_i'\hat{\gamma}\right)$. Then $\hat{f}\left(y_i - \mathbf{x}_i'\hat{\beta}_\ell, x_i, q_i, z_i\right)$ in Step 3 is replaced by $\frac{1}{\hat{\sigma}_\ell}\hat{f}_e\left(\frac{y_i - \mathbf{x}_i'\hat{\beta}_\ell}{\hat{\sigma}_\ell}\right)$. Note also that when ε is independent of $(x', q, z')'$, the discreteness of x and/or z is out of consideration in the estimation of f. In practice, it is strongly suggested to employ this simplification to alleviate the curse of dimensionality. As to the bandwidth selection in the estimation of $\hat{f}_\ell$ and $\hat{f}_e$, see the discussions in Section 6. Finally, uniform consistency of $\hat{f}$ can be shown as in Silverman (1978) or Hansen (2008), and the validity of the NPI can be shown along the line of Theorem 4 of Yu (2015), so the details are omitted.

Another important inference method is proposed by Seo and Linton (2007) in a similar framework as in this paper. As mentioned in the introduction, their method is based on smoothing the least squares objective function and the convergence rate is slower than n. When the threshold is constant, the simulation studies in Yu (2015) show that the performance of their CI for γ is not as good as Hansen's LR-CI or our NPI, while those in Seo and Linton (2007) show that their CI outperforms Hansen's LR-CI for large threshold effects. Seo and Linton (2007) did not study the finite-sample performance of their CI when the threshold is not constant; we will study this scenario in Section 6.2. To our knowledge, the CI based on the SLSE and our NPI are the only two available CIs for γ in the framework of this paper.

A corollary of Algorithm G is the semiparametric empirical Bayes estimator (SEBE) of Yu (2015), e.g., the posterior mean or median based on $\hat{p}_n(\cdot)$. As shown in Yu (2015), the SEBE is an adaptive estimator of γ in the fixed-threshold-effect framework of the usual threshold regression. Such an adaptiveness result is ready to extend to the general case. The efficiency improvement of SEBE relative to the LSE is confirmed in the simulation studies of Section 6.2.

5 Testing For A Threshold Effect

To use the LSE to estimate γ in model (2), we must first guarantee that there indeed exist a threshold effect. For such a testing problem, it is more convenient to reparametrize the model (2) as

$$y = \mathbf{x}'\beta_o + \mathbf{x}'\delta 1(q \le \mathbf{z}'\gamma) + \varepsilon,\ \mathbb{E}\left[\varepsilon \mid x, q, z\right] = 0.$$

where the true threshold parameter γ_0 is unknown, $\beta_o = \beta_2$ and $\delta = \beta_1 - \beta_2$. The null hypothesis is

$$H_0 : \beta_{10} - \beta_{20} = 0 \text{ or } \delta_0 = 0$$

and correspondingly, the alternative is

$$H_1 : \beta_{10} - \beta_{20} \neq 0 \text{ or } \delta_0 \neq 0$$

and the local alternative is

$$H_1^c : \delta_n = \beta_{10} - \beta_{20} = n^{-1/2} c.$$

Under the null, the model is linear, so this is also called a testing for linearity problem. Usually, a Wald-type test is suggested, that is, estimate β_1 and β_2 for each possible sample splitting of $\gamma \in \Gamma$ and then test whether $\beta_1 - \beta_2 = 0$ based on the supremum or average of the Wald test statistics among all $\gamma \in \Gamma$. Specifically, define W_n as a random function on Γ,

$$W_n(\gamma) = \left(\hat{V}_1(\gamma) + \hat{V}_2(\gamma)\right)^{-1/2} \sqrt{n}\left(\hat{\beta}_1(\gamma) - \hat{\beta}_2(\gamma)\right), \gamma \in \Gamma,$$

where $\hat{\beta}_1(\gamma)$ and $\hat{\beta}_2(\gamma)$ are defined in Section 2, and

$$\hat{V}_1(\gamma) = \left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' 1(q_i \leq \mathbf{z}_i'\gamma)\right]^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \hat{\varepsilon}_i^2 1(q_i \leq \mathbf{z}_i'\gamma)\right] \left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' 1(q_i \leq \mathbf{z}_i'\gamma)\right]^{-1},$$

$$\hat{V}_2(\gamma) = \left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' 1(q_i > \mathbf{z}_i'\gamma)\right]^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \hat{\varepsilon}_i^2 1(q_i > \mathbf{z}_i'\gamma)\right] \left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' 1(q_i > \mathbf{z}_i'\gamma)\right]^{-1}.$$

The test statistic is a functional of W_n. In practice, two test statistics are most popular. The first is the Kolmogorov-Smirnov sup-type statistic

$$KS = \sup_{\gamma \in \Gamma} \left\| W_n(\gamma) \right\|^2,$$

and the second is the Cramér–von Mises average-type statistic

$$C_v M = \int_{\Gamma} \left\| W_n(\gamma) \right\|^2 d\gamma.$$

Both test statistics can be written as functionals $g(W_n)$, where $g(\cdot)$ maps functions on Γ to $\mathbb{R}$. A key problem associated with such tests is that $\hat{\beta}_\ell(\gamma)$ for all $\gamma \in \Gamma$ need be computed, which is quite time-consuming especially when the dimension of Γ is large. When the threshold boundary is nonparametric such as $q \le g(z)$, we can think of the dimension of Γ as infinity, and the computational problem is particularly severe; see Yu et al. (2018) for the test for a threshold effect in this nonparametric case. The LR test in Lee et al. (2018) suffers a similar problem (and also different from the Wald test and our score test, their null asymptotic distribution is not a functional of a chi-square process when the error is heteroskedastic so their test may suffer from a power loss). In the following, we propose a score-type test which can avoid this problem.

5.1 Test Construction and Asymptotics

Our test statistics are based on the score of the LS objective function under the null. For the testing purpose, it is more convenient to rewrite the objective function as

$$Q_n(\delta;\gamma,\beta_o) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i'\delta 1(q_i \le \mathbf{z}_i'\gamma) - \mathbf{x}_i'\beta_o\right)^2.$$

The score function of $Q_n(\delta;\gamma,\beta_o)$ with respect to δ and evaluated at $\delta = 0$ is

$$S_n(\gamma) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i 1(q_i \le \mathbf{z}_i'\gamma)\varepsilon_{oi}$$

after discarding the constant terms, where ε_{oi} equals ε_i under H_0 and contains some extra bias under H_1, and can be estimated by $\hat{\varepsilon}_{oi} = y_i - \mathbf{x}_i'\hat{\beta}_o$ with $\hat{\beta}_o$ being the coefficients in the regression of y_i on $\mathbf{x}_i$. Our score-type tests are based on

$$T_n(\gamma)=\left[n^{-1}\sum_{i=1}^{n}\left(\mathbf{x}_i 1(q_i \le \mathbf{z}_i'\gamma)-\hat{Q}_1(\gamma)\hat{Q}^{-1}\mathbf{x}_i\right)\left(\mathbf{x}_i 1(q_i \le \mathbf{z}_i'\gamma)-\hat{Q}_1(\gamma)\hat{Q}^{-1}\mathbf{x}_i\right)' \hat{\varepsilon}_{oi}^2 \right]^{-1/2}$$
$$\cdot n^{-1/2}\sum_{i=1}^{n}\left[\mathbf{x}_i 1(q_i \le \mathbf{z}_i'\gamma)-\hat{Q}_1(\gamma)\hat{Q}^{-1}\mathbf{x}_i\right]\hat{\varepsilon}_{oi}, \gamma \in \Gamma,$$

where $\hat{Q}_1(\gamma)=n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' 1(q_i \le \mathbf{z}_i'\gamma)$ and $\hat{Q}=n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'$. Different from W_n, we need only run one regression of y on $\mathbf{x}$ to construct T_n, so the computation burden is significantly lightened.

Note that our score test is different from the Lagrange multiplier (LM) test in Hansen (1996). The LM test there is still a Wald-type test since the test statistic is constructed under the alternative, and only the residuals in the regression score are constructed under the null. On the other hand, our score test is similar in spirit to the LM test in Hansen (1990) where the test statistic is constructed in the structural change context. Note also that although $\hat{Q}_1(\gamma)\hat{Q}^{-1}n^{-1/2}\sum_{i=1}^{n}\mathbf{x}_i\hat{\varepsilon}_{oi}=o_p(1), \mathbf{x}_i 1(q_i \le \mathbf{z}_i'\gamma)$ is recentered by $\hat{Q}_1(\gamma)\hat{Q}^{-1}\mathbf{x}_i$. This is because the effect of $\hat{\beta}_o$ will not disappear asymptotically so the asymptotic distribution of $n^{-1/2}\sum_{i=1}^{n}\mathbf{x}_i 1(q_i \le \mathbf{z}_i'\gamma)\hat{\varepsilon}_{oi}$ differs from $n^{-1/2}\sum_{i=1}^{n}\mathbf{x}_i 1(q_i \le \mathbf{z}_i'\gamma)\varepsilon_{oi}$ under H_0. Recentering is to offset the effect of $\hat{\beta}_o$.

Given $T_n(\cdot)$, we can similarly construct the Kolmogorov-Smirnov sup-type statistic or the Cramér–von Mises average-type statistic. Define $g_n = g(T_n)$, where g is the functional defined in KS or C_vM. The following theorem states the weak limit of g_n under H_1^c, which implies the asymptotic null distribution and the consistency of the tests.

Theorem 3. *Under* H_1^c,

$$g_n \xrightarrow{d} g_c = g(T^c),$$

where

$$T^c(\gamma) = H(\gamma,\gamma)^{-1/2}\left\{S(\gamma) + \left[Q_1(\gamma \wedge \gamma_0) - Q_1(\gamma)Q^{-1}Q_1(\gamma_0)\right]c\right\},$$

$$Q_1(\gamma \wedge \gamma_0) = \mathbb{E}\left[\mathbf{x}\mathbf{x}'1(q \le \mathbf{z}'\gamma \wedge \mathbf{z}'\gamma_0)\right], Q_1(\gamma) = Q_1(\gamma \wedge \gamma), Q = \mathbb{E}\left[\mathbf{x}\mathbf{x}'\right],$$ *and* $S(\gamma)$ *is a mean zero Gaussian process with covariance kernel*

$$H(\gamma_1,\gamma_2) = \mathbb{E}\left[\left(\mathbf{x}1(q \le \mathbf{z}'\gamma_1) - Q_1(\gamma_1)Q^{-1}\mathbf{x}\right)\left(\mathbf{x}1(q \le \mathbf{z}'\gamma_2) - Q_1(\gamma_2)Q^{-1}\mathbf{x}\right)'\varepsilon^2\right].$$

In some special cases, T_n can be simplified. For example, if $\mathbf{x}_i$ is independent of $(q_i, z_i')'$, then $T_n(\gamma)$ can be simplified as

$$T_n(\gamma) = \left[n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\left(1(q_i \le \mathbf{z}_i'\gamma) - \overline{1(q \le \mathbf{z}'\gamma)}\right)^2\hat{\varepsilon}_{oi}^2\right]^{-1/2} n^{-1/2}\sum_{i=1}^n\left(1(q_i \le \mathbf{z}_i'\gamma) - \overline{1(q \le \mathbf{z}'\gamma)}\right)\mathbf{x}_i\hat{\varepsilon}_{oi},$$

where $\overline{1(q \le \mathbf{z}'\gamma)} = n^{-1}\sum_{i=1}^n 1(q_i \le \mathbf{z}_i'\gamma)$.

5.2 Simulating Critical Values

The asymptotic distribution in Theorem 3 is not pivotal, but the simulation method in Hansen (1996) can be extended to the current case to obtain critical values. Given that T_n is easier to compute than W_n, the simulation method for T_n is also computationally less intensive. More specifically, let $\{\xi_i^*\}_{i=1}^n$ be i.i.d. $N(0,1)$ random variables, and set

$$T_n^*(\gamma) = \left[n^{-1}\sum_{i=1}^n\left(\mathbf{x}_i1(q_i \le \mathbf{z}_i'\gamma) - \hat{Q}_1(\gamma)\hat{Q}^{-1}\mathbf{x}_i\right)\left(\mathbf{x}_i1(q_i \le \mathbf{z}_i'\gamma) - \hat{Q}_1(\gamma)\hat{Q}^{-1}\mathbf{x}_i\right)'\hat{\varepsilon}_{oi}^2\right]^{-1/2} \cdot n^{-1/2}\sum_{i=1}^n\left[\mathbf{x}_i1(q_i \le \mathbf{z}_i'\gamma) - \hat{Q}_1(\gamma)\hat{Q}^{-1}\mathbf{x}_i\right]\hat{\varepsilon}_{oi}\xi_i^*, \gamma \in \Gamma. \quad (7)$$

Our test rejects H_0 if g_n is greater than the $(1-\alpha)$ th conditional quantile of $g(T_n^*)$. Equivalently, the p-value transformation can be employed. Define $p_n^* = 1 - F_n^*(g_n)$, and $p_n = 1 - F_0(g_n)$, where F_n^* is the conditional distribution of $g(T_n^*)$ given the original data, and F_0 is the asymptotic distribution of $g(T_n)$ under the null. Our test rejects H_0 if $p_n^* \le \alpha$. The following theorem states the validity of the above procedure.

Theorem 4. *Under the assumptions of Theorem 3,* $p_n^* = p_n + o_p(1)$ *under both* H_0 *and* H_1^c. *Hence* $p_n^* \xrightarrow{d} p_c = 1 - F_0(g_c)$ *under* H_1^c, *and* $p_n^* \xrightarrow{d} U[0,1]$ *under* H_0.

By stochastic equicontinuity of the $T_n(\gamma)$ process, we can replace Γ by finite grids Γ_n with the distance between adjacent grid points going to zero as $n \to \infty$ (see Section 6 for more implementation details). Also, the conditional distribution can be approximated by standard simulation techniques. More specifically, the following algorithm is used.

Algorithm S:

Step 1: Generate i.i.d. $N(0, 1)$ random variables $\{\xi_{ij}^*\}_{i=1}^n$.

Step 2: Set $T_n^{j*}(\gamma_l)$ as in (7), where $\gamma_l \in \Gamma_n$. Note here that the same $\{\xi_{ij}^*\}_{i=1}^n$ are used for all γ_l, $l = 1, \cdots, L$.

Step 3: Set $g_n^{j*} = g(T_n^{j*})$.

Step 4: Repeat Step 1-3 J times to generate $\{g_n^{j*}\}_{j=1}^J$.

Step 5: If $p_n^{J*} = J^{-1} \sum_{j=1}^J 1(g_n^{j*} \ge g_n) \le \alpha$, we reject H_0; otherwise, accept H_0.

6 Simulations

In this section, we conduct some Monte Carlo simulations to check the performance of the estimators and tests described in the previous sections.

Given that Algorithm M, G and S are quite time-consuming, we will consider only the following simple DGP

$$y = \delta_n 1(q \le \gamma_1 + \gamma_2 z) + \varepsilon,$$

where $\epsilon = q - (\gamma_1 + \gamma_2 z) \sim U[-0.5, 0.5]$, $z \sim U[0,1]$, $\varepsilon \sim N(0,1)$, and $(z, \epsilon, \varepsilon)$ are independent of each other. This setup is the same as (3) except that an error term ε is added in. We use this simple setup to emphasize the new feature of the model in this paper, i.e., the threshold depends on covariates, and neglect the other popular features of TR models, e.g., threshold effects depend on covariates, error variances depend on regimes, etc. The sample size n = 200, and the number of repetitions is set as 500. The true threshold parameter $\gamma_0 = (1,1)'$. In specification testing, $\delta_n = cn^{-1/2}$ with $c = 0,1,\cdots,10$. In estimation, we consider two δ_n's with c = 10 and 20 which roughly correspond to the small-threshold-effect framework in Section 3.2 and the fixed-threshold-effect framework in Section 3.1.

The following simulation study serves three purposes: (i) check the effect of the specification of Γ_n on size and power of the specification tests; (ii) check the performance of our LSE in different frameworks; (iii) check the coverage of the NPI in different frameworks. In other words, we concentrate on the new aspects of the threshold model considered in this paper; other simulation results which are much expected are referred to the existing literature in the references.

We provide more details on our implementation here. First, we specify our Γ as follows. From Section 2, the set of all possible γ's is unbounded, so we first restrict $\gamma_2 \in [0,2]$ – a compact set. Note that this range of γ_2 is actually $[\gamma_2^m - 1, \gamma_2^m + 1]$, where γ_2^m is the slope in the (population) median or mean regression of q on $(1, z)$. We also tried $[\hat{\gamma}_2^m - 1, \hat{\gamma}_2^m + 1]$ as the parameter space for γ_2 and all the simulation results barely change, where $\hat{\gamma}_2^m$ is the estimate of γ_2^m in

sample. As to γ_1, we implicitly restrict it such that each regime contains at least 30 data points. In the test for linearity, we use a MCMC algorithm to simulate from the uniform distribution on Γ (i.e., replace $\exp\{-M_n(\gamma)\}$ in $p_n(\gamma)$ of Algorithm M by 1) and the resulting approximation set of Γ is used as Γ_n. Second, a MCMC algorithm must be employed in simulating Γ_n, $p_n(\gamma)$ in Algorithm M and $\hat{p}_n(\gamma)$ in Algorithm G. We use the Matlab function slicesample for our purpose. In the function slicesample, the arguments are some initial value and a posterior function form that is not necessarily normalized as a density; unlike in the Metropolis-Hastings sampler, the transition probability function is not required. We use the LADE of q on $(1, z)$ as the starting value, and draw 1000 samples from the posterior after discarding the first 200 "burn-in" draws. We refer to Neal (2003) for a concrete description of the slice sampling. Third, in Algorithm G, since ε is independent of (q, z), we need only estimate the density of ε. We use the Matlab function ksdensity to carry out this estimation. The function ksdensity uses by default the standard normal kernel and the optimal bandwidth when the true density is normal to get a smoothing density; see Section 3.4.2 of Silverman (1986) for details of this bandwidth selection. To improve the finite-sample performance, we instead use the Mablab function kde.m provided by Zdravko Botev to select the bandwidth adaptively and then plug in ksdensity to estimate f_ε; see Botev et al. (2010) for the details.

We also compare the performance of the SLSE and the associated CI of Seo and Linton (2007) with our LSE/SEBE and NPI. Note that the objective function of SLSE in our DGP is

$$\tilde{M}_n(\gamma) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \delta_n K\left(\frac{\gamma_1 + \gamma_2 z_i - q_i}{h}\right)\right]^2,$$

which just replaces $1(\gamma_1 + \gamma_2 z_i - q_i \geq 0)$ in $M_n(\gamma)$ by $K\left(\frac{\gamma_1 + \gamma_2 z_i - q_i}{h}\right)$, where as suggested by Seo and Linton (2007), $K(x) = \Phi(x) + x\phi(x)$ with Φ and ϕ being

the standard normal cdf and pdf respectively, and the bandwidth $h = \log n / \sqrt{n}$. Seo and Linton (2007) propose another SLSE, but their theoretical analysis and simulation studies show that this SLSE is better when the above $K(\cdot)$ and h are used. In searching for the minimizer of $\tilde{M}_n(\gamma)$, we can utilize Algorithm M, replacing $M_n(\gamma)$ by $\tilde{M}_n(\gamma)$ and neglecting Steps 4 and 5 since $\arg\min_\gamma \tilde{M}_n(\gamma)$ is unique. However, if we bootstrap the SLSE, then Algorithm M is too time-consuming. As an alternative, we just grid search $\tilde{M}_n(\gamma)$ over 201 × 201 points uniformly distributed on $[0,2]\times[0,2]$. This grid search is possible since $\dim(\gamma) = 2$ in the current DGP. The $K(\cdot)$ function in $\tilde{M}_n(\cdot)$ allows that each regime need not contain at least $k + 1$ data points, which is very different from the LSE. In CI construction, we report only the CIs based on the bootstrap-t method to gain the finite-sample refinement as suggested in Seo and Linton (2007), where the number of bootstrap repetitions is set to be 399. Compared with our LSE/SEBE and NPI, the SLSE and the associated CI have three drawbacks. First, they are less practical. When $\dim(\gamma)$ is large, only Algorithm M can be used, but then only the asymptotic CI can be used while the bootstrap CI is too time-consuming. Second, their performance critically depends on the choice of h which is not easily determined in real applications. Third, the SLSE has a slower convergence rate than our LSE, which implies higher risk and longer CI.

6.1 Testing for Linearity

In the testing for linearity, we let $\#(\Gamma_n)$, the number of γ in Γ_n, be $O(n)$. Three choices are checked – $n/2, n, 2n$. This specification of Γ_n is motivated by the fact that in the usual threshold regression, the number of all possible sample splittings is $O(n)$. This simulation is to check whether different approximations of Γ have significant effects on size and power. Given that the specification of Γ_n may have a large impact on the average-type statistic, we consider only the sup-type statistic here. In Algorithm S, $J = 500$. The size and power are evaluated at

the 5% nominal level. In this simple setup, $\mathbf{x}_i 1(q_i \le \mathbf{z}_i'\gamma) - \hat{Q}_1(\gamma)\hat{Q}^{-1}\mathbf{x}_i$ in T_n reduces to $1(q_i \le \mathbf{z}_i'\gamma) - \overline{1(q \le \mathbf{z}'\gamma)}$, and $\hat{\varepsilon}_{oi}$ reduces to $y_i - n^{-1}\sum_{i=1}^{n} y_i$.

We report the simulation results in the left panel of Figure 4; two results of interest are as follows. First and importantly, $\#(\Gamma_n)$ does not have significant effects on the size and power of our score test. Actually, the power curves associated with the three $\#(\Gamma_n)$ are almost identical. Second, the size of our test matches the nominal level and the power is reasonably good. For comparison, we also report the performance of the sup-Wald test in the right panel of Figure 4. Comparing with the score test, the Wald test is oversized and $\#(\Gamma_n)$ seems to have relatively larger impacts on the power. The phenomenon of oversizedness in the Wald test also appears in, e.g., Hansen (1996), and the power difference between these two tests seems due to the size distortion in the Wald test. Also, the Wald test indeed takes much longer time to execute in our simulation. One practical implication of this simulation is that we do not need to pay much attention to the specification of Γ_n in our score test as long as $\#(\Gamma_n)$ is reasonably large.

6.2 Estimation

In this simple setup, the asymptotic distribution of $\hat{\gamma}$ in Theorems 1 and 2 can be much simplified. In Theorem 1, $J_{\ell i} = \mathcal{J}_{\ell i}$ follows a Gamma distribution since $f_{c|z}(0 \mid z_i) = 1$, $Z_{1i} = \frac{\delta_n}{2} + \varepsilon_i^-$, and $Z_{2i} = \frac{\delta_n}{2} - \varepsilon_i^+$, where ε_i^- and ε_i^+ follow the same distribution as ε. Because z_i is always positive, the number of jumps in the first term of $D(v)$ goes to infinity when $v_1 \wedge v_2 \to -\infty$, and the number of jumps in the second term of $D(v)$ goes to infinity when $v_1 \vee v_2 \to \infty$. In Theorem 2, $I_1(v) = \mathbb{E}\left[|\mathbf{z}'v| 1(\mathbf{z}'v < 0)\right]$, $I_2(v) = \mathbb{E}\left[\mathbf{z}'v 1(\mathbf{z}'v > 0)\right]$, $Var(B_1(v)) = \sigma^2 I_1(v)$, $Var(B_2(v)) = \sigma^2 I_2(v)$, $Cov\left(B_1(v_1), B_1(v_2)\right) = \sigma^2 \mathbb{E}\left[|\mathbf{z}'v_1 \vee \mathbf{z}'v_2| 1(\mathbf{z}'v_1 \vee \mathbf{z}'v_2 < 0)\right]$, and $Cov\left(B_2(v_1), B_2(v_2)\right) = \sigma^2 \mathbb{E}\left[\left(\mathbf{z}'v_1 \wedge \mathbf{z}'v_2\right) 1(\mathbf{z}'v_1 \wedge \mathbf{z}'v_2 > 0)\right]$, where $\sigma^2 = Var(\varepsilon) = 1$.

The RMSEs of the SLSE, LSE and SEBE are reported in Table 1, where we use the posterior mean to represent the SEBE, and the performance of the posterior median is similar. From Table 1, the following conclusions can be drawn. First, as expected, when δ_n gets larger, the risks of all γ estimators are smaller. Second, the risk of the updated estimator $\hat{\gamma}_N$ are smaller than that of the original estimator $\hat{\gamma}_I$, but the difference is only marginal, so at most one updating in Algorithm M is needed in practice. Third, as expected, the LSE has a smaller risk than the SLSE, especially when δ_n is large (when $c = 10$, $\hat{\gamma}_{1,SLSE}$ has a smaller risk than $\hat{\gamma}_{1,I}$, but its risk is similar to $\hat{\gamma}_{1,N}$ and larger than the SEBE). Fourth, the risk of the SEBE are smaller than those of the LSEs and SLSE, especially when δ_n is large. In summary, the SEBE performs the best in all circumstances and is suggested in practice. As to the computational time, we find Algorithm M is faster than Algorithm G; both algorithms take seconds for each repetition.

The length and coverage of the bootstrap-t CI based on the SLSE and our NPI are summarized in Table 2, where the coverage level is set as 95%. From Table 2, a few conclusions can be drawn. First, as expected, when δ_n gets larger, the coverage of both CIs for both γ_1 and γ_2 is better and the length is shorter. Second, the bootstrap- t CI has a similar coverage as the NPI, but is much longer than the latter, which matches the efficiency comparison in Table 1. Third, it seems that the slope γ_2 is harder to estimate than the intercept γ_1 – larger risks and longer CIs.

7 Empirical Applications

In this section, we apply our estimation and testing methods to two examples in labor economics. The first application is about tipping point in dynamic segregation and the second is about CEO compensation; $z \in \mathbb{R}$ in both applications are continuous. These two applications have been among the most debated topics in the general public as well as academics.

7.1 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.

We use Pan (2015) to illustrate the econometric methodology in this paper. The basic empirical specification in Pan (2015) is

$$Dm_{isrj,t} = p\left(f_{isrj,t-10} - f^{*}_{rj,t-10} \right) + d1(f_{isrj,t-10} > f^{*}_{rj,t-10}) + X'_{isrj,t-10}\beta + \varepsilon_{isrj,t}, \qquad (8)$$

where $m_{isrj,t}$ Z $m_{isrj,t}$ 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 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 (2015) adopts a two-step method as in Card et al. (2008). First, she estimates the time specific discontinuity point $f^{*}_{rj,t-10}$ from the data. Specifically, she uses two different methods with similar results; the first is a structural break method as in Hansen (2000a) and the second method is a "fixed-point" procedure suggested in Card et al. (2008). Second, she feeds the estimates of $f^{*}_{rj,t-10}$ to (8) and carries out the regression analysis.

From the results for the pooled sample in Table 6 of Pan (2015), the higher the male prejudice, the lower the tippling point, where higher male prejudice represents more male-prejudiced attitude toward the appropriate role of women among men. So we modify the model (8) as

$$Dm_{isrj,t} = p\left(f_{isrj,t-10}\right) + d1(f_{isrj,t-10} > \mathbf{z}'_{rj,t-10}\gamma) + X'_{isrj,t-10}\beta + \varepsilon_{isrj,t}, \qquad (9)$$

where we estimate the tipping point as a function of $\mathbf{z}_{rj,t-10}$. Following Pan (2015), we include a constant and a male prejudice variable in $\mathbf{z}_{rj,t-10}$.

First, we use our score test to check whether the tipping point indeed exists. It turns out that the *p*-values are all zero for every ten-year period between 1940 and 1990, so there are strong evidences that threshold effects are present. Second, we report our estimates of γ and d. Table 3 contains the γ estimates and the associated 95% NPIs, where only the updated LSE of γ in Algorithm M is reported as the LSE. From Table 3, the estimated threshold location is negatively correlated with the male prejudice for each decade, and the negative correlation is statistically significant. This is consistent with the results in Table 6 of Pan (2015). Table 4 contains the estimates of d in model (9). It estimates a discontinuous decline in net male employment growth at tipping points for all decadal periods. The magnitude and the time trend of the decline are mostly consistent with the results in Table 3 of Pan (2015).

In summary, the results from the application of our methods confirm the tipping behavior of the occupation gender segregation in the labor market, and the tipping points are lower in regions with higher male prejudice.

7.2 Executive Compensation

Executive compensation has been one of the most debated topics among the general public as well as academics (see, e.g., Murphy (1999) and Edmans et al. (2017) for extensive review and discussion). In determining the executive

compensation, the classic principal-agent theory suggests Relative Performance Evaluation (RPE). That is, the risk neutral principal (shareholders) should bear all risks which are out of the executive's control, called "luck" for short. The executive should be compensated for her performance relative to a benchmark to filter out the effects of such luck component. However, some empirical researches find executives are paid for performance beyond their control. Bertrand and Mullainathan (2001) find CEO compensation is positively and significantly correlated with luck. They also find such "pay for luck" might be asymmetric: CEOs might not be punished by bad luck as much as rewarded for good luck. Findings in Garvey and Milbourn (2006) confirm this asymmetry.

Using the methodology developed in this paper, we investigate the pay to luck asymmetry in a more general framework. Specifically, we test and estimate the following econometric model,

$$y = \begin{cases} L\alpha_1 + X'\beta_1 + \varepsilon_1, & L \le \mathbf{z}'\gamma, \\ L\alpha_2 + X'\beta_2 + \varepsilon_2, & L > \mathbf{z}'\gamma, \end{cases} \quad (10)$$

where y is the executive compensation, L is luck, X contains controls for skills, gender, age, tenure and total market value of the firm, and ε_ℓ is the unobservable noise term in each regime. The variable $\mathbf{z}$ can be a constant or a vector specified below. If we restrict $\beta_1 = \beta_2$ and $\mathbf{z}'\gamma = 0$, then it is essentially the model used in Garvey and Milbourn (2006) where they find $\alpha_1 < \alpha_2$.

We use the compensation data drawn from Standard and Poor's ExecuComp on 2306 executives over the 1992-2012 period. We decompose the firm performance P (measured by one-year percentage stock return) into two components, luck and skill, following Bertrand and Mullainathan (2001) and Garvey and Milbourn (2006). As in Garvey and Milbourn (2006) when estimating the executive compensation (10), the dependent variable is the change in the logarithm of total compensation, which includes salary and bonus.

We first use our score test to check whether there are threshold effects for the two specifications of **z**. First, $\mathbf{z} = 1$, i.e., the threshold is constant. The *p*-value is zero in this case. Second, we assume the threshold may vary with the firm size, $\mathbf{z} = (1, z)'$, where z is is the total market value of the firm measured in millions of dollars. The *p*-value is also zero in this case. In both cases, our tests indicate that there are strong threshold effects.

Table 5 reports the point estimates and 95% NPIs of γ for the two specifications of **z**, where we only report the updated LSE of γ in Algorithm M. The results for the first specification are reported in the first column of Table 5. Combined with the first two columns of Table 6, these results are consistent with the asymmetric compensation for luck as found in Bertrand and Mullainathan (2001) and Garvey and Milbourn (2006), i.e., there are two regimes of luck as in (10) such that $\alpha_1 < \alpha_2$. On the other hand, we find the benchmark luck level is 0.108 or 42% of one standard deviation above zero. That is, only if luck is good enough are executives rewarded, so the assumption that the threshold is zero in Garvey and Milbourn (2006) is not justified in our data. The results for the second specification are reported in the second column of Table 5, which shows a negative correlation between the firm size and the threshold, after controlling for the firm fixed effects. Such negative correlation is both statistically and economically significant. When the firm's market value increases one million dollars, the threshold for rewarding good luck decreases by 0.044 or 17.2% of the standard deviation. If the firm's market value increases one standard deviation which is ten million dollars, then the threshold decreases by more than one and a half standard deviation. Bertrand and Mullainathan (2000) discuss a positive correlation between larger firms and poorer governance, while poorer governance indicates more vulnerable to skimming from executives (see, e.g., Bertrand and Mullainathan (2001)). Our results are consistent with their discussions and findings.

Table 6 reports the point estimates and 95% CIs for $(\alpha_\ell, \beta_\ell')$ under the two specifications of $\mathbf{z}$, where the splitting is based on the posterior mean of γ estimation. We will not discuss the details of these estimates but emphasize that different from the specification in Garvey and Milbourn (2006), β_1 and β_2 are quite different in both specifications of $\mathbf{z}$. Rigorously, we test whether $\beta_1 = \beta_2$ using the Wald test. The p-values in both specifications are zero, so incorporating the threshold effects in X is necessary in this study. We further test $\alpha_1 = \alpha_2$ against $\alpha_1 < \alpha_2$ using the t-test; the resulting p-values are 0.0047 and 0.0013, respectively, i.e., α_1 is indeed less than α_2, which confirms the result in Garvey and Milbourn (2006).

In summary, our results confirm the existence of asymmetric benchmark in the executive compensation and find this asymmetric benchmark is significantly higher than the presumed level of zero in the previous literature. We also find a statistically significant and economically large difference in this benchmark by the firm size, which could be explained by the skimming model.

8 Conclusion

This paper discusses the computation, estimation, inference and specification testing in threshold regression with a threshold boundary. The contribution of this paper is better understood in comparison with the results in the usual threshold regression. First, different from the usual threshold regression, computation of the threshold boundary is nontrivial, so we develop an algorithm for this purpose. Second, the asymptotic distribution of the LSE is not related to a compound Poisson process or a two-sided Brownian motion as in the usual threshold regression, but an extension of them – a compound Poisson field or a two-sided Brownian field. Third, unlike in the usual threshold regression, the method of inverting the LR statistics is not easy to apply in constructing confidence sets for the threshold parameters, while the NPI is still applicable. Fourth, in specification testing, the computational burden of the Wald-type test is much heavier in the

general threshold regression, so we develop a score-type test to alleviate the problem.

There are many interesting extensions of the model studied in this paper which are not covered due to space limitations. We list only a few here. First, the analysis for the LSE in this paper can be easily extended to the maximum likelihood estimator of Yu (2012) and the integrated quantile threshold regression estimator of Yu (2013). Second, combining with the analysis in Porter and Yu (2015), Yu and Phillips (2018a) and Yu et al. (2018), we can estimate the nonparametric threshold regression with a nonparametric threshold boundary,

$$y = \begin{cases} m_1(x,q) + \varepsilon_1, & q \le g(z), \\ m_2(x,q) + \varepsilon_2, & q > g(z). \end{cases}$$
$$\mathbb{E}[\varepsilon_\ell \mid x, q, z] = 0, \ell = 1, 2;$$

see Knight (2001) for estimation of the usual nonparametric boundary. Note that if we linearly approximate *g* in each neighborhood of *z*, then the model locally has a parametric linear boundary as discussed in this paper. On the other hand, different from the parametric model in this paper, we need to carefully handle the bias in linearly approximating $g(\cdot)$ locally; see Wang and Lee (2019) for a detailed analysis where $g(\cdot)$ is locally approximated by a constant. Third, we can extend our analysis to the case with multiple boundaries. For illustration, suppose there are only two boundaries. Then these two boundaries take the form $q_\ell = \mathbf{z}_\ell' \gamma_\ell, \ell = 1, 2$. If $q_1 = q_2$ and $\mathbf{z}_1 = \mathbf{z}_2$, then this is a natural extension of Bai (1997) and Bai and Perron (1998). If $q_1 \neq q_2$, i.e., there are two threshold variables, then this is an extension of Chen et al. (2012) and Chong and Yan (2015). Fourth, our arguments can be extended to the threshold autoregressive (TAR) model with a threshold boundary. For example, the TAR(1) model with a threshold boundary is like

$$y_t = \begin{cases} \beta_{11} + \beta_{12} y_{t-1} + \varepsilon_{1t}, & y_{t-1} \le \gamma_1 + \gamma_2 y_{t-2}, \\ \beta_{21} + \beta_{22} y_{t-1} + \varepsilon_{2t}, & y_{t-1} > \gamma_1 + \gamma_2 y_{t-2}, \end{cases}$$

where $\varepsilon_{\ell t}, t=1,\cdots,T$, are i.i.d. innovations. Fifth, the boundary setup of our model can be used in an alternative of threshold regression – smooth transition regression (see Teräsvirta (1998), Teräsvirta et al. (2010) and van Dijk et al. (2002) for surveys). For example, the transition function can take the form $G\left(\frac{q-\mathbf{z}'\gamma}{\eta}\right)$, where G is a cumulative distribution function. As $\eta\to\infty$, this smooth transition model reduces to the sharp transition model in this paper. Sixth, we did not discuss how to select relevant threshold variables z among many potential variables. In practice, the choice of z is usually based on economic intuition. A statistical method such as an information criterion or a penalization approach to choose z is an interesting research topic. Seventh, we did not consider how to conduct inference when q and z are estimated rather than observed. Now, the threshold boundary $q\le\mathbf{z}'\gamma$ would contain an estimation noise which makes the asymptotics more difficult. For these last two issues, see Lee et al. (2018).

References

Abrevaya, J. and J. Huang, 2005, On the Bootstrap of the Maximum Score Estimator, *Econometrica*, 73, 1175-1204.

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

Andrews, D.W.K., 1994, Empirical Process Methods in Econometrics, *Handbook of Econometrics*, Vol. 4, R.F. Eagle and D.L. McFadden, eds., New York: Elsevier Science B.V., Ch. 37, 2247- 2294.

Andrews, D.W.K. and W. Ploberger, 1994, Optimal Tests when a Nuisance Parameter is Present Only Under an Alternative, *Econometrica*, 62, 1383-1414.

Bai, J., 1997, Estimating Multiple Breaks One At a Time, *Econometric Theory*, 13, 315-352.

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

Bertrand, M. and S. Mullainathan, 2000, Agents With and Without Principals, *American Economic Review*, 90, 203-208.

Bertrand, M. and S. Mullainathan, 2001, Are CEOs Rewarded for Luck? The Ones Without Principals Are, *Quarterly Journal of Economics*, 116, 901-932.

Billingsley, P., 1968, *Convergence of Probability Measures*, New York: Wiley.

Botev, Z.I., J.F. Grotowski and D.P. Kroese, 2010, Kernel Density Estimation via Diffusion, *The Annals of Statistics*, 38, 2916–2957.

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

Chan, K.S., 1993, Consistency and Limiting Distribution of the Least Squares Estimator of a Threshold Autoregressive Model, *The Annals of Statistics*, 21, 520-533.

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, H., T.T.-L. Chong and J. Bai, 2012, Theory and Applications of TAR Model with Two Threshold Variables, *Econometric Reviews*, 31, 142-170.

Chernozhukov, V. and H. Hong, 2003a, An MCMC Approach to Classical Estimation, *Journal of Econometrics*, 115, 293-346.

Chernozhukov, V. and H. Hong, 2003b, Likelihood Estimation and Inference in a Class of nonregular Econometric Models, MIT Department of Economics Working Paper.

Chernozhukov, V. and H. Hong, 2004, Likelihood Estimation and Inference in a Class of nonregular Econometric Models, *Econometrica*, 1445-1480.

Chong, T.T.L. and I.K.M. Yan, 2015, A New Threshold Regression Approach to Predict Currency Crises, mimeo.

Davies, R.B., 1977, Hypothesis Testing when a Nuisance Parameter is Present only Under the Alternative, *Biometrika*, 64, 247-254.

Davies, R.B., 1987, Hypothesis Testing when a Nuisance Parameter is Present only Under the Alternative, *Biometrika*, 74, 33-43.

Edmans, A., X. Gabaix, and D. Jenter, 2017, Executive Compensation: A Survey of Theory and Evidence, mimeo.

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

Garvey, G.T. and T.T. Milbourn, 2006, Asymmetric Benchmarking in Compensation: Executives are Rewarded for Good Luck but not Penalized for Bad, *Journal of Financial Economics*, 82, 197-225.

Geyer, C.J., 1994, On the Asymptotics of Constrained M-Estimation, *The Annals of Statistics*, 22, 1993-2010.

Hansen, B.E., 1990, Lagrange Multiplier Tests for Parameter Instability in Non-Linear Models, mimeo.

Hansen, B.E., 1996, Inference when a Nuisance Parameter is not Identified under the Null Hypothesis, *Econometrica*, 64, 413-430.

Hansen, B.E., 2008, Uniform Convergence Rates for Kernel Estimation with Dependent Data, *Econometric Theory*, 24, 726-748.

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

Hansen, B.E., 2000b, Testing for Structural Change in Conditional Models, *Journal of Econometrics*, 97, 93-115.

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.

Hirano, K. and J. R. Porter, 2003, Asymptotic Efficiency in Parametric Structural Models with Parameter-dependent Support, *Econometrica*, 71, 1307-1338.

Jun, S.J., J. Pinkse and Y. Wan, 2015, Classical Laplace Estimation for $\sqrt[3]{n}$ -consistent estimators: Improved Convergence Rates and Rate-Adaptive Inference, *Journal of Econometrics*, 187, 201-216.

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

Klein, R.W. and R.H. Spady, 1993, An Efficient Semiparametric Estimator for Binary Response Models, *Econometrica*, 61, 387-421.

Knight, K., 1999, Epi-convergence and Stochastic Equisemicontinuity, mimeo.

Knight, K., 2001, Limiting Distributions of Linear Programming Estimators, *Extremes*, 4, 87–103.

Knight, K., 2006, Asymptotic Theory for M-Estimators of Boundaries, *The Art of Semiparametric*, Heidelberg: Physica-Verlag.

Lee, S., Y. Liao, M.H. Seo and Y. Shin, 2018, Factor-Driven Two-Regime Regression, mimeo.

Meyer, R.M., 1973, A Poisson-Type Limit Theorem for Mixing Sequences of Dependent Rare Events, *The Annals of Probability*, 1, 480-483.

Mincer, J., 1974, *Schooling, Experience and Earnings*, New York: National Bureau of Economic Research.

Murphy, K.J., 1999, Executive Compensation, in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics*, Vol. 3, Elsevier, Chapter 38, 2485-2563.

Neal, R.M., 2003, Slice Sampling, *The Annals of Statistics*, 31, 705-767.

Newey, W.K. and D.L. McFadden, 1994, Large Sample Estimation and Hypothesis Testing, *Handbook of Econometrics*, Vol. 4, R.F. Eagle and D.L. McFadden, eds., Elsevier Science B.V., Ch. 36, 2113-2245.

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

Pakes, A., and D. Pollard, 1989, Simulation and the Asymptotics of Optimization Estimators, *Econometrica*, 57, 1027-1057.

Porter, J. and P. Yu, 2015, Regression Discontinuity with Unknown Discontinuity Points: Testing and Estimation, *Journal of Econometrics*, 189, 132-147

Potter, S.M., 1995, A Nonlinear Approach to U.S. GNP, *Journal of Applied Econometrics*, 2, 109-125.

Racine, J. and Q. Li, 2004, Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data, *Journal of Econometrics*, 119, 99-130.

Resnick, S.I., 1986, Point Processes, Regular Variation and Weak Convergence, *Advances in Applied Probability*, 18, 66-138.

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

Robert, C.P. and G. Casella, 2004, *Monte Carlo Statistical Methods*, 2nd edition, New York: Springer.

Rockafellar, R.T. and R.J.B. Wets, 1998, *Variational Analysis*, Berlin: Springer.

Rubinstein, R.Y. and D.P. Kroese, 2017, *Simulation and the Monte Carlo Method*, 3rd edition, Hoboken, N.J.: Wiley.

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

Seo, M.H. and O. Linton, 2007, A Smoothed Least Squares Estimator for Threshold Regression Models, *Journal of Econometrics*, 141, 704-735.

Silverman, B.W., 1978, Weak and Strong Uniform Consistency of the Kernel Estimate of a Density and Its Derivatives, *The Annals of Statistics*, 6, 177-184.

Silverman, B.W., 1986, *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

Teräsvirta, T., 1998, Modeling Economic Relationships with Smooth Transition Regressions, *Handbook of Applied Economic Statistics*, A. Ullah and D.E.A. Giles, eds., New York: Dekker, 507-552.

Teräsvirta, T., D. Tjøstheim, and C.W.J. Granger, 2010, *Modelling Nonlinear Economic Time Series*, Oxford: Oxford University Press.

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

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

Van Dijk, D., T. Teräsvirta, and P.H. Franses, 2002, Smooth Transition Autoregressive Models – A Survey of Recent Developments, *Econometric Reviews*, 21, 1-47.

Wang, Y. and Y. Lee, 2019, Nonparametric Sample Splitting, mimeo.

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

Yu, P., 2013, Integrated Quantile Threshold Regression and Distributional Threshold Effects, mimeo.

Yu, P., 2014, The Bootstrap in Threshold Regression, *Econometric Theory*, 30, 676-714.

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

Yu. P., Q. Liao and P.C.B. Phillips, 2018, Inferences and Specification Testing in Threshold Regression with Endogeneity, mimeo.

Yu, P. and P.C.B. Phillips, 2018a, Threshold Regression with Endogeneity, *Journal of Econometrics*, 203, 50-68.

Yu, P. and P.C.B. Phillips, 2018b, Threshold Regression Asymptotics: From Compound Poisson Process to Two-Sided Brownian Motion, *Economics Letters*, 172, 123-126.

Yu, P. and Y.Q. Zhao, 2013, Asymptotics for Threshold Regression Under General Conditions, *Econometrics Journal*, 16, 430-462.

Appendix: Mathematical Proofs

First, some notations are collected for reference in all lemmas and proofs. $a_n = n\delta_n'\delta_n$. P_n is the empirical measure.

$$m(w|\theta) = \left(y - \mathbf{x}'\beta_1 1(q \le \mathbf{z}'\gamma) - \mathbf{x}'\beta_2 1(q > \mathbf{z}'\gamma)\right)^2.$$
$$Q_n(\theta) = P_n\left(m(\cdot|\theta)\right),\ Q(\theta) = P\left(m(\cdot|\theta)\right),\ \mathbb{G}_n m = \sqrt{n}\left(Q_n - Q\right).$$
$$\bar{Z}_1\left(w|\beta_2,\tilde{\beta}_1\right) = \left(\tilde{\beta}_1 - \beta_2\right)' \mathbf{x}\mathbf{x}'\left(\tilde{\beta}_1 - \beta_2\right) + 2\left(\tilde{\beta}_1 - \beta_2\right)\mathbf{x}\varepsilon_1, \text{ so } \bar{Z}_{1i} = \bar{Z}_1\left(w_i|\beta_{20},\beta_{10}\right).$$
$$\bar{Z}_2\left(w|\beta_1,\tilde{\beta}_2\right) = \left(\tilde{\beta}_2 - \beta_1\right)' \mathbf{x}\mathbf{x}'\left(\tilde{\beta}_2 - \beta_1\right) + 2\left(\tilde{\beta}_2 - \beta_1\right)\mathbf{x}\varepsilon_2, \text{ so } \bar{Z}_{2i} = \bar{Z}_2\left(w_i|\beta_{10},\beta_{20}\right).$$

The following formulas are used repetitively in the following analysis:

$$m(w|\theta) = \left(\mathbf{x}'\left(\beta_{10} - \beta_1\right) + \varepsilon_1\right)^2 1(q \le \mathbf{z}'\gamma \wedge \mathbf{z}'\gamma_0) + \left(\mathbf{x}'\left(\beta_{20} - \beta_2\right) + \varepsilon_2\right)^2 1(q > \mathbf{z}'\gamma \vee \mathbf{z}'\gamma_0)$$
$$+\left(\mathbf{x}'\left(\beta_{10} - \beta_2\right) + \varepsilon_1\right)^2 1(\mathbf{z}'\gamma \wedge \mathbf{z}'\gamma_0 < q \le \mathbf{z}'\gamma_0) + \left(\mathbf{x}'\left(\beta_{20} - \beta_1\right) + \varepsilon_2\right)^2 1(\mathbf{z}'\gamma_0 < q \le \mathbf{z}'\gamma \vee \mathbf{z}'\gamma_0),$$

so

$$\begin{aligned}
&\mathrm{m}(\mathrm{w}|\theta)\text{-}\mathrm{m}(\mathrm{w}|\theta_0) = \left[\left(\beta_{10}\text{-}\beta_1\right)' \mathbf{x}\mathbf{x}'\left(\beta_{10}\text{-}\beta_1\right) + 2\left(\beta_{10}\text{-}\beta_1\right)' \mathbf{x}\varepsilon_1\right] 1(\mathrm{q} \le \mathbf{z}'\gamma \wedge \mathbf{z}'\gamma_0) \\
&+\left[\left(\beta_{20}\text{-}\beta_2\right)' \mathbf{x}\mathbf{x}'\left(\beta_{20}\text{-}\beta_2\right) + 2\left(\beta_{20}\text{-}\beta_2\right)' \mathbf{x}\varepsilon_2\right] 1(\mathrm{q} > \mathbf{z}'\gamma \vee \mathbf{z}'\gamma_0) \\
&+\bar{\mathrm{Z}}_1\left(\mathrm{w}|\beta_2,\beta_{10}\right) 1(\mathbf{z}'\gamma \wedge \mathbf{z}'\gamma_0 < \mathrm{q} \le \mathbf{z}'\gamma_0) + \bar{\mathrm{Z}}_2\left(\mathrm{w}|\beta_1,\beta_{20}\right) 1(\mathbf{z}'\gamma_0 < \mathrm{q} \le \mathbf{z}'\gamma \vee \mathbf{z}'\gamma_0) \\
&:= \mathrm{T}\left(\mathrm{w}|\beta_1,\beta_{10}\right) 1(\mathrm{q} \le \mathbf{z}'\gamma \wedge \mathbf{z}'\gamma_0) + \mathrm{T}\left(\mathrm{w}|\beta_2,\beta_{20}\right) 1(\mathrm{q} > \mathbf{z}'\gamma \vee \mathbf{z}'\gamma_0) \\
&+\bar{\mathrm{Z}}_1\left(\mathrm{w}|\beta_2,\beta_{10}\right) 1(\mathbf{z}'\gamma \wedge \mathbf{z}'\gamma_0 < \mathrm{q} \le \mathbf{z}'\gamma_0) + \bar{\mathrm{Z}}_2\left(\mathrm{w}|\beta_1,\beta_{20}\right) 1(\mathbf{z}'\gamma_0 < \mathrm{q} \le \mathbf{z}'\gamma \vee \mathbf{z}'\gamma_0) \\
&:= \mathrm{A}(\mathrm{w}|\theta) + \mathrm{B}(\mathrm{w}|\theta) + \mathrm{C}(\mathrm{w}|\theta) + \mathrm{D}(\mathrm{w}|\theta).
\end{aligned}$$

Proof of Theorem 1. We use Theorem 1 of Knight (1999) to show this theorem. From Lemma 5,

$$Z_n := \left(Z_{\gamma n}, Z_{\beta n}\right) = \arg\min_{(v,u)} \left\{ D_n(v) + u_1' \mathbb{E}\left[\mathbf{x}\mathbf{x}' 1\left(\epsilon \le 0\right)\right] u_1 + u_2' \mathbb{E}\left[\mathbf{x}\mathbf{x}' 1\left(\epsilon > 0\right)\right] u_2 - 2W_n(u) + o_p(1) \right\},$$

where $u = (u_1, u_2), Z_{\beta n} = \left(Z_{\beta_1 n}, Z_{\beta_2 n}\right)$, and $W_n(u) = W_{1n}(u_1) + W_{2n}(u_2)$ with $W_{1n}\left(u_1\right) = u_1'\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{x}_i \varepsilon_{1i} 1\left(\epsilon_i \le 0\right)\right)$ and $W_{2n}\left(u_2\right) = u_2'\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{x}_i \varepsilon_{2i} 1\left(\epsilon_i > 0\right)\right)$. It will be proved that

$$Z_n \xrightarrow{d} Z := \left(Z_\gamma, Z_\beta\right)$$

with $Z_\beta = \left(Z_{\beta_1}, Z_{\beta_2}\right)$. We need to check three conditions to finish the proof.

(a) epi-convergence in distribution of $D_n(v) - 2W_n(u)$ to its finite-dimensional limit $D(v) - 2W(u)$, where

$$W(u) = u_1' \mathbb{E}\left[\mathbf{x}\mathbf{x}' 1\left(\epsilon \le 0\right)\right] Z_{\beta_1} + u_2' \mathbb{E}\left[\mathbf{x}\mathbf{x}' 1\left(\epsilon > 0\right)\right] Z_{\beta_2}.$$

(b) $Z_n = O_p(1)$, and

(c) uniqueness of *Z*.

(b) is proved in Lemma 3, and the uniqueness is implied by Assumption D8 and discussed in the main text. The most difficult task is to show (a) which we now turn to. The proof idea is inspired by Part II of the Technical Addendum in Chernozhukov and Hong (2003b), but is simpler because of the speciality of the threshold model. Because $D_n(v) - 2W_n(u)$ is separable in *u* and *v*, the marginal epi-convergence in distribution of $D_n(v)$ and $W_n(u)$ implies the joint epi-convergence. From the form of $W_n(u)$, it converges in distribution to $W(u)$ with respect to the topology of uniform convergence on compacts, and $W(u)$ is continuous. Such a uniform convergence in distribution implies the epi-convergence in distribution; see the discussion on page 5 of Knight (1999) for more details. If we can show $D_n(v)$ epi-converges in distribution to $D(v)$, by

Theorem 4 of Knight (1999), (a) is proved. In Chernozhukov and Hong (2003b), the parameters of the regular and nonregular components have some overlaps, so the proof there is messier.

The remaining task is to show that $D_n(v)$ epi-converges in distribution to $D(v)$. Recall that D_n epi-converges in distribution to D if for any closed rectangles $R_1,\cdots,R_K$ in $\mathbb{R}^{k+1}$ with open interiors $R_1^o,\cdots,R_K^o$, and any real $r_1,\cdots,r_K$:

$$
\begin{aligned}
&P\left(\inf_{v\in R_1} D(v) > r_1,\cdots,\inf_{v\in R_K} D(v) > r_K\right)\\
&\overset{(1)}{\le} \underline{\lim}_{n\to\infty} P\left(\inf_{v\in R_1} D_n(v) > r_1,\cdots,\inf_{v\in R_K} D_n(v) > r_K\right)\\
&\overset{(2)}{\le} \overline{\lim}_{n\to\infty} P\left(\inf_{v\in R_1^o} D_n(v) \ge r_1,\cdots,\inf_{v\in R_K^o} D_n(v) \ge r_K\right)\\
&\overset{(3)}{\le} P\left(\inf_{v\in R_1^o} D(v) \ge r_1,\cdots,\inf_{v\in R_K^o} D(v) \ge r_K\right),
\end{aligned}
$$

where (2) is implied by the lower-semi-continuity of D_n (recall that a function f is lower semi-continuous if $f(x)\le \underline{\lim}_{n\to\infty} f(x_n)$ for any sequence $\{x_n\}$ whose limit is x. Lower-semi-continuity of D_n is to guarantee that $\arg\min_v D_n(v)$ is well defined if the minimizer is unique. But this is not the case in our setup.). Note that the original D_n need not be lower semi-continuous (l-sc). However, from the definition of our LSE, we can adjust the values of D_n on its jumping locations to make it l-sc without affecting $\arg\min_v D_n(v)$. Because we have already proved the fidi-convergence of $D_n(v)$ in Lemma 7, by Theorem 2 of Knight (1999), we need only prove that D_n is stochastically equi-lower-semicontinuous (e-l-sc). Roughly speaking, stochastic equi-lower-semicontinuity allows us to approximate the distribution of the infimum of D_n over a bounded set B by the distribution of the minimum of D_n over an approximate fixed finite set contained in B for all n sufficiently large (as analogs, the epi-convergence in distribution is the counterpart of the weak convergence when the empirical process is not continuous, and the stochastic equi-lower-semicontinuity plays the role of the

stochastic equicontinuity.). Such a result is given in Lemma 9. More specifically, for each bounded R and $\delta > 0$, there exists a set of fixed points $\{v_{kj}\}$ and neighborhoods $\{V_{kj}\}$ which "center" at $\{v_{kj}\}$ and cover R such that

$$\overline{\lim_{n\to\infty}} P\left(\bigcup_{k,j}\left\{\inf_{v\in V_{kj}} D_n(v) \neq D_n(v_{kj})\right\}\right) < \delta. \qquad (11)$$

For a fixed R and $\delta > 0$, we can always pick $\varphi(\delta)$ in Lemma 9 small enough such that (11) holds, where φ is the sup-distance between the adjacent points of $\{v_{kj}\}$. □

Proof of Theorem 2. The consistency of $\hat{\gamma}$ is proved in Lemma 2, and the convergence rate is shown in Lemma 4. From Lemma 6, $a_n(\hat{\gamma} - \gamma_0)$ has the same asymptotic distribution as $\arg\min_v C_n(v)$, where $C_n(v)$ is defined in (14). We now apply Theorem 2.7 of Kim and Pollard (1990) to find the asymptotic distribution of $a_n(\hat{\gamma} - \gamma_0)$. We need only check the first two conditions of their Theorem 2.7 since the third condition automatically holds.

(i) $C_n(v) \rightsquigarrow C(v) \in \mathbf{C}_{\min}\left(\mathbb{R}^{k+1}\right)$, where $\mathbf{C}_{\min}\left(\mathbb{R}^{k+1}\right)$ is defined as the subset of continuous functions $g(\cdot) \in \mathbf{B}_{\text{loc}}\left(\mathbb{R}^{k+1}\right)$ for which (i) $g(t) \to \infty$ as $\|t|\| \to \infty$ and (ii) $g(t)$ achieves its minimum at a unique point in $\mathbb{R}^{k+1}$, and $\mathbf{B}_{\text{loc}}(\mathbb{R}^{k+1})$ is the space of all locally bounded real functions on $\mathbb{R}^{k+1}$, endowed with the uniform metric on compacta. The weak convergence is proved in Lemma 8. We now check $C(v) \in \mathbf{C}_{\min}\left(\mathbb{R}^{k+1}\right)$. It is not hard to check $C(v)$ is continuous, has a unique minimum (see Lemma 2.6 of Kim and Pollard (1990)), and $\lim_{\|v\|\to\infty} C(v) = \infty$ almost surely (which is true since $B(v)$ is stochastically similar to a two-sided Brownian motion indexed by $\|v|\|$ and $I(v) = O(\|v|\|)$, and for a Brownian motion $W(v)$,

$\lim_{v\to\infty} W(v)/v = 0$ almost surely by virtue of the law of the iterated logarithm for Brownian motion).

(ii) $a_n(\hat{\gamma} - \gamma_0) = O_p(1)$. This is proved in Lemma 4. □

Proof of Theorem 3. First, under H_1^c,

$$
\begin{aligned}
&n^{-1/2}\sum_{i=1}^{n}\mathbf{x}_i 1(q_i \le \mathbf{z}_i'\gamma)\hat{\varepsilon}_{oi} = n^{-1/2}\sum_{i=1}^{n}\mathbf{x}_i 1(q_i \le \mathbf{z}_i'\gamma)\left(y_i - \mathbf{x}_i'\hat{\beta}_o\right)\\
&= n^{-1/2}\sum_{i=1}^{n}\mathbf{x}_i 1(q_i \le \mathbf{z}_i'\gamma)\left(y_i - \mathbf{x}_i'\beta_o\right) - n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' 1(q_i \le \mathbf{z}_i'\gamma)\sqrt{n}\left(\hat{\beta}_o - \beta_o\right)\\
&= n^{-1/2}\sum_{i=1}^{n}\mathbf{x}_i 1(q_i \le \mathbf{z}_i'\gamma)\left(\mathbf{x}_i'\delta_n 1(q_i \le \mathbf{z}_i'\gamma_0) + \varepsilon_i\right) - \hat{Q}_1(\gamma)\sqrt{n}\left(\hat{\beta}_o - \beta_o\right),
\end{aligned}
$$

where $n^{-1/2}\sum_{i=1}^{n}\mathbf{x}_i 1(q_i \le \mathbf{z}_i'\gamma)\mathbf{x}_i'\delta_n 1(q_i \le \mathbf{z}_i'\gamma_0) \xrightarrow{p} Q_1(\gamma\wedge\gamma_0)c$ uniformly in $\gamma\in\Gamma$, and $\hat{Q}_1(\gamma)\xrightarrow{p} Q_1(\gamma)$ uniformly in $\gamma\in\Gamma$. Next,

$$
\begin{aligned}
&n^{-1/2}\textstyle\sum_{i=1}^{n}\hat{Q}_1(\gamma)\hat{Q}^{-1}\mathbf{x}_i\hat{\varepsilon}_{oi}\\
&= \hat{Q}_1(\gamma)\hat{Q}^{-1}n^{-1/2}\textstyle\sum_{i=1}^{n}\mathbf{x}_i\left(-\mathbf{x}_i'(\hat{\beta}_o - \beta_o) + \mathbf{x}_i'\delta_n 1(q_i \le \mathbf{z}_i'\gamma_0) + \varepsilon_i\right)\\
&= -\hat{Q}_1(\gamma)\sqrt{n}\left(\hat{\beta}_o - \beta_o\right) + \hat{Q}_1(\gamma)\hat{Q}^{-1}\hat{Q}_1(\gamma_0)c + \hat{Q}_1(\gamma)\hat{Q}^{-1}\left(n^{-1/2}\textstyle\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i\right),
\end{aligned}
$$

where the second term in the last equality converges in probability to $Q_1(\gamma)Q^{-1}Q_1(\gamma_0)c$ uniformly in $\gamma\in\Gamma$. In summary,

$$
\begin{aligned}
&n^{-1/2}\sum_{i=1}^{n}\left[\mathbf{x}_i 1(q_i \le \mathbf{z}_i'\gamma) - \hat{Q}_1(\gamma)\hat{Q}^{-1}\mathbf{x}_i\right]\hat{\varepsilon}_{oi}\\
&= n^{-1/2}\sum_{i=1}^{n}\left[\mathbf{x}_i 1(q_i \le \mathbf{z}_i'\gamma) - Q_1(\gamma)Q^{-1}\mathbf{x}_i\right]\varepsilon_i + \left[Q_1(\gamma\wedge\gamma_0) - Q_1(\gamma)Q^{-1}Q_1(\gamma_0)\right]c + o_p(1)\\
&\rightsquigarrow S(\gamma) + \left[Q_1(\gamma\wedge\gamma_0) - Q_1(\gamma)Q^{-1}Q_1(\gamma_0)\right]c.
\end{aligned}
$$

It is standard to show that under H_1^c, $\hat{\beta}_o$ is consistent and

$$H_n(\gamma_1,\gamma_2)=n^{-1}\sum_{i=1}^{n}\left(\mathbf{x}_i 1(q_i\le \mathbf{z}_i'\gamma_1)-\hat{Q}_1(\gamma_1)\hat{Q}^{-1}\mathbf{x}_i\right)\left(\mathbf{x}_i 1(q_i\le \mathbf{z}_i'\gamma_2)-\hat{Q}_1(\gamma_2)\hat{Q}^{-1}\mathbf{x}_i\right)'\hat{\varepsilon}_{oi}^2$$
$$\xrightarrow{p} H(\gamma_1,\gamma_2)$$

uniformly over $(\gamma_1,\gamma_2)\in\Gamma\times\Gamma$, which implies $H_n(\gamma,\gamma)\xrightarrow{p}H(\gamma,\gamma)$ uniformly over $\gamma\in\Gamma$ under H_1^c, so the results of the theorem follow. □

Proof of Theorem 4. Conditional on the original sample path, $n^{-1/2}\sum_{i=1}^{n}\left[\mathbf{x}_i 1(q_i\le \mathbf{z}_i'\gamma)-\hat{Q}_1(\gamma)\hat{Q}^{-1}\mathbf{x}_i\right]\hat{\varepsilon}_{oi}\xi_i^*$ is a zero-mean Gaussian process with covariance function $H_n(\gamma_1,\gamma_2)$. From the last theorem, $H_n(\gamma_1,\gamma_2)\xrightarrow{p}H(\gamma_1,\gamma_2)$ uniformly over $(\gamma_1,\gamma_2)\in\Gamma\times\Gamma$ under H_1^c. Also, $H_n(\gamma,\gamma)\xrightarrow{p}H(\gamma,\gamma)$ uniformly over $\gamma\in\Gamma$ under H_1^c. In summary, $T_n^*(\gamma)\overset{*}{\rightsquigarrow}H(\gamma,\gamma)^{-1/2}S(\gamma)=T^0(\gamma)$ in $\ell^\infty(\Gamma)$, where $\overset{*}{\rightsquigarrow}$ signifies the weak convergence in probability. □

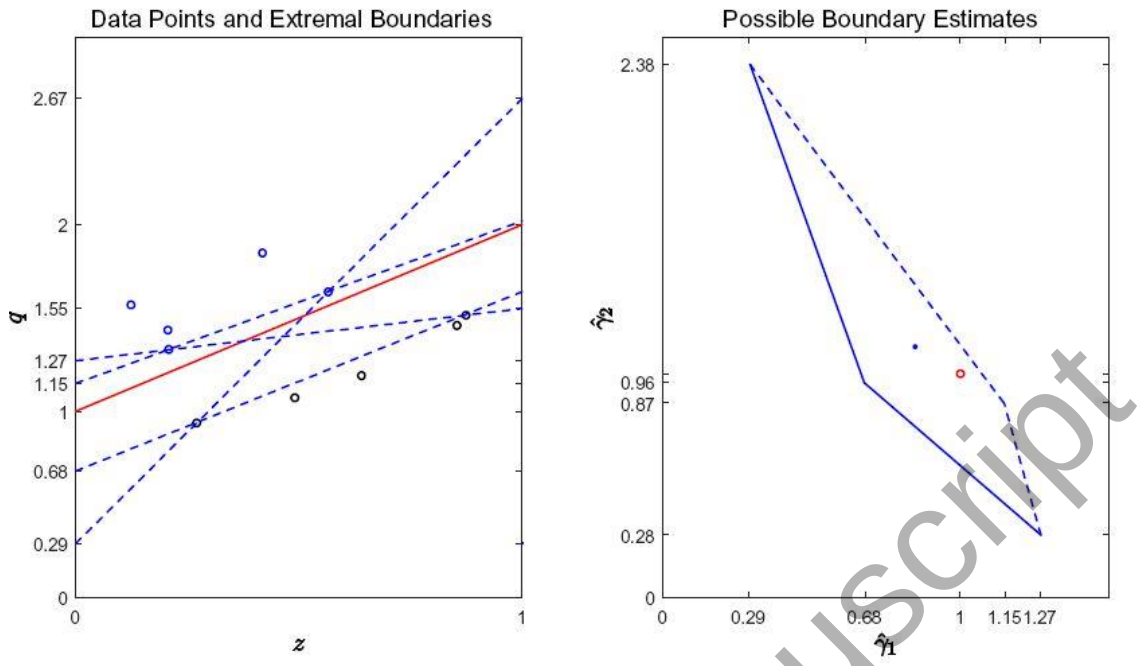

Fig. 1 The Threshold Boundary in Finite Samples in a Simple Example

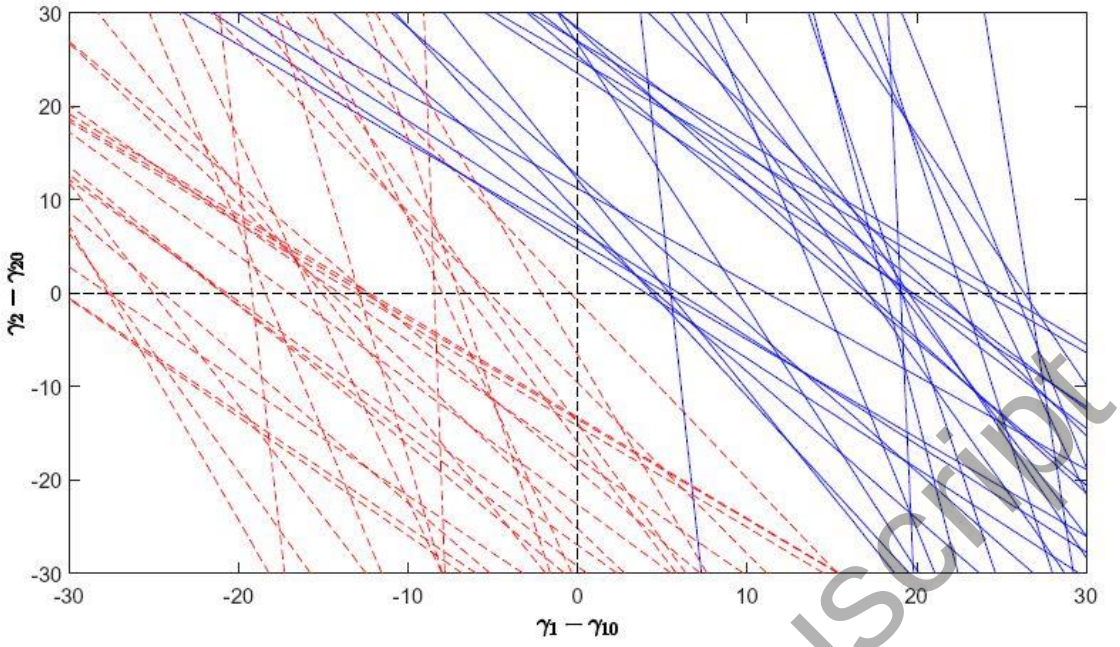

Fig. 2 Jumping Locations of $n\left[M_n(\gamma) - M_n(\gamma_0)\right]$

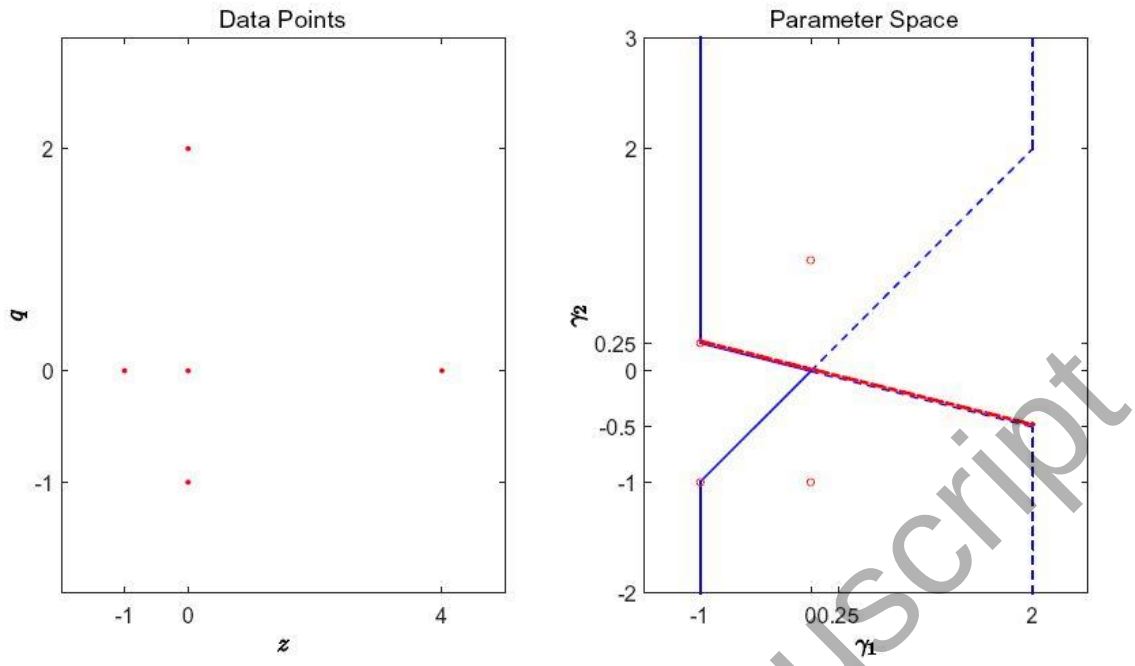

Fig. 3 Parameter Space in a Simple Example

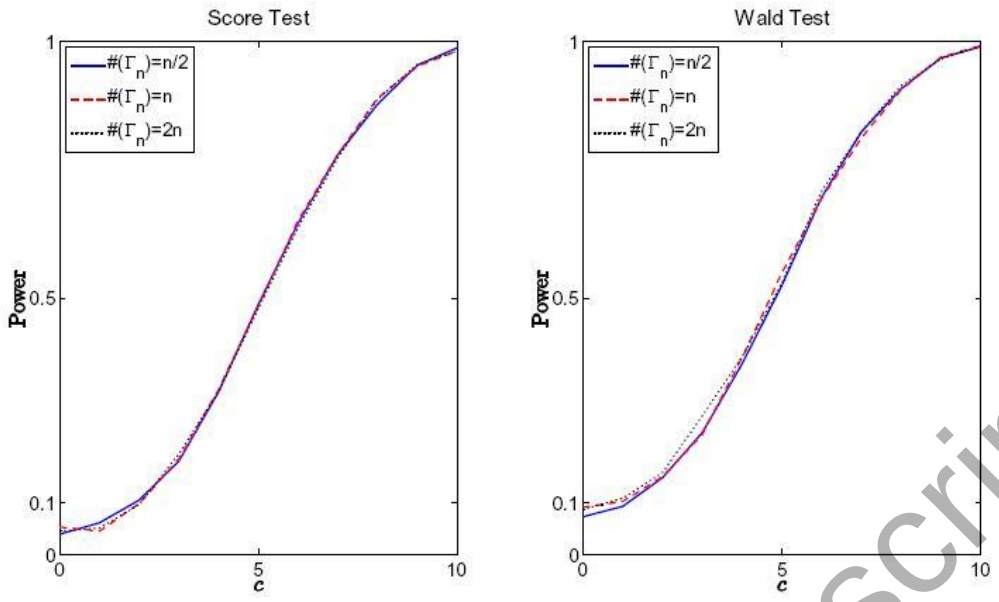

Fig. 4 Power Comparison with Different $^{\#(\Gamma_n)}$

Table 1 Risk of the Estimators of γ

$\delta_n = cn^{-1/2} \rightarrow$	c = 10		c = 20	
RMSE ($\times 10^{-2}$) $\searrow$	γ_1	γ_2	γ_1	γ_2
$\hat{\gamma}_{SLSE}$	14.416	25.027	6.949	11.787
$\hat{\gamma}_I$	14.478	23.628	4.637	8.116
$\hat{\gamma}_N$	14.427	23.515	4.629	8.113
Posterior Mean	13.089	21.639	3.631	6.266

Table 2 Coverage and Length of the 95% CIs for γ

$\delta_n = cn^{-1/2} \rightarrow$	c = 10		c = 20	
Coverage and Length $\searrow$	γ_1	γ_2	γ_1	γ_2
Coverage of Bootstrap SLSE	0.928	0.912	0.950	0.968
Coverage of NPI	0.908	0.920	0.950	0.954
Length of Bootstrap SLSE	0.609	1.066	0.302	0.522
Length of NPI	0.411	0.738	0.129	0.236

Table 3 Estimates of γ and the 95% NPI

Time Period	1940-1950	1950-1960	1960-1970	1970-1980	1980-1990
Constant	0.247 0.245 (0.231, 0.264)	0.325 0.368 (0.318, 0.414)	0.328 0.327 (0.303, 0.346)	0.174 0.183 (0.162, 0.194)	0.277 0.280 (0.265, 0.300)
Male Prejudice	–1.067 –1.057 (–1.172, –0.972)	–0.517 –0.598 (–0.856, –0.376)	–0.763 –0.752 (–0.882, –0.624)	–0.490 –0.527 (–0.607, –0.387)	–0.197 –0.202 (–0.346, –0.073)

Note: the upper value for each coefficient is the LSE, the middle value is the posterior mean, and the lower value is the NPI

Table 4 Estimates of d

Time Period	1940-1950	1950-1960	1960-1970	1970-1980	1980-1990
d	-0.358 $(-0.468, -0.247)$	-0.452 $(-0.582, -0.322)$	-0.534 $(-0.668, -0.400)$	-0.257 $(-0.318, -0.195)$	-0.193 $(-0.228, -0.158)$

Note: nominal 95% CIs are reported in parentheses.

Table 5 Estimates of γ and the 95% NPI

γ	$\mathbf{z} = 1$	$\mathbf{z} = (1, z)'$
Constant	0.108 0.102 $[0.091, 0.111]$	0.141 0.130 $[0.127, 0.143]$
Market Value $(\text{inmillion[dollar]})$	– – –	-0.044 -0.040 $[-0.042, -0.029]$

Note: the upper value for each coefficient is the LSE, the middle value is the posterior mean, and the lower value is the NPI

Table 6 Estimates of $(\alpha_\ell, \beta_\ell')'$ and Number of Observations in Each Regime

	$\mathbf{z}=1$		$\mathbf{z}=(1,z)'$	
	Regime I	Regime II	Regime I	Regime II
Constant	–0.007 (–0.017, 0.003)	0.015 (–0.003, 0.033)	0.005 (–0.006, 0.016)	–0.022 (–0.041, –0.002)
Luck	0.013 (–0.040, 0.066)	0.102 (0.061, 0.143)	0.053 (–0.004, 0.111)	0.168 (0.120, 0.215)
Skill	0.168 (0.144, 0.191)	0.067 (0.044, 0.089)	0.166 (0.143, 0.189)	0.063 (0.041, 0.086)
Female	–0.012 (–0.085, 0.061)	0.106 (–0.065, 0.278)	–0.022 (–0.105, 0.061)	0.124 (0.007, 0.242)
Age $\times 10$	–0.164 (–0.282, –0.045)	0.095 (–0.087, 0.277)	–0.143 (–0.261, –0.024)	0.022 (–0.172, 0.217)
Age^2 $\times 10^2$	0.015 (0.006, 0.025)	–0.001 (–0.016, 0.014)	0.014 (0.005, 0.024)	0.003 (–0.012, 0.019)
Tenure $\times 10$	–0.067 (–0.108, –0.026)	–0.136 (–0.193, –0.079)	–0.076 (–0.114, –0.038)	–0.110 (–0.182, –0.038)
Tenure^2 $\times 10^2$	0.005 (–0.009, 0.020)	0.025 (0.008, 0.041)	0.008 (–0.004, 0.020)	0.016 (–0.011, 0.043)
Market Value	–0.003 (–0.006, 0.000)	0.001 (–0.002, 0.004)	–0.002 (–0.007, 0.003)	–0.0004 (–0.002, 0.001)
N %ofN	19,336 74.3%	6,703 25.7%	19,137 73.5%	6,902 26.5%

Note: nominal 95% CIs are reported in parentheses.