Inference and Specification Testing in Threshold Regression with Endogeneity*

Ping Yu†
The University of Hong Kong

Qin Liao‡
The University of Hong Kong

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

First Version: June 2017
This Version: February 2021

Abstract

We propose three new methods of inference for the threshold point in endogenous threshold regression and two specification tests designed to assess the presence of endogeneity and threshold effects without necessarily relying on instrumentation of the covariates. The first inferential method is a parametric two-stage least squares method and is suitable when instruments are available. The second and third methods are based on smoothing the objective function of the integrated difference kernel estimator in different ways and these methods do not require instrumentation. All three methods are applicable irrespective of endogeneity of the threshold variable. The two specification tests are constructed using a score-type principle. The threshold effect test extends conventional parametric structural change tests to the nonparametric case. A wild bootstrap procedure is suggested to deliver finite sample critical values for both tests. Simulations show good finite sample performance of these procedures and the methods provide flexibility in testing and inference for practitioners working with threshold models.

KEYWORDS: Threshold regression, Endogeneity, Identification, Confidence interval, 2SLS, IDKE, Secification testing, Bootstrap, U-statistic

JEL-CLASSIFICATION: C21, C24, C26,

*We thank Yonghong An, Bruce Hansen, Zheng Fang, Chirok Han, Shengjie Hong, Yingyao Hu, Zhongxiao Jia, Hongyi Li, Qi Li, Jaimie Lien, R.M. Nayga Jr., Robin Sickles, Xiaojun Song, Wenwu Wang, Chuancun Yin, ChuancJun Yu and seminar participants at CMES2018, CUHK, ESAM2018, ESEM2018, HU-HUE-SMU Tripartite 2018, NASMES2018, QNU, TAMU, TEC2018 and Tsinghua for helpful comments. Phillips and Yu acknowledge support from the GRF of Hong Kong Government under Grant No. 17520716. Phillips acknowledges support from the NSF under Grant No. SES 18-50860 and a Kelly Fellowship at the University of Auckland.

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

‡Faculty of Business and Economics, The University of Hong Kong, Pokfulam Road, Hong Kong; email: liaoq@connect.hku.hk.

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

1 Introduction

In recognition of potential shifts in economic relationships, threshold models have become increasingly popular in recent econometric practice. Hansen (2011) provides an overview of the methods and their various applications in economics and finance. One typical application of the threshold model in time series is to illustrate asymmetric effects of shocks over the business cycle (e.g., Potter, 1995). Threshold models are also useful in cross section applications. For example, Hansen (2000) applied a threshold model to show that depending on the starting point, rich countries and poor countries have different growth patterns. All this literature assumes exogenous regressors and an exogenous threshold variable. But in practical work there is often uncertainty about exogeneity and threshold model applications have commonly encountered endogeneity issues. For example, the empirical growth models used in Papageorgiou (2002) and Tan (2010) both suffer from endogenous regressor problems, as argued in Frankel and Romer (1999) and Acemoglu et al. (2001).

The standard model formulation for endogenous threshold regression is

$$y = \mathbf{x}'\beta_1 1\left(q \leq \gamma\right) + \mathbf{x}'\beta_2 1\left(q > \gamma\right) + \varepsilon =: \mathbf{x}'\beta + \mathbf{x}'\delta 1\left(q \leq \gamma\right) + \varepsilon, \tag{1}$$

with $\mathbb{E}[\varepsilon|\mathbf{x}] \neq 0$, where $\mathbf{x} = (1, x', q)' \in \mathbb{R}^{d+1} =: \mathbb{R}^{\overline{d}}$, and where d and $\overline{d}$ are the dimensions of the nonconstant covariates $(x', q)'$ and all covariates including the constant. The parameter of interest is $\theta = \left(\beta_1', \beta_2', \gamma\right)'$ or equivalently, $\theta := \left(\beta', \delta', \gamma\right)'$ with $\beta = \beta_2$, $\delta = \beta_1 - \beta_2$ and $\gamma \in \Gamma$. This setup is similar to endogenous linear regression except that the regression coefficients depend on whether the threshold variable q crosses the threshold point γ.

The literature on estimation of this model includes the following three main contributions. First, Caner and Hansen (2004) (CH hereafter) use a two-stage least squares (2SLS) method to estimate γ in the small-threshold-effect framework of Hansen (2000), but assuming q is exogenous so that $\mathbb{E}[\varepsilon|\mathbf{x}] = \mathbb{E}[\varepsilon|x]$ holds. Second, working in Hansen's (2000) framework, Kourtellos, Stengos and Tan (2016) (KST hereafter) use a control function approach to deal with the case where q is also endogenous.[1] Their setup is parametric (see Kourtellos et al. (2017) for a semiparametric extension) and the asymptotic theory is flawed. Specifically, Yu, Liao and Phillips (2018) (YLP hereafter) show that the structural threshold regression (STR) estimator of the threshold point in KST is not consistent unless the endogeneity level of the threshold variable is low compared to the threshold effect. Third, Yu and Phillips (2018) (YP hereafter) use an integrated difference kernel estimator (IDKE) to estimate γ in the fixed-threshold-effect framework of Chan (1993). Their estimator can be applied irrespective of whether q is endogenous or whether instruments are available (as required in the previous two methods). Even when there are no instruments available and the model reduces to a *nonparametric* threshold regression, their estimator is still n-consistent, just as in the parametric setup.[2] The endogeneity problem is also considered in the related structural change literature, where the threshold variable is a simple time index and is always exogenous; see YP for a detailed literature review.

In spite of the theoretical developments on the estimation of γ in endogenous threshold regression, infer-

[1] If q is exogenous, then KST's estimator is asymptotically equivalent to CH's estimator.

[2] There are two other estimators of γ in nonparametric threshold regression with different motivations and objective functions from the IDKE. The first estimator is the semiparametric M-estimator of Henderson et al. (2017). That estimator can be treated as an extension of the partial linear estimator of Porter (2003) (see also Yu (2016)) in regression discontinuity designs to the case with unknown discontinuity point and extra covariates (beyond q), but this estimator can be applied only to the case with constant threshold effects; readers are referred to the supplementary materials of YP to see why the authors avoid using a generalized version of this estimator. The second estimator is the least squares estimator of Chiou et al. (2018). Chiou et al. (2018) can be treated as a nonparametric parallel of Bai and Perron (1998); for example, they allow for multiple regimes, use sequential tests to determine the number of regimes, and $q \notin \mathbf{x}$ (because q in structural change models is the time index which is usually not a covariate in $\mathbf{x}$); again, readers are referred to the supplementary materials of YP to see why the authors avoid using this estimator.

ence concerning the threshold parameter γ still presents practical difficulties especially when q is endogenous. First, the CH method should be applied only if q is exogenous. For as shown in Yu (2013a), the CH estimator is generally inconsistent when q is endogenous. Second, as mentioned above, the KST approach is not generically applicable. Third, the asymptotic distribution of the IDKE in YP is too complicated to be readily applied in empirical work. This paper seeks to alleviate these practical difficulties by proposing three new methods of confidence interval (CI) construction for γ.[3] All three methods can be applied regardless of whether q is endogenous. To our knowledge, these methods are the only valid and applicable inferential tools that are robust to endogeneity of q in the sense that the procedures need no modification when q is endogenous. The first method is a parametric two stage least squares (2SLS) method and requires instruments, while the second and third methods are based on smoothing the objective function of the IDKE in different ways so that instruments are unnecessary. In discussing the first method of inference, we also discuss the identification issue of γ using moment conditions. Of the two remaining methods, the second method assumes fixed threshold effects and uses the data around the threshold point marginally, while the third method assumes shrinking threshold effects and makes full use of data around the threshold point. So the second and third methods are similar in spirit to the smoothed maximum score (SMS) estimator of Horowitz (1992). On the other hand, the two methods differ from the SMS estimator in the sense that they have slower convergence rates than the IDKE in YP while the SMS estimator improves the convergence rate over the original maximum score estimator of Manski (1975, 1985). This feature of the methods is in some sense similar to the smoothed least squares estimator (SLSE) of Seo and Linton (2007) which also has a slower convergence rate than the usual least squares estimator (LSE) in, e.g., Yu and Fan (2019). Like the original IDKE approach, both of these IDKE-smoothing methods are nonparametric and require kernel and bandwidth selection. Practitioners can select an approach to inference from among these three methods based on their suitability to the data and on data availability. For example, if instruments are available, then the first method can be used; otherwise, the second and third methods may be preferable.

This paper also proposes two specification tests. The first tests for the existence of endogeneity, and the second tests for the presence of threshold effects with and without instruments. Both tests are score-type tests in the sense that they are constructed under the null, and their asymptotic properties are therefore easy to develop. More importantly, these tests of structural shifts are easier to implement in practice than the popular Wald test especially when instruments are unavailable. Because the Wald and score tests of structural shifts when instruments are available are standard extensions of existing tests and are well understood in the literature, these tests are relegated to an online supplement and the main text of the paper concentrates on the test without instruments. Both specification tests discussed in the main text take a nonparametric form and have an asymptotic normal (null) distribution. We suggest a wild bootstrap procedure to obtain critical values for these tests. Practitioners are encouraged to give greater attention to the inference methods and specification tests that are developed without instrumentation because good instruments are often hard to find and justify in practical work.

The rest of this paper is organized as follows. Section 2 provides an overview of the three inference methods and the two specification tests. Sections 3 to 5 analyze the three inference methods in turn and derive the corresponding asymptotic theory for constructing CIs. Section 6 presents the limit theory of the two specification tests and shows how to bootstrap the critical values. Section 7 includes some simulation results and Section 8 concludes.[4] Proofs of theorems with supporting propositions and lemmas are given in Supplements A, B and C. Additional discussion on parametric tests for the presence of threshold effects

[3] We will not discuss inference on regular parameters such as β and δ because these cases fall within the standard literature; see, e.g., CH, YP and YLP. The first inference method in this paper also covers β and δ.

[4] The dissertation of Qin Liao also includes a serious empirical application using the techniques in this paper, but to restrain the length of this paper, we decide to discuss the application in a separate paper.

when instruments are available is given in Supplement D. These supplements are collected together for online access in Yu et al. (2019).

A word on notation. The three inference methods in the paper are labeled Methods I, II and III. The symbol ℓ is used to indicate the two regimes in (1) or the two specification tests, and is not always written explicitly as '$\ell = 1, 2$'. For matrices, $A > 0$ means that A is positive definite, $span\,(A)$ denotes the column space of A, and I_m is the $m \times m$ identity matrix. For a random sequence Z_n, plimZ_n means the probability limit of Z_n as the sample size $n \to \infty$. For any random vector $\mathbf{x}$, $\mathbf{x}_{\leq\gamma} := \mathbf{x}1\,(q \leq \gamma)$ and $\mathbf{x}_{>\gamma}$ is similarly defined. For any two random vectors $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{x} \perp \mathbf{y}$ means $\mathbf{x}$ and $\mathbf{y}$ are independent. A parameter with a subscript 0 is the true value of the parameter.

2 Overview of Inferential Methods and Specification Testing

This section briefly overviews the three inferential methods and the two specification tests, introduces notation useful in the subsequent development, and details assumptions employed in the asymptotic theory.

2.1 Methods of Inference for the Threshold Point

If we write the model (1) as $y = G\,(x, q; \theta) + \varepsilon$, with $\mathbb{E}\,[\varepsilon|x, q] \neq 0$, where $G\,(x, q; \theta) = \mathbf{x}'\beta + \mathbf{x}'\delta 1\,(q \leq \gamma)$ is a nonlinear function of (x, q), then estimation of γ can be treated as in a nonlinear regression model with endogeneity. As argued in Section 2.1.6 of Blundell and Powell (2003), the fitted-value method of 2SLS relies heavily on linearity of the regression function, a feature that can explain why the 2SLS estimators in Yu (2013a) are not consistent. To restore consistency of 2SLS, we need to maintain the linear structure of the model. In other words, instead of projecting (x, q) on instruments $\mathbf{z}$, we first project $(\mathbf{x}, \mathbf{x}_{\leq\gamma})$ for a fixed γ on $\mathbf{z}$ to get the predictors $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{x}}_{\leq\gamma}$; then we can find $\widehat{\beta}(\gamma)$ and $\widehat{\delta}(\gamma)$ by regressing y on $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{x}}_{\leq\gamma}$; finally, $\widehat{\gamma}$ is obtained by minimizing $\sum_{i=1}^{n}(y_i - \widehat{\mathbf{x}}_i'\widehat{\beta}(\gamma) - \widehat{\mathbf{x}}_{\leq\gamma,i}'\widehat{\delta}(\gamma))^2$, from which we obtain $\widehat{\beta} = \widehat{\beta}(\widehat{\gamma})$ and $\widehat{\delta} = \widehat{\delta}\,(\widehat{\gamma})$. It is easy to see that this procedure is equivalent to the instrumental variable (IV) extremum estimation problem

$$\left(\widehat{\beta}, \widehat{\delta}, \widehat{\gamma}\right) = \arg\min_{\beta,\delta,\gamma}\,(Y - X\beta - X_{\leq\gamma}\delta)'P_Z(Y - X\beta - X_{\leq\gamma}\delta), \tag{2}$$

where Y, X, $X_{\leq\gamma}$ and Z are matrices stacking y_i, $\mathbf{x}_i'$, $\mathbf{x}_{\leq\gamma,i}'$ and $\mathbf{z}$ respectively, and P_Z is the projection matrix onto the instrument space $span\,(Z)$. This method, labeled as Method I, treats γ as a regular parameter and is just nonlinear 2SLS, as in Amemiya (1974). We will also show that this 2SLS estimator may be viewed as a version of the GMM estimator considered in Hall, Han and Boldea (2012) (HHB hereafter; see also Andrews (1993)) but one that turns out to have more desirable asymptotic properties, including consistency, in the endogenous threshold regression case.

To better understand the estimator $\widehat{\gamma}$, we consider the case where $\mathbf{x} = 1$, $\beta_0 = 0$ and $\delta_0 = 1$ are known, and $z = 1$. For this simple case, $y = 1\,(q \leq \gamma_0) + \varepsilon$, and the moment condition is $\mathbb{E}[z\varepsilon] = \mathbb{E}[\varepsilon] = \mathbb{E}[y] - F_q(\gamma_0) = 0$, where $F_q(\cdot)$ is the cdf of q. In other words, $\gamma_0 = F_q^{-1}(\mathbb{E}[y])$ is the $\mathbb{E}[y]$'th quantile of q, and $\widehat{\gamma} = \widehat{F}_q^{-1}(\overline{y})$. Following this intuition, we will show that: (i) $\widehat{\gamma}$ is $\sqrt{n}$-consistent, asymptotically normal, and the asymptotic variance involves the density of q at γ_0 (i.e., $f_q(\gamma_0)$) as in quantile regression; (ii) different from the usual threshold regression estimators where $\widehat{\gamma}$ is asymptotically independent of $\left(\widehat{\beta}, \widehat{\delta}\right)$, this new estimator $\widehat{\gamma}$ is correlated with $\left(\widehat{\beta}, \widehat{\delta}\right)$ asymptotically. Given these two results, a valid CI for γ can be constructed jointly with (β, δ) by means of the bootstrap, just as in quantile regression to avoid nonparametric estimation of the density $f_q(\gamma_0)$ in the asymptotic distribution.

Differing from CH estimation, this 2SLS estimator can be applied irrespective of whether q is endogenous.

Yu (2013a) shows that when q is exogenous, the CH estimator is inconsistent if the first stage predictor is a projection rather than a conditional mean. By contrast our 2SLS estimator requires only a linear projection in the first stage, so it is more robust in this respect.

Moving away from instrument-based estimation, we next introduce instrument-free estimators in Methods II and III by extending the IDKE of YP in different directions. Without instruments, the model reduces to a nonparametric threshold regression that can be written in the form

$$\begin{aligned} y &= m(x,q) + u = m_-(x,q)1\left(q \leq \gamma\right) + m_+(x,q)1\left(q > \gamma\right) + u \\ &= g(x,q) + \Delta\left(x,q\right)1\left(q \leq \gamma\right) + u, \end{aligned}$$

where $u = \varepsilon - \mathbb{E}[\varepsilon|x,q]$, $m_-(x,q) = \mathbf{x}'\beta_1 + \mathbb{E}\left[\varepsilon|x,q\right]$, $m_+(x,q) = \mathbf{x}'\beta_2 + \mathbb{E}\left[\varepsilon|x,q\right]$, $g(x,q) = m_+(x,q)$ when $q > \gamma$ and is the smooth extension of $m_+(x,q)$ when $q \leq \gamma$, and $\Delta\left(x,q\right) = m_-(x,q) - m_+(x,q)$. This setup allows $\mathbb{E}[\varepsilon|x,q]$ to be kinked or discontinuous at γ. When $\mathbb{E}[\varepsilon|x,q]$ is smooth, then $g(x,q) = \mathbf{x}'\beta + \mathbb{E}[\varepsilon|x,q]$ and $\Delta\left(x,q\right) = \mathbf{x}'\delta$; otherwise, $g(x,q) \neq \mathbf{x}'\beta + \mathbb{E}[\varepsilon|x,q]$ for $q \leq \gamma$ and $\Delta\left(x,q\right) \neq \mathbf{x}'\delta$. When $\mathbb{E}[\varepsilon|x,q]$ is continuous $\Delta\left(x,\gamma\right) = \left(1, x', \gamma\right)\delta$.

To construct the IDKE of γ, we start by defining a generalized kernel function, following Müller (1991).

Definition: *$k_h(\cdot,\cdot)$ is called a univariate generalized kernel function of order p if $k_h(u,t) = 0$ when $u > t$ or $u < t-1$ and for all $t \in [0,1]$,*

$$\int_{t-1}^{t} u^j k_h(u,t)du = \begin{cases} 1, & \text{if } j = 0, \\ 0, & \text{if } 1 \leq j \leq p-1. \end{cases}$$

A popular example of the generalized kernel function is obtained as follows. Define the space

$$\mathcal{M}_p\left([a,b]\right) = \left\{ g \in \text{Lip}\left([a,b]\right), \int_a^b x^j g(x)dx = \begin{cases} 1, & \text{if } j = 0, \\ 0, & \text{if } 1 \leq j \leq p-1 \end{cases} \right\},$$

where Lip$([a,b])$ denotes the space of Lipschitz continuous functions on $[a,b]$. Define $k_+(\cdot,\cdot)$ and $k_-(\cdot,\cdot)$ as follows:

(i) The support of $k_-(x,r)$ is $[-1,r]\times[0,1]$ and the support of $k_+(x,r)$ is $[-r,1]\times[0,1]$.

(ii) $k_-(\cdot,r) \in \mathcal{M}_p\left([-1,r]\right)$ and $k_+(\cdot,r) \in \mathcal{M}_p\left([-r,1]\right)$.

(iii) $k_+(x,r) = k_-(-x,r)$.

(iv) $k_-(-1,r) = k_+(1,r) = 0$.

Condition (iv) implies that $k_-(\cdot,r)$ is Lipschitz on $(-\infty,r]$ and $k_+(\cdot,r)$ is Lipschitz on $[-r,\infty)$. This assumption is important in deriving the asymptotic distribution of the IDKE of γ. Readers are referred to Appendix A of Porter and Yu (2015) for related discussion in the DKE case.

To simplify the construction of $k_h(u,t)$, the following constraints are imposed on the support of x and on the parameter space.

Assumption S: $(y,x',q)' \in \mathbb{R}\times\mathcal{X}\times\mathcal{Q} \subset \mathbb{R}^{\overline{d}}$, $\mathcal{X} = [0,1]^{d-1}$, $\mathcal{Q} = [\underline{q},\overline{q}]$, *and* $\gamma \in \Gamma = [\underline{\gamma},\overline{\gamma}] \subset \mathcal{Q}$, $\beta \in B \subset \mathbb{R}^{\overline{d}}$, $\delta \in \Xi \subset \mathbb{R}^{\overline{d}}$, *where $\underline{q}$ can be $-\infty$ and $\overline{q}$ can be ∞, and Γ, B and Ξ are compact.*

We do not restrict δ_0 to be fixed or to shrink to zero in all cases. Rather, δ_0 is taken as fixed in Method II and shrinks to zero in Method III. We assume x is continuously distributed, but note that continuous and discrete

components may be accommodated, at least in a conceptually straightforward manner but at the expense of additional notational complexity, by using the continuous covariate estimator within samples homogeneous in the discrete covariates. Requiring the support of x to be $[0,1]^{d-1}$ is not restrictive as this support can be achieved by use of a suitable monotone transformation such as the empirical percentile transformation. The compactness assumption on $\mathcal{X}$ simplifies the proof and may be relaxed by imposing restrictions on the tail of the distribution of x.

Define

$$\begin{aligned} k(\cdot) &= k_+(\cdot,1) = k_-(\cdot,1) \in \mathcal{M}_p\left([-1,1]\right), k_h(u) = k(u/h)/h, \\ k_+(\cdot) &= k_+(\cdot,0) \in \mathcal{M}_p\left([0,1]\right), k_h^+(u) = k_+(u/h)/h, \\ k_-(\cdot) &= k_-(\cdot,0) \in \mathcal{M}_p\left([-1,0]\right), k_h^-(u) = k_-(u/h)/h, \end{aligned}$$

and

$$k_h(u,t) = \begin{cases} \frac{1}{h}k\left(\frac{u}{h}\right), & \text{if } h \le t \le 1-h, \\ \frac{1}{h}k_+\left(\frac{u}{h},\frac{t}{h}\right), & \text{if } 0 \le t \le h, \\ \frac{1}{h}k_-\left(\frac{u}{h},\frac{1-t}{h}\right), & \text{if } 1-h \le t \le 1. \end{cases} . \tag{3}$$

Then $k_h(u,t)$ is a generalized kernel function of order p. We may construct a corresponding multivariate generalized kernel function of order p by taking the product of univariate generalized kernel functions of order p. We only require $k_h(u,t)$ to be a first order kernel function in Method II but may require it to be a higher order kernel function in Method III.[5] In particular, we use the following two conditions. For Method II we use:

Assumption K: *$k_h(u,t)$ takes the form of (3) with $p=1$, $k_+(0) = k_-(0) = 0$, and $k_+'(0) > 0$, $k_-'(0) < 0$.*

For Method III we use:

Assumption K′: *$k_h(u,t)$ takes the form of (3) with $p = s$, and $k_+(0) = k_-(0) > 0$, where s is the smoothness index of $g(x,q)$ and will be defined in Assumption G′ below.*

Assumption K mimics Assumptions B2 and B3 of Delgado and Hidalgo (2000) (DH hereafter) and Assumption K′ is Assumption K in YP with the additional requirement that $p = s$. Higher order kernels are required in Assumption K′ only to achieve the optimal convergence rate of γ. In practice, $p = 1$ is sufficient.

Given $k_h(u,t)$, the IDKE of γ is constructed as the extremum estimator which satisfies

$$\begin{aligned} \widehat{\gamma} &= \arg\max_{\gamma} \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{n-1}\sum_{j=1,j\neq i}^{n} y_j K_{h,ij}^{\gamma-} - \frac{1}{n-1}\sum_{j=1,j\neq i}^{n} y_j K_{h,ij}^{\gamma+}\right]^2 \\ &=: \arg\max_{\gamma} \frac{1}{n}\sum_{i=1}^{n}\widehat{\Delta}_i^2(\gamma) =: \arg\max_{\gamma} \widehat{Q}_n(\gamma), \end{aligned} \tag{4}$$

where

$$\begin{aligned} K_{h,ij}^{\gamma-} &= \textstyle\prod_{l=1}^{d-1} k_h(x_{lj}-x_{li},x_{li}) \cdot k_h^-(q_j-\gamma) =: K_{h,ij}^x k_h^-(q_j-\gamma), \\ K_{h,ij}^{\gamma+} &= \textstyle\prod_{l=1}^{d-1} k_h(x_{lj}-x_{li},x_{li}) \cdot k_h^+(q_j-\gamma) =: K_{h,ij}^x k_h^+(q_j-\gamma), \\ \widehat{\Delta}_i(\gamma) &= \tfrac{1}{n-1}\textstyle\sum_{j=1,j\neq i}^{n} y_j K_{h,ij}^{\gamma-} - \tfrac{1}{n-1}\sum_{j=1,j\neq i}^{n} y_j K_{h,ij}^{\gamma+}. \end{aligned} \tag{5}$$

[5] Note here that the usual symmetric kernel is a second order kernel, but the boundary kernel is only a first order kernel because $\int u k_h(u,t) \neq 0$,

For notational convenience, we here use the same bandwidth for each dimension of $(x', q)'$, although there may be some finite sample improvement from using different bandwidths in each dimension. As suggested in Yu (2012, 2015b), we need only check the mid-points of the contiguous q_i's in the optimization process of (4).[6] In other words, the argmax operator is a mid-point operator. The summation in the parenthesis of (4) excludes $j = i$, which is a standard strategy in converting a V-statistic to a U-statistic. Also, the normalization factor $\sum_{j=1,j\neq i}^{n} K_{h,ij}^{\gamma\pm}$ does not appear in the construction of $\widehat{\gamma}$, thereby avoiding random denominator issues in conditional mean estimation and simplifying the derivation of the limit distribution of $\widehat{\gamma}$, a technique that dates back at least to Powell et al. (1989). This form of $\widehat{\gamma}$ has some practical advantages especially when d is large. Since the conditional mean is estimated at the boundary point $q = \gamma$, the local linear smoother (LLS) or local polynomial estimator (LPE) may be considered to ameliorate bias. However, when d is large, there are not many data points in a h neighborhood of $(x_i', \gamma)'$. As a result, not only does the LLS lose degrees of freedom (by estimating more parameters) but its denominator matrix can be close to singular, which disrupts finite sample performance. Further, differing from regular parameter estimation (such as conditional mean estimation), use of the LLS in this context does not affect the first-order asymptotic distribution of $\widehat{\gamma}$.

CI construction based on the limit distribution of $\widehat{\gamma}$ under Assumption K$'$ and fixed threshold effects (i.e., in the framework of YP) is challenging because the asymptotics involve a compound Poisson process, making simulation awkward. Methods II and III use different smoothing schemes to achieve more convenient asymptotic distributions. Method II assumes fixed threshold effects but uses data in the neighborhood of γ_0 only marginally. The resulting asymptotic theory is normal and the CI can be constructed by inverting either the t or likelihood ratio (LR) statistic. Method III fully utilizes data in the neighborhood of γ_0 but assumes shrinking threshold effects. The limit distribution then involves a two-sided Brownian motion. As suggested in Hansen (2000) we can invert the LR statistic (rather than the t-statistic) to improve finite-sample performance. As expected, due to insufficient usage of data information in the neighborhood of γ_0, the convergence rates of the IDKE in both these methods are slower than the $O(n)$ rate in YP. In Method II we also require $k_{\pm}'(0) \neq 0$, or else the convergence rate of the IDKE is even slower.

We next provide some intuition that helps to justify the extremum estimator $\widehat{\gamma}$. For this purpose we impose the following Assumption F on the distribution of $(x', q)'$ in Method II and Assumptions G and G$'$ on $g(x, q)$ in Method III.

Assumption F: *The density $f(x, q)$ of $(x', q)'$ is second order continuously differentiable and satisfies $0 < \underline{f} \leq f(x, q) \leq \overline{f} < \infty$ for $(x', q)' \in X \times \Gamma_\epsilon$, where $\Gamma_\epsilon := \left(\underline{\gamma} - \epsilon, \overline{\gamma} + \epsilon\right)$ for some $\epsilon > 0$ and* and $(\underline{f}, \overline{f})$ are some fixed quantities.

Assumption G: *$g(x, q)$ is second order continuously differentiable on $\mathcal{X} \times \Gamma_\epsilon$.*

Assumption G$'$: *$g(x, q)$ is s'th order continuously differentiable on $X \times \Gamma_\epsilon$ with $s \geq d$.*

Assumption F implies that $f_q(\gamma)$ is continuous, and $0 < \underline{f}_q \leq f_q(\gamma) \leq \overline{f}_q < \infty$ for $\gamma \in \Gamma_\epsilon$ and fixed $\left(\underline{f}_q, \overline{f}_q\right)$, and the conditional density $f_{x|q}(x|q)$ is bounded below and above for $(x', q)' \in \mathcal{X} \times \Gamma_\epsilon$; see Yu and Zhao (2013) for relaxation of these conditions. The first part of Assumption F implies that there are no discrete covariates in x. As mentioned earlier in the remarks following Assumption S, this assumption is made for simplicity, just as in Robinson (1988), and is not critical to the methodology or the limit theory. The second

[6] Although in the fixed-threshold-effect framework with $k_{\pm}(0) > 0$ the asymptotic distribution of $\widehat{\gamma}$ depends on whether the left endpoint or the middle point of the maximizing interval is taken as the maximizer, the asymptotic distributions of the two IDKEs in the present paper are both invariant to such choices of $\widehat{\gamma}$.

part of Assumption F implies that γ_0 is not on the boundary of $\mathcal{Q}$. Under these two assumptions, we can expect the objective function $\widehat{Q}_n(\gamma)$ to converge to

$$\mathbb{E}\left[\{\mathbb{E}[y|x,q=\gamma-]f(x,\gamma)-\mathbb{E}[y|x,q=\gamma+]f(x,\gamma)\}^2\right]=\int\left(\mathbb{E}[y|x,q=\gamma-]-\mathbb{E}[y|x,q=\gamma+]\right)^2 f(x,\gamma)^2 f(x)dx.$$

Since $f(x)$ and $f(x,\gamma)$ are continuous in x and γ, there will be a jump in the limit only if $\gamma=\gamma_0$ which provides identifying information. In view of these properties, the threshold point can then be identified and consistently estimated by maximizing $\widehat{Q}_n(\gamma)$ under an additional requirement on the differential

$$\Delta(x,\gamma_0):=\mathbb{E}[y|x,q=\gamma_0-]-\mathbb{E}[y|x,q=\gamma_0+], \tag{6}$$

which enables identification of γ_0.

Assumption I: $\Delta(x,\gamma_0)\neq 0$ *for* x *in some set of positive Lebesgue measure in* $\mathcal{X}$.

In Method I, $\Delta(x,\gamma_0)=(1,x',\gamma_0)\,\delta_0$, so we can replace $\Delta(x,\gamma_0)$ by $(1,x',\gamma_0)\,\delta_0$ in Assumption I. For comparison, we state the following Assumption I′.

Assumption I′: $\delta_0\neq\mathbf{0}$, *where* $\neq$ *here means that at least one element is unequal.*

Note that Assumption I is stronger than Assumption I′ when $\Delta(x,\gamma_0)=(1,x',\gamma_0)\,\delta_0$. For example,

$$\delta_0=\begin{cases}\left(1,\mathbf{0},-\frac{1}{\gamma_0}\right)', & \text{if } \gamma_0\neq 0,\\ (0,\mathbf{0},1)', & \text{if } \gamma_0=0,\end{cases}$$

is nonzero but does not satisfy Assumption I. Assumption I implies that $P((1,x',\gamma_0)\,\delta_0\neq 0)>0$, which excludes the continuous threshold regression (CTR) of Chan and Tsay (1998) (see also Hansen (2017)).

For comparison, we also review the DKE in DH here. Define the DKE

$$\begin{aligned}\widetilde{\gamma} &= \arg\max_{\gamma}\left[\frac{1}{n}\sum\nolimits_{j=1}^{n}y_jK_{h,j}^{\gamma-}-\frac{1}{n}\sum\nolimits_{j=1}^{n}y_jK_{h,j}^{\gamma+}\right]^2\\ &=: \arg\max_{\gamma}\widehat{\Delta}_o^2(\gamma)=:\arg\max_{\gamma}\widetilde{Q}_n(\gamma),\end{aligned}$$

where

$$K_{h,j}^{\gamma-}=\prod\nolimits_{l=1}^{d-1}k_h(x_{lj}-x_{ol},x_{ol})\cdot k_h^-(q_j-\gamma),\ K_{h,j}^{\gamma+}=\prod\nolimits_{l=1}^{d-1}k_h(x_{lj}-x_{ol},x_{ol})\cdot k_h^+(q_j-\gamma),$$

and x_o is some fixed point in the interior of $\mathcal{X}$.[7] As explained in YP, selection of x_o is difficult from both theoretical and practical perspectives. As distinct from the DKE, the IDKE procedure integrates the jump information over all the x_i, thereby removing the problem of choosing x_o. Further, usage of all the data ensures that the IDKE has greater identifying capability than the DKE in both Methods II and III.

[7]Strictly speaking, DH normalize the first term of $\widehat{\Delta}_o(\gamma)$ by $\widehat{f}_o^{\gamma-}:=\frac{1}{n}\sum_{j=1}^{n}K_{h,j}^{\gamma-}$ and the second term by $\widehat{f}_o^{\gamma+}:=\frac{1}{n}\sum_{j=1}^{n}K_{h,j}^{\gamma+}$. However, their estimator is asymptotically equivalent to $\arg\max_{\gamma}\widehat{\Delta}_o^2(\gamma)/f(x_o,\gamma_o)^2$ and has the same asymptotic distribution as $\widetilde{\gamma}$.

2.2 Overview of Two Specification Tests

The first specification test addresses potential endogeneity and the corresponding hypotheses $H^{(1)}$ are formulated as follows

$$\begin{aligned} H_0^{(1)} &: \mathbb{E}[\varepsilon|x,q]=0, \\ H_1^{(1)} &: \mathbb{E}[\varepsilon|x,q]\neq 0. \end{aligned}$$

This exogeneity test can be conducted prior to model estimation. When instruments are available, the Hausman test in Kapetanios (2010) can be applied. In the present paper we therefore consider only the case without instruments and apply the techniques developed in Fan and Li (1996) and Zheng (1996) to test the null $H_0^{(1)}$. In the second test, the hypotheses $H^{(2)}$ are

$$\begin{aligned} H_0^{(2)} &: \beta_1=\beta_2 \text{ or } \delta=0, \\ H_1^{(2)} &: \beta_1\neq\beta_2 \text{ or } \delta\neq 0. \end{aligned}$$

If $H_0^{(1)}$ is not rejected, i.e., there is no evidence of endogeneity, then $H^{(2)}$ involve a conventional parametric structural change test, such as that considered in Davies (1977, 1987), Andrews (1993), Andrews and Ploberger (1994) and Hansen (1996), among others. If $H_0^{(1)}$ is rejected, the ensuing situation is more complex. When there are instruments, Wald-type test statistics such as the sup-statistic in Section 5 of Caner and Hansen (2004) or score-type statistics such as those in Yu (2013b) can be used. Since the asymptotic distributions of both these types of test statistics are not pivotal, the simulation method of Hansen (1996) and De Jong (1996) can be applied to obtain critical values. Details concerning these tests are given in Supplement D of the paper because techniques for these tests are nowadays standard.

When there are no instruments, the Wald-type statistic is hard to implement since its asymptotic distribution is hard to derive given that $\widehat{\delta}$ can only be estimated at a nonparametric rate – see Section 3.3 of Porter and Yu (2015) for discussion.[8] However, the score-type test of Porter and Yu (2015) can be extended to this case with some technical complications. Importantly, the hypotheses $H^{(2)}$ relate to whether $m(x,q)$ is continuous, so $H_0^{(2)}$ encompasses more data generating processes (DGPs) than the null hypothesis in the usual structural change literature where $m(x,q)$ has a simple parametric form. In other words, the usual parametric tests have power against alternatives in which $m(x,q)$ does not take the form $\mathbf{x}'\beta+\mathbf{x}'\delta 1(q\leq\gamma)$ (see, e.g., Section 5.4 of Andrews (1993))[9], but our test has only trivial asymptotic power in such continuous $m(x,q)$ cases. A simple example may clarify the point. Suppose $m(x,q)=\alpha+\beta q$, in contrast to the specifications employed for our tests, which are based on (1), or $y=\alpha+\delta 1(q\leq\gamma)+\varepsilon$. It is easy to see that the usual tests have power against $m(x,q)$, which is very smooth in this case. In summary, the usual tests have power against both misspecification and certain types of structural change, whereas our test has non-trivial power only against threshold structural change, which may be more relevant in practical work.[10] But this

[8] Gao et al. (2008) discuss an average form of such a test in the time series context. But their test is not easy to extend to the case with a nonparametric threshold boundary as in the present framework. See also Hidalgo (1995) for a nonparametric conditional moment test for structural stability in a fully nonparametric environment, which focuses on global stability rather than local stability as here.

[9] In this framework and assuming $m(x,q)=\mathbf{x}'\beta(q)$, the structural change tests focus on whether $\beta(q)=\beta$. See, e.g., Chen and Hong (2012), Kristensen (2012) and references therein for related tests in the time series context using nonparametric techniques. Actually, we can test whether $\beta(q)$ is continuous by extending the tests below, e.g., we can construct residuals $\widehat{e}_i$ in $I_n^{(2)}$ by estimating $\beta(q)$ using estimation techniques from the varying coefficient model (VCM) literature - see Robinson (1989, 1991), Cleveland et al. (1992) and Hastie and Tibshirani (1993) for early developments, and Fan and Zhang (2008) for a summary of recent developments.

[10] In the same way, there are also cases where the parametric test does not have power when there is a nonparametric threshold effect; see Example 1 of Hidalgo (1995).

advantage does not come for free: the usual tests have power against $n^{-1/2}$ local alternatives, while our test needs a larger (than $n^{-1/2}$) local alternative to generate non-trivial power. Understandably so, because our test is essentially nonparametric whereas the usual tests are parametric.

In the following discussion of the two specification tests, H_0 indicates both $H_0^{(1)}$ and $H_0^{(2)}$, and H_1 indicates both $H_1^{(1)}$ and $H_1^{(2)}$, $1_q^\Gamma = 1(q \in \Gamma)$, $1_i^\Gamma = 1(q_i \in \Gamma)$, $m_i = m(x_i, q_i) = \mathbb{E}[y_i|x_i, q_i]$, $f_i = f(x_i, q_i)$, $K_{h,ij} = K_{h,ij}^x \cdot k_h(q_j - q_i)$, and $L_{b,ij} = L_{b,ij}^x \cdot l_b(q_j - q_i)$ with $l_b(\cdot)$ similarly defined as $k_h(\cdot)$. Denote the class of probability measures under $H_0^{(\ell)}$ as $\mathcal{H}_0^{(\ell)}$ and under $H_1^{(\ell)}$ as $\mathcal{H}_1^{(\ell)}$. Both $\mathcal{H}_0^{(\ell)}$ and $\mathcal{H}_1^{(\ell)}$ are characterized by $m(\cdot)$, so we acknowledge the dependence of the distribution of y given $(x', q)'$ upon $m(x, q)$ by denoting probabilities and respective expectations as P_m and $\mathbb{E}_m$. To unify notation, we define $u_i = y_i - \mathbb{E}[y_i|x_i, q_i] = y_i - m_i$ under both the null and alternative in these tests.

For the first test, we use the statistic

$$I_n^{(1)} = \frac{nh^{d/2}}{n(n-1)} \sum_i \sum_{j \neq i} K_{h,ij} \widehat{e}_i \widehat{e}_j,$$

and, for the second, we use

$$I_n^{(2)} = \frac{nh^{d/2}}{n(n-1)} \sum_i \sum_{j \neq i} 1_i^\Gamma 1_j^\Gamma K_{h,ij} \widehat{e}_i \widehat{e}_j.$$

The exact forms of $\widehat{e}_i$ in these two tests are defined later. To motivate the statistics, let $e = y - \overline{m}(x)$, where

$$\begin{aligned} \overline{m}(\cdot) &= \arg \inf_{\widetilde{m}(x,q) = \mathbf{x}'\beta + \mathbf{x}'\delta 1(q \leq \gamma)} \mathbb{E}\left[(y - \widetilde{m}(x,q))^2 \right] \\ &= \arg \inf_{\widetilde{m}(x,q) = \mathbf{x}'\beta + \mathbf{x}'\delta 1(q \leq \gamma)} \mathbb{E}\left[(m(x,q) - \widetilde{m}(x,q))^2 \right] \end{aligned} \tag{7}$$

in the first test, and

$$\begin{aligned} \overline{m}(\cdot) &= \arg \inf_{\widetilde{m} \in \mathcal{C}_s(B, \mathcal{X} \times \mathcal{Q})} \mathbb{E}\left[(y - \widetilde{m}(x,q))^2 1_q^\Gamma \right] \\ &= \arg \inf_{\widetilde{m} \in \mathcal{C}_s(B, \mathcal{X} \times \mathcal{Q})} \mathbb{E}\left[(m(x,q) - \widetilde{m}(x,q))^2 1_q^\Gamma \right], \end{aligned} \tag{8}$$

in the second test, where $\mathcal{C}_s(B, \mathcal{X} \times \mathcal{Q})$ is the class of s times continuously differentiable functions on $\mathcal{X} \times \mathcal{Q}$ with all derivatives up to order s bounded by B. In other words, we use $\widetilde{m}(x,q) = \mathbf{x}'\beta + \mathbf{x}'\delta 1(q \leq \gamma)$ to approximate $m(x,q)$ in the first test and use $\widetilde{m} \in \mathcal{C}_s(B, \mathcal{X} \times \mathcal{Q})$ to approximate $m(x,q)$ in the second test. Note that in the first test the model need not have a threshold effect. The reason is that the class of functions $\{\mathbf{x}'\beta + \mathbf{x}'\delta 1(q \leq \gamma)\}$ includes the linear function where $\delta = \mathbf{0}$, the CTR of Chan and Tsay (1998) where $\delta \neq \mathbf{0}$ but $\delta_x = \mathbf{0}$ and $\delta_\alpha + \delta_q \gamma = 0$, and the usual threshold regression where $\delta_x \neq \mathbf{0}$ or $\delta_\alpha + \delta_q \gamma \neq 0$; see Yu (2017) for more discussion on misspecified threshold regression. Here, δ is partitioned according to the partition of $\mathbf{x} = (1, x', q)'$ as $(\delta_\alpha, \delta_x', \delta_q')'$.

Note further that $e = u$ under H_0, so e has the same meaning in both $I_n^{(1)}$ and $I_n^{(2)}$ under H_0. Observe that $\mathbb{E}[e\mathbb{E}[e|x,q]f(x,q)] = \mathbb{E}\left[\mathbb{E}[e|x,q]^2 f(x,q)\right] \geq 0$ in the first test and $\mathbb{E}\left[e\mathbb{E}[e|x,q] f(x,q) 1_q^\Gamma\right] = \mathbb{E}\left[\mathbb{E}[e|x,q]^2 f(x,q) 1_q^\Gamma\right] \geq 0$ in the second test where the equalities hold if and only if H_0 holds. So we can construct the statistic based on the moment $\mathbb{E}[e\mathbb{E}[e|x,q]f(x,q)]$ in the first test and the moment $\mathbb{E}[e\mathbb{E}[e|x,q] f(x,q) 1_q^\Gamma]$ in the second test. Here, $f(x,q)$ is added in to avoid the random denominator problem in kernel estimation, and 1_q^Γ appears in the second test because threshold effects can occur only on $q \in \Gamma$.

To construct a feasible test statistic, we need sample analogues of e and $\mathbb{E}[e|x,q]f(x,q)$. For the first

test, the sample counterpart of e is

$$\widehat{e}_i = y_i - \widehat{y}_i = y_i - \left[\mathbf{x}_i'\widehat{\beta} + \mathbf{x}_i'\widehat{\delta}1\left(q_i \leq \widehat{\gamma}\right)\right], \tag{9}$$

where $\left(\widehat{\beta}', \widehat{\delta}', \widehat{\gamma}\right)'$ is the LSE. For the second test, let

$$\widehat{e}_i = y_i - \widehat{y}_i = \left(m_i - \widehat{m}_i\right) + \left(u_i - \widehat{u}_i\right), \tag{10}$$

where

$$\widehat{y}_i = \frac{1}{n-1}\sum\nolimits_{j\neq i} y_j L_{b,ij} \Big/ \widehat{f}_i \tag{11}$$

and $\widehat{f}_i$ is the corresponding kernel estimator of f_i given by

$$\widehat{f}_i = \frac{1}{n-1}\sum\nolimits_{j\neq i} L_{b,ij},$$

and $\widehat{m}_i$ and $\widehat{u}_i$ are defined in the same way as $\widehat{y}_i$ in (11) with y_j replaced by m_j and u_j, respectively. Under H_0, $\widehat{e}_i$ is a good estimate of u_i, while under H_1, $\widehat{e}_i$ includes a bias term which generates power. Now, $\mathbb{E}[e|x,q]f(x,q)$ at $(x_i', q_i)'$ is estimated by $\frac{1}{n-1}\sum_{j\neq i}\widehat{e}_j K_{h,ij}$ in the first test and by $\frac{1}{n-1}\sum_{j\neq i}\widehat{e}_j K_{h,ij}1_j^{\Gamma}$ in the second test. Hence, we may regard $I_n^{(1)}$ and $I_n^{(2)}$ as the sample analogues of $\mathbb{E}[e\mathbb{E}[e|x,q]f(x,q)]$ and $\mathbb{E}[e\mathbb{E}[e|x,q]f(x,q)1_q^{\Gamma}]$. The statistics are constructed under the null, mimicking the idea of score tests. For example, the construction of $I_n^{(2)}$ does not involve $H_1^{(2)}$ at all (see Figure 1 of Porter and Yu (2015) for an intuitive illustration in a simple case without x), whereas the usual test statistics in the structural change literature typically involve $H_1^{(2)}$ in one way or another.

3 Inference Based on the 2SLS Estimator

In this section, we derive the asymptotic distribution of the 2SLS estimator of θ and discuss some identifiability results for γ when estimation is based on moment conditions. First, note that the 2SLS estimator of γ can be written in GMM form as

$$\widehat{\gamma} = \arg\min_{\gamma} \widehat{Q}_n\left(\gamma\right),$$

where

$$\widehat{Q}_n\left(\gamma\right) = \min_{\beta,\delta} \widehat{Q}_n\left(\theta\right) := \min_{\underline{\theta}} \widehat{g}_n\left(\theta\right)' \widehat{W}\widehat{g}_n\left(\theta\right) \tag{12}$$

with $\underline{\theta} = \left(\beta', \delta'\right)'$, $\widehat{W} = \left(n^{-1}Z'Z\right)^{-1}$ and

$$\widehat{g}_n\left(\theta\right) = \frac{1}{n}\sum_{i=1}^{n} g_i\left(\theta\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i\left(y_i - \mathbf{x}_i'\beta - \mathbf{x}_{\leq\gamma,i}'\delta\right).$$

To develop asymptotic properties of $\widehat{\theta}$ we make the following assumption. First, throughout our analysis we use the notation δ_n for the true value of δ when we allow δ to shrink to zero, as in Hansen (2000), and we use δ_0 to denote the true value of δ when δ is fixed to signify this difference.

Assumption IV: $\mathbb{E}\left[\mathbf{z}\varepsilon\right] = \mathbf{0}$, $\dim\left(\mathbf{z}\right) = l \geq 2\overline{d}+1$, $\delta_n / \left\|\delta_n\right\| \to c$,

$$G = \left(\mathbb{E}\left[\mathbf{z}\mathbf{x}'\right], \mathbb{E}\left[\mathbf{z}\mathbf{x}_{\leq\gamma_0}'\right], \mathbb{E}\left[\mathbf{z}\mathbf{x}'|q=\gamma_0\right] c f_q\left(\gamma_0\right) \right)$$

is of full column rank, $\underline{G}_\gamma = E\left[\mathbf{z}\left(\mathbf{x}', \mathbf{x}'_{\leq\gamma}\right)\right] =: \left(\underline{G}_1, \underline{G}_{2,\gamma}\right)$ *is of full column rank for any* $\gamma \in \Gamma$, $W := E\left[\mathbf{z}\mathbf{z}'\right] > 0$, *and* $\Omega := E\left[\mathbf{z}\mathbf{z}'\varepsilon^2\right] > 0$. *Also, there does not exist a vector* $a \in \mathbb{R}^{2\overline{d}}$ *such that* $\underline{G}_\gamma a = \underline{G}_{2,\gamma_0} c$ *for any* $\gamma \neq \gamma_0$.

When δ is fixed, the parameter c is just the normalized form of δ. When $\|\delta_n\| \to 0$, only the components of c that correspond to the lowest shrink rate of δ_n are nonzero. Full column rankness of G excludes CTR models where $\mathbf{x}'\delta_n | q = \gamma_0$ is always zero so that the third part of G is a zero matrix.[11] But this assumption is nonetheless weaker than full column rankness of $\mathbb{E}\left[\mathbf{z}\mathbf{x}' | q = \gamma_0\right]$. This is because if $\mathbb{E}\left[\mathbf{z}\mathbf{x}' | q = \gamma_0\right]$ has full column rank, then 1 and q cannot be elements of $\mathbf{x}$ simultaneously; otherwise, the first and the last columns of $\mathbb{E}\left[\mathbf{z}\mathbf{x}' | q = \gamma_0\right]$ would be collinear.

All other conditions in Assumption IV are standard except the last condition. This condition is required for the identification of γ_0. Take the fixed-δ case as an example where c can be taken as δ_0 and no normalization on δ is required. Note that if $\mathbb{E}\left[\mathbf{z}\varepsilon\right] = \mathbf{0}$, then $\mathbb{E}\left[\mathbf{z}y\right] = \underline{G}_{\gamma_0}\theta_0$. If there exists an $a := \left(a_1', a_2'\right)'$ such that $\underline{G}_\gamma a = \underline{G}_{2,\gamma_0}\delta_0$ for some $\gamma \neq \gamma_0$, then under this γ, we can still let $\underline{\theta} = (\beta_0 + a_1, a_2)$ satisfy the moment conditions, in which case the model is not identified by the moment conditions. This condition requires that the l-dimensional vector $\underline{G}_{2,\gamma_0}\delta_0 \notin \bigcup_{\gamma\neq\gamma_0} span\left(\underline{G}_1, \underline{G}_{2,\gamma}\right)$, where $span\left(\underline{G}_1, \underline{G}_{2,\gamma}\right)$ is a $2\overline{d} < l$ dimensional space. There is an important case where this condition is violated. If q is independent of $\left(\mathbf{z}', \mathbf{x}'\right)'$ as in the structural change model where q is the time index, then $\underline{G}_{2,\gamma_0} = \underline{G}_1 F_q\left(\gamma_0\right)$ and $a = \left(F_q\left(\gamma_0\right)\delta_0', \mathbf{0}'\right)'$ satisfy $\underline{G}_\gamma a = \underline{G}_{2,\gamma_0}\delta_0$. In the TR context, if q is independent of the rest of the system, then q should be included in $\mathbf{z}$, and cannot be independent of $\left(\mathbf{z}', \mathbf{x}'\right)'$. This condition also implies the usual assumption that $\mathbf{z}$ cannot be independent of the endogenous variables $\left(x', q\right)'$.[12] If this were the case, then $\underline{G}_{2,\gamma_0} = \mathbb{E}\left[\mathbf{z}\right]\mathbb{E}\left[\mathbf{x}'_{\leq\gamma_0}\right]$ would span a one-dimensional space, which can obviously be spanned by $\underline{G}_1 = \mathbb{E}\left[\mathbf{z}\right]\mathbb{E}\left[\mathbf{x}'\right]$ and $\underline{G}_{2,\gamma} = \mathbb{E}\left[\mathbf{z}\right]\mathbb{E}\left[\mathbf{x}'_{\leq\gamma}\right]$.

Theorem 1 *Under Assumptions F, I, IV and S,* $\widehat{\theta}$ *and especially* $\widehat{\gamma}$ *are consistent and have limit distribution given by*

$$\begin{pmatrix} \sqrt{n} I_{2\overline{d}} & \mathbf{0} \\ \mathbf{0} & \sqrt{n}\left\|\delta_n\right\| \end{pmatrix} \begin{pmatrix} \widehat{\underline{\theta}} - \underline{\theta}_0 \\ \widehat{\gamma} - \gamma_0 \end{pmatrix} \xrightarrow{d} N\left(\mathbf{0}, V\right)$$

where $V = (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}$.

Note that we need only Assumption I′ to show the consistency of $\widehat{\gamma}$. But to derive the asymptotic distribution we need Assumption I. Otherwise, G need not be of full column rank.[13] Also, as predicted in Section 2.1, $f_q\left(\gamma_0\right)$ appears in V and $\widehat{\gamma}$ is not asymptotically independent of $\left(\widehat{\beta}, \widehat{\delta}\right)$. $\widehat{W}$ can be any positive definite matrix besides $\left(n^{-1}Z'Z\right)^{-1}$. We still use $\widehat{W}$ to denote such a general weight matrix and use W to denote its limit.

To provide some intuition on the asymptotic variance of $\widehat{\gamma}$, consider the simple example in Section 2.1 again. In this example, $G = f_q\left(\gamma_0\right)$, $\Omega = Var\left(\varepsilon\right)$ and W is irrelevant, so $V = Var\left(\varepsilon\right)/f_q\left(\gamma_0\right)^2$. In fact, $\widehat{\gamma} = \widehat{F}_q^{-1}(\overline{y})$, so $\sqrt{n}\left(\widehat{\gamma} - \gamma_0\right) = \sqrt{n}\left[\widehat{F}_q^{-1}(\overline{y}) - F_q^{-1}(\overline{y}) - \left(\widehat{F}_q^{-1}(\mathbb{E}\left[y\right]) - F_q^{-1}(\mathbb{E}\left[y\right])\right)\right] + \sqrt{n}\left(\widehat{F}_q^{-1}(\mathbb{E}\left[y\right]) - F_q^{-1}(\mathbb{E}\left[y\right])\right) + \sqrt{n}\left(F_q^{-1}(\overline{y}) - F_q^{-1}(\mathbb{E}\left[y\right])\right)$. The first term is, roughly speaking, $\sqrt{n}\left(\sum_{i=1}^n \left(\psi_i\left(\overline{y}\right) - \psi_i\left(\mathbb{E}\left[y\right]\right)\right)\right)$ with $\psi_i\left(\tau\right) =$

[11] Note that zero $\mathbf{x}'\delta_n | q = \gamma_0$ does not imply $E\left[\mathbf{z}\mathbf{x}'_{\leq\gamma_0}\right]\delta_n = \mathbf{0}$ or $E\left[\mathbf{z}\mathbf{x}'_{>\gamma_0}\right]\delta_n = \mathbf{0}$.

[12] In the nonlinear scenario, uncorrelatedness in the linear scenario should be strengthened to independence. Also, all elements of $\left(x', q\right)'$ should be endogenous; otherwise, $\mathbf{z}$ should include the exogenous elements of $\left(x', q\right)'$ and cannot be independent of $\left(x', q\right)'$.

[13] In Remark 2 of Seo and Shin (2016) where their FD-GMM estimator, which is similar to our estimator, is used to estimate the dynamic panel threshold regression, they claim that the asymptotic distribution of $\widehat{\theta}$ is invariant to whether the model is CTR. That statement is not correct as can be seen by noting that their $G_\gamma\left(\gamma_0\right) = \mathbf{0}$ in CTR so that the asymptotic variance matrix of $\widehat{\theta}$ is undefined. The zeroness of $G_\gamma\left(\gamma_0\right)$ is due to some redundancy in the parameter θ when the model is CTR. If we rewrite the CTR as $y = \mathbf{x}'\beta + (q - \gamma)\delta 1\left(q \leq \gamma\right) + \varepsilon$, then the corresponding G under the moment conditions $E\left[\mathbf{z}\varepsilon\right] = \mathbf{0}$ is $\left(\ \mathbb{E}\left[\mathbf{z}\mathbf{x}'\right], \mathbb{E}\left[\mathbf{z}_{\leq\gamma_0}\left(q - \gamma_0\right)\right], \left(\mathbb{E}\left[\mathbf{z}\left(q - \gamma_0\right)|q = \gamma_0\right] f_q\left(\gamma_0\right) - E\left[\mathbf{z}_{\leq\gamma_0}\right]\right)\delta_0\ \right)$, which is of full column rank.

$\frac{\tau-1(q_i\leq F_q^{-1}(\tau))}{f_q(\gamma_0)}$. By a stochastic equicontinuity argument this term is $o_p(1)$, and so the asymptotic distribution of $\sqrt{n}(\widehat{\gamma}-\gamma_0)$ is the same as that of $\sqrt{n}\left(\widehat{F}_q^{-1}(\mathbb{E}[y])-F_q^{-1}(\mathbb{E}[y])\right)+\sqrt{n}\left(F_q^{-1}(\overline{y})-F_q^{-1}(\mathbb{E}[y])\right)$, where the first term represents the randomness in $\widehat{F}_q^{-1}$ and the second term represents the randomness in $\overline{y}$ (recall that $\widehat{\gamma}=\widehat{F}_q^{-1}(\overline{y})$). By the Bahadur representation, $\sqrt{n}\left(\widehat{F}_q^{-1}(\mathbb{E}[y])-F_q^{-1}(\mathbb{E}[y])\right)\approx\frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{F_q(\gamma_0)-1(q_i\leq\gamma_0)}{f_q(\gamma_0)}$, and by the Delta method, $\sqrt{n}\left(F_q^{-1}(\overline{y})-F_q^{-1}(\mathbb{E}[y])\right)\approx\frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{y_i-\mathbb{E}[y]}{f_q(\gamma_0)}=\frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{\varepsilon_i+1(q_i\leq\gamma_0)-F_q(\gamma_0)}{f_q(\gamma_0)}$. Hence, by the continuous mapping theorem (CMT), $\sqrt{n}(\widehat{\gamma}-\gamma_0)\approx\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\frac{F_q(\gamma_0)-1(q_i\leq\gamma_0)}{f_q(\gamma_0)}+\frac{\varepsilon_i+1(q_i\leq\gamma_0)-F_q(\gamma_0)}{f_q(\gamma_0)}\right)=\frac{\varepsilon_i}{f_q(\gamma_0)}$, and the asymptotic variance is $V=Var(\varepsilon)/f_q(\gamma_0)^2$. To consider the effect of δ_n on $\widehat{\gamma}$, suppose $y=\delta_n 1(q\leq\gamma_0)+\varepsilon$ with δ_n known. Then, by a similar argument, we can show $\sqrt{n}(\widehat{\gamma}-\gamma_0)\approx\frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{\varepsilon_i}{\delta_n f_q(\gamma_0)}$, so that the asymptotic variance of $\widehat{\gamma}$ is $O\left(\frac{1}{n\delta_n^2}\right)$. When δ_n is smaller, the asymptotic variance of $\widehat{\gamma}$ is larger, and when δ_n shrinks to zero the convergence rate of $\widehat{\gamma}$ is $\sqrt{n}|\delta_n|$.

By choosing $\widehat{W}=\widehat{\Omega}^{-1}$ with $\widehat{\Omega}=n^{-1}\sum_{i=1}^n\mathbf{z}_i\mathbf{z}_i'\widehat{\varepsilon}_i^2$ and $\widehat{\varepsilon}_i=y_i-\mathbf{x}_i'\widehat{\beta}-\mathbf{x}_{\leq\gamma,i}'\widehat{\delta}$, we get the asymptotically efficient estimator of θ_0 and the following result holds.

Corollary 1 *Under the same assumptions as in Theorem 1, if $\widehat{\theta}$ is estimated using $\widehat{Q}_n(\theta)$ with $\widehat{W}=\widehat{\Omega}^{-1}$, then*

$$\begin{pmatrix}\sqrt{n}I_{2\overline{d}} & \mathbf{0}\\ \mathbf{0} & \sqrt{n}\|\delta_n\|\end{pmatrix}\begin{pmatrix}\widehat{\underline{\theta}}-\underline{\theta}_0\\ \widehat{\gamma}-\gamma_0\end{pmatrix}\xrightarrow{d}N\left(\mathbf{0},(G'\Omega^{-1}G)^{-1}\right).$$

When the model is homoskedastic, i.e., $\mathbb{E}\left[\varepsilon^2|\mathbf{z}\right]=\sigma^2$, our 2SLS estimator is efficient. For inference concerning γ we suggest use of bootstrap methods such as in Hall and Horowitz (1996), Brown and Newey (2002) or Lee (2014) to avoid estimating $\mathbb{E}[\mathbf{z}\mathbf{x}'|q=\gamma_0]$ and $f_q(\gamma_0)$, for instance by numerical derivatives, as in Section 7.3 of Newey and McFadden (1994), or by some kernel or series method.

3.1 Comparison with the GMM of HHB and the 2SLS of CH

In the structural change context, HHB show that the GMM estimator based on the following criterion is generally inconsistent:

$$\widetilde{\gamma}=\arg\min_{\gamma}\widetilde{Q}_n(\gamma),$$

where

$$\widetilde{Q}_n(\gamma)=\min_{\beta_1,\beta_2}\widetilde{Q}_n(\theta):=\min_{\beta_1,\beta_2}\widetilde{m}_n(\theta)'\widetilde{W}\widetilde{m}_n(\theta)$$

with

$$\widetilde{m}_n(\theta)=\frac{1}{n}\sum_{i=1}^n m_i(\theta)=\frac{1}{n}\sum_{i=1}^n\begin{pmatrix}m_{1i}(\theta)\\ m_{2i}(\theta)\end{pmatrix}:=\frac{1}{n}\sum_{i=1}^n\mathbf{z}_i\begin{pmatrix}(y_i-\mathbf{x}_i'\beta_1)1(q_i\leq\gamma)\\ (y_i-\mathbf{x}_i'\beta_2)1(q_i>\gamma)\end{pmatrix}.$$

As commented by HHB, inconsistency of $\widetilde{\gamma}$ stems from the fact that the minimand is a quadratic form in the sample moment, thereby taking the form of a square of sums. This "square of sums" structure provides an opportunity for the effects of misspecification associated with the selection of the wrong threshold point to be an offsetting balancing factor in the minimand, leading to inconsistency. In contrast, the objective function of the 2SLS estimator in CH takes a "sum of squares" form, which generates a consistent estimator of γ. Specifically, the objective function of the 2SLS of CH is

$$\widehat{S}_n(\theta)=\frac{1}{n}\sum_{i=1}^n\left(y_i-\beta_1'\left(\widehat{\Pi}'\mathbf{z}_i\right)1(q_i\leq\gamma)-\beta_2'\left(\widehat{\Pi}'\mathbf{z}_i\right)1(q_i>\gamma)\right)^2,\tag{13}$$

where $\widehat{\Pi}'\mathbf{z}_i$ delivers a first-stage prediction of $\mathbf{x}_i$.[14] Given the comments by HHB, it may seem surprising that our GMM estimator is consistent even if the minimand is also a square of sums. The key point, however, is not the distinction between the "square of sums" and "sum of squares" criteria in this case, but rather the fact that the threshold variable q in the structural change model is a time index which is independent of the other components of the system (so that offsetting is possible, resulting in inconsistency).

Before a formal discussion on these points, note first that our 2SLS estimator is a special GMM estimator of HHB. Specifically, it is easy to check that when

$$\widetilde{W} = \begin{pmatrix} I_l \\ I_l \end{pmatrix} \widehat{W} \begin{pmatrix} I_l & I_l \end{pmatrix}, \tag{14}$$

$\widetilde{Q}_n(\theta) = \widehat{Q}_n(\theta)$. This $\widetilde{W}$ is only positive semidefinite, not positive definite. In other words, our 2SLS estimator does not fully explore the information in $\widetilde{m}_n(\theta)$. This is why we need $l > 2\overline{d}$ instruments, whereas HHB's GMM estimator needs only $l > \overline{d}$ instruments. This is also why our 2SLS estimator is not consistent when q has properties like a time index (because the general GMM estimator is *not* consistent). The moment conditions in $\widetilde{m}_n(\theta)$ explore the special structure of threshold regression - β_1 and β_2 are involved only in one regime of the system, while the moment conditions in $\widehat{g}_n(\theta)$ are designed for any nonlinear system $y = G(x, q; \theta) + \varepsilon$ with $\mathbb{E}[\varepsilon|x, q] \neq 0$. Essentially, the moment conditions $\widetilde{m}_n(\theta)$ explore the validity of the moments $\mathbb{E}\left[\mathbf{z}\varepsilon_{\leq\gamma_0}\right] = \mathbf{0}$ and $\mathbb{E}\left[\mathbf{z}\varepsilon_{>\gamma_0}\right] = \mathbf{0}$, whereas $\widehat{g}_n(\theta)$ explores only $\mathbb{E}[\mathbf{z}\varepsilon] = \mathbb{E}\left[\mathbf{z}\varepsilon_{\leq\gamma_0}\right] + \mathbb{E}\left[\mathbf{z}\varepsilon_{>\gamma_0}\right] = \mathbf{0}$.

We can now formally state the consistency of $\widetilde{\gamma}$ when q is not independent of $(\mathbf{z}', \mathbf{x}', \varepsilon)'$. First, we impose the following assumption. Because we concentrate on the identification issue below, we here assume that δ is fixed for notational simplicity.

Assumption IV′: $\dim(\mathbf{z}) = l \geq \overline{d} + 1$, $\widetilde{W} \xrightarrow{p} W > 0$,[15] *and* $E\left[\mathbf{z}\mathbf{x}'_{\leq\gamma}\right]$ *and* $E\left[\mathbf{z}\mathbf{x}'_{>\gamma}\right]$ *are of full column rank for any* $\gamma \in \Gamma$. *(i) If* q *is exogenous (i.e.,* q *is included in* z*, and* $E[\varepsilon|\mathbf{z}] = 0$*), then there does not exist* $a = (a_1', a_2')' \in \mathbb{R}^{2\overline{d}}$ *such that* $E\left[\mathbf{z}\mathbf{x}'_{>\gamma}\right] a_2 = E\left[\mathbf{z}\mathbf{x}'_{>\gamma_0}\right] \delta_0$ *for any* $\gamma < \gamma_0$ *or* $E\left[\mathbf{z}\mathbf{x}'_{\leq\gamma}\right] a_1 = E\left[\mathbf{z}\mathbf{x}'_{\leq\gamma_0}\right] \delta_0$ *any* $\gamma > \gamma_0$. *(ii) If* q *is endogenous and only* $E\left[\mathbf{z}\varepsilon_{\leq\gamma_0}\right] = 0$ *and* $E\left[\mathbf{z}\varepsilon_{>\gamma_0}\right] = 0$ *hold, then there does not exist* $a = (a_1', a_2')' \in \mathbb{R}^{2\overline{d}}$ *such that* $E\left[\mathbf{z}\mathbf{x}'_{\leq\gamma}\right] a_1 = E\left[\mathbf{z}\varepsilon_{\leq\gamma}\right], E\left[\mathbf{z}\mathbf{x}'_{>\gamma_0}\right] \delta_0 + E\left[\mathbf{z}\mathbf{x}'_{>\gamma}\right] a_2 = E\left[\mathbf{z}\varepsilon_{>\gamma}\right]$ *for any* $\gamma < \gamma_0$ *or* $E\left[\mathbf{z}\mathbf{x}'_{\leq\gamma}\right] a_1 - E\left[\mathbf{z}\mathbf{x}'_{\leq\gamma_0}\right] \delta_0 = E\left[\mathbf{z}\varepsilon_{\leq\gamma}\right]$ *and* $E\left[\mathbf{z}\mathbf{x}'_{>\gamma}\right] a_2 = E\left[\mathbf{z}\varepsilon_{>\gamma}\right]$ *for any* $\gamma > \gamma_0$.

Theorem 2 *Under Assumptions F, I′, IV′ and S,* $\widetilde{\gamma}$ *is consistent.*

HHB assume $\widetilde{W} = \operatorname{diag}\left(\widetilde{W}_1, \widetilde{W}_2\right)$ with $\widetilde{W}_1 \xrightarrow{p} W_1 > 0$ and $\widetilde{W}_2 \xrightarrow{p} W_2 > 0$, but we do not need such a restriction to show the consistency of $\widetilde{\gamma}$ or inconsistency of $\widetilde{\gamma}$ in the HHB setup. Similar to $\widehat{\gamma}$, we only require Assumption I′ rather than the stronger Assumption I to prove the consistency of $\widetilde{\gamma}$. In contrast to Assumption IV, we need different assumptions here for the identification of γ_0 depending on whether q is exogenous or not. In this sense, reducing $\widetilde{m}_n(\theta)$ to $\widehat{g}_n(\theta)$ makes the treatment of identification more uniform although there is some loss of information in doing so. When q is exogenous, Assumption IV′(i) requires some extra variation in $\mathbb{E}[\mathbf{z}\mathbf{x}'|q = \gamma]$ when γ moves away from γ_0. This condition implicitly precludes the possibility that q is independent of $(\mathbf{z}', \mathbf{x}')'$ because if this is the case, then $\mathbb{E}\left[\mathbf{z}\mathbf{x}'_{>\gamma}\right] = \mathbb{E}[\mathbf{z}\mathbf{x}'](1 - F_q(\gamma))$ and $\mathbb{E}\left[\mathbf{z}\mathbf{x}'_{>\gamma_0}\right] = \mathbb{E}[\mathbf{z}\mathbf{x}'](1 - F_q(\gamma_0))$, so a_2 can be chosen as $\frac{1-F_q(\gamma_0)}{1-F_q(\gamma)}\delta_0$ and, similarly, a_1 can be chosen as $\frac{F_q(\gamma_0)}{F_q(\gamma)}\delta_0$. When q is endogenous, we need also to take account of the variation in $\mathbb{E}[\mathbf{z}\varepsilon_{\leq\gamma}]$ and $\mathbb{E}[\mathbf{z}\varepsilon_{>\gamma}]$ as γ moves away from γ_0. We provide more intuition on such identifying information in the following discussion.

[14] Following the general setup of this paper we do not assume a threshold effect in the first stage.

[15] To save notation, we still use W to denote the limit of $\widetilde{W}$. This should not introduce any confusion.

The proof of the theorem also establishes the following results. First, if q is exogenous and also independent of $(\mathbf{z}', \mathbf{x}')'$, then γ_0 cannot be identified by $\widetilde{Q}_n(\gamma)$. This is essentially the case considered by HHB, and we label it case (o). Second, in case (i), both groups of moment conditions in $m_i(\theta)$ are required to identify γ_0; using only $m_{1i}(\theta)$ or $m_{2i}(\theta)$ is not enough. Third, in case (ii), either group of moment conditions in $m_i(\theta)$ can identify γ_0.[16] For example, if there does not exist $a_1 \in \mathbb{R}^{\overline{d}}$ such that $\mathbb{E}\left[\mathbf{z}\mathbf{x}'_{\leq\gamma}\right] a_1 = \mathbb{E}\left[\mathbf{z}\varepsilon_{\leq\gamma}\right]$ for any $\gamma < \gamma_0$ and $\mathbb{E}\left[\mathbf{z}\mathbf{x}'_{\leq\gamma}\right] a_1 - \mathbb{E}\left[\mathbf{z}\mathbf{x}'_{\leq\gamma_0}\right] \delta_0 = \mathbb{E}\left[\mathbf{z}\varepsilon_{\leq\gamma}\right]$ for any $\gamma > \gamma_0$, then γ_0 can be identified by only $m_{1i}(\theta)$. In comparison with case (i), we can see that the identifying power in either group of moment conditions in $m_i(\theta)$ comes solely from the correlation between q and ε. In other words, endogeneity is helpful in identifying γ_0 by moment conditions.

It seems that the correlation of q with the rest of the system is critical for the identification of γ_0. When q is independent of $(\mathbf{z}', \mathbf{x}', \varepsilon)'$, then even the combination of m_1 and m_2 cannot identify γ_0; if q is independent of ε but not $(\mathbf{z}', \mathbf{x}')'$, then combination of m_1 and m_2 can (but m_1 or m_2 individually cannot) identify γ_0; if q is correlated with all of $(\mathbf{z}', \mathbf{x}', \varepsilon)'$, then either m_1 or m_2 can identify γ_0.[17] What is the intuition here? We can understand these results by using Lemma 2.3 of Newey and McFadden (1994) which states that as long as $W\mathbb{E}[m_i(\theta)] \neq \mathbf{0}$ for $\theta \neq \theta_0$, then θ_0 is identified. In case (o), for any W, $W\mathbb{E}[m_i(\theta)] \neq \mathbf{0}$ for $\theta \neq \theta_0$ cannot hold. In case (i), when $W > 0$ or $W = \begin{pmatrix} I_l \\ I_l \end{pmatrix} W_0 \begin{pmatrix} I_l & I_l \end{pmatrix}$ for some $W_0 > 0$, $W\mathbb{E}[m_i(\theta)] \neq \mathbf{0}$ for $\theta \neq \theta_0$. In case (ii), when $W > 0$ or $W = \begin{pmatrix} I_l \\ I_l \end{pmatrix} W_0 \begin{pmatrix} I_l & I_l \end{pmatrix}$ for some $W_0 > 0$ or $W = \begin{pmatrix} W_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$ with $W_1 > 0$ or $W = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & W_2 \end{pmatrix}$ with $W_2 > 0$, $W\mathbb{E}[m_i(\theta)] \neq \mathbf{0}$ for $\theta \neq \theta_0$. To be specific, we identify γ_0 from the fact that

$$W\mathbb{E}\left[\mathbf{z}\begin{pmatrix} (y - \mathbf{x}'\beta_1) 1(q \leq \gamma) \\ (y - \mathbf{x}'\beta_2) 1(q > \gamma) \end{pmatrix}\right] = \mathbf{0}$$

only if $\gamma = \gamma_0$ for any β_1 and β_2, or equivalently,

$$W\begin{pmatrix} \mathbb{E}\left[\mathbf{z}\mathbf{x}'_{\leq\gamma}\right]\beta_1 \\ \mathbb{E}\left[\mathbf{z}\mathbf{x}'_{>\gamma}\right]\beta_2 \end{pmatrix} \neq W\mathbb{E}\begin{bmatrix} \mathbf{z}y_{\leq\gamma} \\ \mathbf{z}y_{>\gamma} \end{bmatrix}$$

when $\gamma \neq \gamma_0$ for any β_1 and β_2. Note that

$$\mathbb{E}\begin{bmatrix} \mathbf{z}y_{\leq\gamma} \\ \mathbf{z}y_{>\gamma} \end{bmatrix} = \mathbb{E}\begin{bmatrix} \mathbf{z}\left(\mathbf{x}'_{\leq\gamma_0}\beta_{10} + \mathbf{x}'_{>\gamma_0}\beta_{20} + \varepsilon\right) 1(q \leq \gamma) \\ \mathbf{z}\left(\mathbf{x}'_{\leq\gamma_0}\beta_{10} + \mathbf{x}'_{>\gamma_0}\beta_{20} + \varepsilon\right) 1(q > \gamma) \end{bmatrix},$$

which is equal to

$$\begin{pmatrix} \mathbb{E}\left[\mathbf{z}\mathbf{x}'_{\leq\gamma}\right]\beta_{10} + \mathbb{E}\left[\mathbf{z}\varepsilon_{\leq\gamma}\right] \\ \mathbb{E}\left[\mathbf{z}\mathbf{x}'_{\gamma<\leq\gamma_0}\right]\beta_{10} + \mathbb{E}\left[\mathbf{z}\mathbf{x}'_{>\gamma_0}\right]\beta_{20} + \mathbb{E}\left[\mathbf{z}\varepsilon_{>\gamma}\right] \end{pmatrix}$$

[16] Assume $\widetilde{W} = \text{diag}\left(\widetilde{W}_1, \widetilde{W}_2\right)$. Because there is no restriction on $\widetilde{W}_1$ and $\widetilde{W}_2$ to obtain the consistency of $\widetilde{\gamma}$ in both case (i) and case (ii), when $\widetilde{W}_1\left(\widetilde{W}_2\right)$ is much larger than $\widetilde{W}_2\left(\widetilde{W}_1\right)$, we are essentially using only $m_{1i}(\theta)(m_{2i}(\theta))$ in $\widetilde{Q}_n(\theta)$. In this sense, it is surprising to see that case (i) requires both $m_{1i}(\theta)$ and $m_{2i}(\theta)$, while case (ii) requires only $m_{1i}(\theta)$ or $m_{2i}(\theta)$. Essentially, the limiting behaviors of $\widetilde{Q}_n(\gamma)$ in these two cases are quite different; see the following discussion and example for more intuition on this point. Importantly, note that either $\widetilde{W}_1 = \mathbf{0}$ or $\widetilde{W}_2 = \mathbf{0}$ violates $\widetilde{W} > 0$, so Assumption IV$'$ does not hold and the identifiability of γ_0 cannot follow from Theorem 2 in this case. The new results here are that in case (ii), $\widetilde{W} > 0$ is not necessary for identification (actually, in case (i), $\widetilde{W} > 0$ is not necessary either, e.g., the $\widetilde{W}$ in (14) is not positive definite).

[17] In cases (o) and (i), we require only $\mathrm{E}[\varepsilon|\mathbf{z}, q] = 0$, and in case (ii), q can be independent of $(\mathbf{z}', \mathbf{x}')'$. Here, we use three sequentially stronger assumptions to distinguish these three cases.

when $\gamma < \gamma_0$, and equal to

$$\begin{pmatrix} \mathbb{E}\left[\mathbf{z}\mathbf{x}'_{\leq\gamma_0}\right]\beta_{10} + \mathbb{E}\left[\mathbf{z}\mathbf{x}'_{\gamma_0<\leq\gamma}\right]\beta_{20} + \mathbb{E}\left[\mathbf{z}\varepsilon_{\leq\gamma}\right] \\ \mathbb{E}\left[\mathbf{z}\mathbf{x}'_{>\gamma}\right]\beta_{20} + \mathbb{E}\left[\mathbf{z}\varepsilon_{>\gamma}\right] \end{pmatrix}$$

when $\gamma > \gamma_0$. In case (o), $\mathbb{E}\left[\mathbf{z}\varepsilon_{\leq\gamma}\right] = \mathbb{E}\left[\mathbf{z}\varepsilon_{>\gamma}\right] = \mathbf{0}$, $\mathbb{E}\left[\mathbf{z}\mathbf{x}'_{\leq\gamma}\right] = \mathbb{E}\left[\mathbf{z}\mathbf{x}'\right]F_q(\gamma)$, $\mathbb{E}\left[\mathbf{z}\mathbf{x}'_{>\gamma}\right] = \mathbb{E}\left[\mathbf{z}\mathbf{x}'\right](1 - F_q(\gamma))$ and $\mathbb{E}\left[\mathbf{z}\mathbf{x}'_{\gamma_1<\leq\gamma_2}\right] = \mathbb{E}\left[\mathbf{z}\mathbf{x}'\right](F_q(\gamma_2) - F_q(\gamma_1))$, where $\gamma_1 < \gamma_2$. So we can choose β_1 and β_2 such that

$$\beta_1 = \beta_{10} \text{ and } (1 - F_q(\gamma))\beta_2 = (F_q(\gamma_0) - F_q(\gamma))\beta_{10} + (1 - F_q(\gamma_0))\beta_{20} \tag{15}$$

when $\gamma < \gamma_0$ and

$$\beta_2 = \beta_{20} \text{ and } F_q(\gamma)\beta_1 = F_q(\gamma_0)\beta_{10} + (F_q(\gamma) - F_q(\gamma_0))\beta_{20} \tag{16}$$

when $\gamma > \gamma_0$ to make the equalities hold.[18] In other words, $\text{plim}\widetilde{Q}_n(\gamma) = 0$ for any γ. In case (i), $\mathbb{E}\left[\mathbf{z}\varepsilon_{\leq\gamma}\right] = \mathbb{E}\left[\mathbf{z}\varepsilon_{>\gamma}\right] = \mathbf{0}$. So when $\gamma < \gamma_0$, we can choose $\beta_1 = \beta_{10}$ to make $\mathbb{E}\left[\mathbf{z}\mathbf{x}'_{\leq\gamma}\right]\beta_1 = \mathbb{E}\left[\mathbf{z}y_{\leq\gamma}\right]$ but cannot choose β_2 such that $\mathbb{E}\left[\mathbf{z}\mathbf{x}'_{>\gamma}\right]\beta_2 = \mathbb{E}\left[\mathbf{z}y_{>\gamma}\right]$, and when $\gamma > \gamma_0$, we can choose $\beta_2 = \beta_{20}$ to make $\mathbb{E}\left[\mathbf{z}\mathbf{x}'_{>\gamma}\right]\beta_2 = \mathbb{E}\left[\mathbf{z}y_{>\gamma}\right]$ but cannot choose β_1 such that $\mathbb{E}\left[\mathbf{z}\mathbf{x}'_{\leq\gamma}\right]\beta_1 = \mathbb{E}\left[\mathbf{z}y_{\leq\gamma}\right]$. In other words, if we use only m_1, then $\text{plim}\widetilde{Q}_n(\gamma) = 0$ for $\gamma \in \left[\underline{\gamma}, \gamma_0\right]$ and if we use only m_2, then $\text{plim}\widetilde{Q}_n(\gamma) = 0$ on $[\gamma_0, \overline{\gamma}]$, while if we use both m_1 and m_2, then $\text{plim}\widetilde{Q}_n(\gamma) = 0$ only if $\gamma = \gamma_0$. In case (ii), $\mathbb{E}\left[\mathbf{z}\varepsilon_{\leq\gamma}\right] \neq \mathbf{0}$ and $\mathbb{E}\left[\mathbf{z}\varepsilon_{>\gamma}\right] \neq \mathbf{0}$. So even if we use only m_1 or m_2, the equalities can hold only at $\gamma = \gamma_0$.

The above arguments also show a key difference between the identification sources of the HHB GMM estimator and the CH 2SLS estimator. In CH,

$$\text{plim}\widehat{S}_n(\theta) = \mathbb{E}\left[\left(y - \beta_1'(\Pi'\mathbf{z})1(q \leq \gamma) - \beta_2'(\Pi'\mathbf{z})1(q > \gamma)\right)^2\right],$$

which assumes that $\mathbb{E}[y|\mathbf{z}, q] = \beta_1'(\Pi'\mathbf{z})1(q \leq \gamma) + \beta_2'(\Pi'\mathbf{z})1(q > \gamma)$ and uses the conditional mean difference of y below γ_0 and above γ_0 to identify γ_0 (just as in standard least squares estimation where $\mathbb{E}[\varepsilon|\mathbf{x}] = 0$ and we can calibrate y against its conditional mean to identify the parameters in the conditional mean). Since $y = \beta_1'(\Pi'\mathbf{z} + \mathbf{u})1(q \leq \gamma) + \beta_2'(\Pi'\mathbf{z} + \mathbf{u})1(q > \gamma) + \varepsilon$, where the first stage regression is assumed to be $\mathbf{x} = \Pi'\mathbf{z} + \mathbf{u}$, we must assume $\mathbb{E}[\mathbf{u}|\mathbf{z}, q] = \mathbf{0}$ and $\mathbb{E}[\varepsilon|\mathbf{z}, q] = 0$ and then the conditional mean of y is $\beta_1'(\Pi'\mathbf{z})1(q \leq \gamma) + \beta_2'(\Pi'\mathbf{z})1(q > \gamma)$. To achieve such conditions, we must assume that q is exogenous so that it can be included in $\mathbf{z}$. Also, as argued in Yu (2013a), the first stage must be a regression rather than a projection, i.e., $\mathbb{E}[\mathbf{u}|\mathbf{z}] = \mathbf{0}$ rather than only $\mathbb{E}[\mathbf{z}\mathbf{u}'] = \mathbf{0}$. In a nonlinear environment, such a requirement does not seem too stringent. On the contrary, the identification of γ_0 by HHB's GMM is based on the matching of covariances just as in the usual linear GMM estimation. If γ_0 were known, we can identify β_1 by matching $\mathbb{E}\left[\mathbf{z}y_{\leq\gamma_0}\right]$ with $\mathbb{E}\left[\mathbf{z}\mathbf{x}'_{\leq\gamma_0}\right]\beta_1$ and β_2 by matching $\mathbb{E}\left[\mathbf{z}y_{>\gamma_0}\right]$ with $\mathbb{E}\left[\mathbf{z}\mathbf{x}'_{>\gamma_0}\right]\beta_2$. It is the nonlinear structure introduced by the unknown γ that necessitates the division of identification into three different cases; in such a nonlinear system, endogeneity of q is helpful rather than harmful to identification as in CH's 2SLS. As far as inference is concerned, the bootstrap is questionable for CH's 2SLS given the negative findings in Yu (2014) where it is shown that the bootstrap for γ when the objective function takes the "sum of squares" form is invalid.

As for the requirement on the number of instruments, CH's assumption that $E[\Pi'\mathbf{z}\mathbf{z}'\Pi|q = \gamma_0] = \Pi'E[\mathbf{z}\mathbf{z}'|q = \gamma_0]\Pi > 0$ implies $l \geq \overline{d}$ instruments are needed. As mentioned in Assumptions IV and IV′, $l \geq 2\overline{d} + 1$ instruments are required in our 2SLS because $g_i(\theta)$ contains $2\overline{d} + 1$ unknown parameters, and $l \geq \overline{d} + 1$ instruments are required in HHB's GMM because each of $m_{1i}(\theta)$ and $m_{2i}(\theta)$ contains $\overline{d} + 1$

[18] Note that we can choose β_1 and β_2 freely, so the choice of β_1 and β_2 depends on γ here.

unknown parameters. On the other hand, more instruments imply more identification power: CH's 2SLS cannot handle the endogenous q case, while the other two estimators can; as discussed before Assumption IV′, HHB's GMM relies on the special structure of (1) while our 2SLS can handle any nonlinear system $y = G(x, q; \theta) + \varepsilon$ with $\mathbb{E}[\varepsilon|x, q] \neq 0$.

It is well known that more moment conditions generally imply higher asymptotic efficiency. Why then is our 2SLS the suggested approach rather than HHB's GMM? The reason is that the derivative $d\mathbb{E}[m_i(\theta_0)]/d\theta'$ does not exist as is normally required in usual GMM asymptotic derivations. Specifically,

$$\left.\frac{\partial \mathbb{E}[m_i(\beta_0, \delta_0, \gamma)]}{\partial \gamma}\right|_{\gamma=\gamma_0+} = \begin{pmatrix} \mathbb{E}[\mathbf{z}\varepsilon|q=\gamma_0] f_q(\gamma_0) - \mathbb{E}[\mathbf{z}\mathbf{x}'|q=\gamma_0]\delta_0 f_q(\gamma_0) \\ -\mathbb{E}[\mathbf{z}\varepsilon|q=\gamma_0] f_q(\gamma_0) \end{pmatrix}, \tag{17}$$

whereas

$$\left.\frac{\partial \mathbb{E}[m_i(\beta_0, \delta_0, \gamma)]}{\partial \gamma}\right|_{\gamma=\gamma_0-} = \begin{pmatrix} \mathbb{E}[\mathbf{z}\varepsilon|q=\gamma_0] f_q(\gamma_0) \\ -\mathbb{E}[\mathbf{z}\varepsilon|q=\gamma_0] f_q(\gamma_0) - \mathbb{E}[\mathbf{z}\mathbf{x}'|q=\gamma_0]\delta_0 f_q(\gamma_0) \end{pmatrix}. \tag{18}$$

Here, note that $\mathbb{E}[\mathbf{z}\varepsilon 1(q \leq \gamma_0)] = \mathbf{0}$ and $\mathbb{E}[\mathbf{z}\varepsilon 1(q > \gamma_0)] = \mathbf{0}$ do not imply $\mathbb{E}[\mathbf{z}\varepsilon|q=\gamma_0] = \mathbf{0}$. Even if $\mathbb{E}[\mathbf{z}\varepsilon|q=\gamma_0] = \mathbf{0}$ as in case (i), if $\mathbb{E}[\mathbf{z}\mathbf{x}'|q=\gamma_0]$ is of full column rank, the derivative $\partial\mathbb{E}[m_i(\beta_0, \delta_0, \gamma_0)]/\partial\gamma$ does not exist. This makes the asymptotic distribution of $\widetilde{\gamma}$ a nonnormal mixture that depends on the one-sided derivatives, rendering inference based on $\widetilde{\gamma}$ difficult.[19] On the contrary, in our 2SLS approach we have

$$\begin{aligned} \left.\frac{\partial \mathbb{E}[g_i(\beta_0, \delta_0, \gamma)]}{\partial \gamma}\right|_{\gamma=\gamma_0+} &= (I_l, I_l) \left.\frac{\partial \mathbb{E}[m_i(\beta_0, \delta_0, \gamma)]}{\partial \gamma}\right|_{\gamma=\gamma_0+} = -\mathbb{E}[\mathbf{z}\mathbf{x}'|q=\gamma_0]\delta_0 f_q(\gamma_0) \\ &= (I_l, I_l) \left.\frac{\partial \mathbb{E}[m_i(\beta_0, \delta_0, \gamma)]}{\partial \gamma}\right|_{\gamma=\gamma_0-}, \end{aligned}$$

which makes bootstrap inference valid.

3.2 Extensions

We discuss two extensions on identification based on moment conditions in this subsection. As mentioned above, more moment conditions typically imply higher asymptotic efficiency, but such results rely in the first place on identification. With 2SLS, even if $\mathbb{E}[\mathbf{z}\varepsilon] = \mathbf{0}$ is replaced by $\mathbb{E}[\varepsilon|\mathbf{z}] = 0$ which implies more moment conditions, the identification results are unaltered. For example, when q is independent of $(\mathbf{z}', \mathbf{x}')'$, γ_0 is not identified even by $\mathbb{E}[\varepsilon|\mathbf{z}] = 0$. Similar identification results apply to HHB's GMM.

In dynamic panel threshold regression,

$$\begin{aligned} y_{it} &= (1, x_{it}')\beta_1 1(q_{it} \leq \gamma) + (1, x_{it}')\beta_2 1(q_{it} > \gamma) + \varepsilon_{it}, \\ i &= 1, \cdots, n, t = 1, \cdots, T, \end{aligned}$$

where $\varepsilon_{it} = \alpha_i + v_{it}$, α_i is the fixed effet, v_{it} is the idiosyncratic disturbance, and x_{it} may contain lagged y_{it}'s, Seo and Shin (2016) apply the FD-GMM estimator of Arellano and Bond (1991) to estimate γ. The moment conditions are

$$E[g_i(\theta)] := E\begin{bmatrix} z_{it_0}\left(\Delta y_{it_0} - \beta_{22}'\Delta x_{it_0} - \delta' X_{it_0} 1_{it_0}(\gamma)\right) \\ \vdots \\ z_{iT}\left(\Delta y_{iT} - \beta_{22}'\Delta x_{iT} - \delta' X_{iT} 1_{iT}(\gamma)\right) \end{bmatrix} = \mathbf{0},$$

[19] We will discuss the asymptotic properties of $\widetilde{\gamma}$ and bootstrap inference based on $\widetilde{\gamma}$ in a separate paper.

where $\beta_2 = \left(\beta_{21}, \beta_{22}'\right)'$, $\delta = \beta_1 - \beta_2$, $\theta = \left(\beta_{22}', \delta', \gamma\right)$, $2 < t_0 \leq T$, $X_{it} = \begin{pmatrix} (1, x_{it}') \\ (1, x_{i,t-1}') \end{pmatrix}$, $1_{it}(\gamma) = \begin{pmatrix} 1(q_{it} \leq \gamma) \\ -1(q_{i,t-1} \leq \gamma) \end{pmatrix}$, and Δ is the first difference operator, and the FD-GMM estimator

$$\widehat{\theta} = \arg\min_{\theta} J_n(\theta),$$

where

$$J_n(\theta) = \overline{g}_n(\theta)' W_n \overline{g}_n(\theta)$$

with $\overline{g}_n(\theta) = \frac{1}{n}\sum_{i=1}^n g_i(\theta)$, and $W_n \xrightarrow{p} W > 0$.

The identification issue still exists even for such general moment conditions. Due to a mistake in the proof of their Theorem 1 (as detailed in the proof of our Theorem 1), Seo and Shin (2016) does not exclude the unidentifiable case in their assumptions. One important unidentifiable case is stated in the following corollary.

Corollary 2 *If q_{it} is independent of $\{z_{it}\}_{t=t_0}^T$ and $\{x_{it}\}_{t=t_0-1}^T$ and has a stationary distribution over $t = t_0 - 1, \cdots, T$, and $E[z_{it}\Delta v_{it}] = 0$ for $t = t_0, \cdots, T$, then γ is not identifiable by $J_n(\theta)$.*

One typical case for q_{it} in Corollary 2 is the time index, i.e., the model considered is the dynamic panel model with structural change. Note also that we do not require q_{it} to be exogenous as in case (o), which implies that our 2SLS estimator is unidentifiable even if q is endogenous as long as q is independent of $\mathbf{z}$ and $\mathbf{x}$. Since q_{it} is independent of other regressors and instruments, it is similar to the Lewbel's special regressor (see, e.g., Lewbel (2014) for a summary) in appearance; however, Lewbel's special regressor is to provide identification while q_{it} here makes identification fail.

3.3 A Simple Illustration

We illustrate the identification results above based on the example in Section 2.1. In this example $y = 1(q \leq \gamma_0) + \varepsilon$, where $q \sim U[0,1]$, $\gamma_0 = 1/2$, $\beta_0 = 0$ and $\delta_0 = 1$ are known, $\mathbf{x} = 1$, and $Var(\varepsilon) = 1$.

First assume $\mathbb{E}[\varepsilon|q] = 0$, so there is no endogeneity. Let $z = 1$, giving case (o) with $\mathbb{E}[\varepsilon|q, z] = 0$ and $q \perp (z, x)$. The moment conditions used for identifying γ_0 are

$$\mathbb{E}\left[\begin{pmatrix} (y-1)\,1(q \leq \gamma) \\ y1(q > \gamma) \end{pmatrix}\right] = \mathbf{0}. \tag{19}$$

Suppose $\widetilde{W} = I_2$. Then

$$\text{plim}\widetilde{Q}_n(\gamma) = \mathbb{E}[(y-1)\,1(q \leq \gamma)]^2 + \mathbb{E}[y1(q > \gamma)]^2.$$

After some algebra,

$$\text{plim}\widetilde{Q}_n(\gamma) = \left[-\left(\gamma - \frac{1}{2}\right)_+\right]^2 + \left(\frac{1}{2} - \gamma\right)_+^2 = \left(\frac{1}{2} - \gamma\right)^2,$$

where for $a \in \mathbb{R}$, $a_+ = \max(a, 0)$, and $\frac{1}{2} - \gamma = -\left(\gamma - \frac{1}{2}\right)_+ + \left(\frac{1}{2} - \gamma\right)_+$. Obviously, $\arg\min_{\gamma \in \Gamma} \text{plim}\widetilde{Q}_n(\gamma) = 1/2$. This seems to contradict the nonidentification result of HHB in case (o). In fact, this outcome is because β_0 and δ_0 are known. In (15) and (16), β_1 and β_2 are fixed at $\beta_{10} = 1$ and $\beta_{20} = 0$. So when $\gamma < \gamma_0$, $(1 - F_q(\gamma))\beta_{20} = (F_q(\gamma_0) - F_q(\gamma))\beta_{10} + (1 - F_q(\gamma_0))\beta_{20}$ or $(F_q(\gamma_0) - F_q(\gamma))\delta_0 = 0$ cannot hold as long as $\delta_0 \neq 0$ and $f_q(\gamma) > 0$ on $\left[\underline{\gamma}, \gamma_0\right]$. Similarly, when $\gamma > \gamma_0$, $F_q(\gamma)\beta_1 = F_q(\gamma_0)\beta_{10} + (F_q(\gamma) - F_q(\gamma_0))\beta_{20}$

or $(F_q(\gamma) - F_q(\gamma_0))\delta_0 = 0$ cannot hold as long as $\delta_0 \neq 0$ and $f_q(\gamma) > 0$ on $[\gamma_0, \overline{\gamma}]$. If they can be chosen freely, then it is obvious that the system cannot be identified - there are two equations and three unknowns.

Next, let $\mathbf{z} = (1, q)'$ as in case (i), for which the moment conditions used for identifying γ_0 are

$$\mathbb{E}\left[\begin{pmatrix} \mathbf{z}(y-1)1(q \leq \gamma) \\ \mathbf{z}y1(q > \gamma) \end{pmatrix}\right] = \mathbf{0}.$$

Suppose $\widetilde{W} = \begin{pmatrix} \widetilde{W}_1 & \widetilde{W}_{12} \\ \widetilde{W}_{12}' & \widetilde{W}_2 \end{pmatrix} = \begin{pmatrix} c_1 I_2 & \widetilde{W}_{12} \\ \widetilde{W}_{12}' & c_2 I_2 \end{pmatrix}$, and then

$$\begin{aligned} \text{plim}\widetilde{Q}_n(\gamma) &= c_1\left[-\left(\gamma - \frac{1}{2}\right)_+\right]^2 + c_1\left[-\left(\frac{1}{2}\gamma^2 - \frac{1}{8}\right)_+\right]^2 + c_2\left(\frac{1}{2} - \gamma\right)_+^2 + c_2\left(\frac{1}{8} - \frac{1}{2}\gamma^2\right)_+^2 \\ &\quad + 2\left(-\left(\gamma - \frac{1}{2}\right)_+, -\left(\frac{1}{2}\gamma^2 - \frac{1}{8}\right)_+\right)\widetilde{W}_{12}\left(\left(\frac{1}{2} - \gamma\right)_+, \left(\frac{1}{8} - \frac{1}{2}\gamma^2\right)_+\right)'. \end{aligned}$$

If $\widetilde{W}_{12} = \mathbf{0}$, $c_1 = 1$ and $c_2 = 0$, then we use only $m_1(\gamma)$ and Figure 1 shows that $\arg\min_{\gamma\in\Gamma} \text{plim}\widetilde{Q}_n(\gamma) = [0, 1/2]$. If $\widetilde{W}_{12} = \mathbf{0}$, $c_1 = 0$ and $c_2 = 1$, then we use only $m_2(\gamma)$ and Figure 1 shows that $\arg\min_{\gamma\in\Gamma} \text{plim}\widetilde{Q}_n(\gamma) = [1/2, 1]$.[20] If $c_1 \neq 0$ and $c_2 \neq 0$, then we use both $m_1(\gamma)$ and $m_2(\gamma)$ and Figure 1 shows that when either $c_1 = 1$, $c_2 = 2$ and $\widetilde{W}_{12} = \mathbf{0}$ or $c_1 = 1$, $c_2 = 1.5$ and $\widetilde{W}_{12} = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$, $\arg\min_{\gamma\in\Gamma} \text{plim}\widetilde{Q}_n(\gamma) = 1/2$.[21]

Section 3.3 of Yu (2015b) considers the following joint distribution of (q, ε):

$$f_{q,\varepsilon}(q, \varepsilon) = \begin{cases} \phi(\varepsilon - 1/2), & \text{if } 0 \leq q < \frac{1}{4}, \\ \phi(\varepsilon + 1/2), & \text{if } \frac{1}{4} \leq q \leq \frac{1}{2}, \\ \phi(\varepsilon - 1/2), & \text{if } \frac{1}{2} < q \leq \frac{3}{4}, \\ \phi(\varepsilon + 1/2), & \text{if } \frac{3}{4} < q \leq 1, \\ 0, & \text{otherwise,} \end{cases} \tag{20}$$

where $\phi(\cdot)$ is the standard normal density. Obviously, $\mathbb{E}[\varepsilon|q] \neq 0$, and $\mathbb{E}[\varepsilon 1(q \leq \gamma_0)] = \mathbb{E}[\varepsilon 1(q > \gamma_0)] = 0$, so this is case (ii). Suppose $z = 1$. Then the moment conditions used for identifying γ_0 are (19). Suppose $\widetilde{W} = \begin{pmatrix} A & C \\ C & B \end{pmatrix}$. Then

$$\text{plim}\widetilde{Q}_n(\gamma) = A\mathbb{E}[(y-1)1(q \leq \gamma)]^2 + B\mathbb{E}[y1(q > \gamma)]^2 + 2C\mathbb{E}[(y-1)1(q \leq \gamma)]\mathbb{E}[y1(q > \gamma)],$$

and with some algebra this reduces to

$$\text{plim}\widetilde{Q}_n(\gamma) = \begin{cases} A\left(\frac{\gamma}{2}\right)^2 + B\left(\frac{1-3\gamma}{2}\right)^2 + 2C\left(\frac{\gamma}{2}\right)\left(\frac{1-3\gamma}{2}\right), & \text{if } 0 \leq \gamma < \frac{1}{4}, \\ (A + B + 2C)\left(\frac{1}{4} - \frac{\gamma}{2}\right)^2, & \text{if } \frac{1}{4} \leq \gamma \leq \frac{3}{4}, \\ A\left(1 - \frac{3\gamma}{2}\right)^2 + B\left(\frac{\gamma-1}{2}\right)^2 + 2C\left(1 - \frac{3\gamma}{2}\right)\left(\frac{\gamma-1}{2}\right), & \text{if } \frac{3}{4} < \gamma \leq 1, \end{cases}$$

Figure 2 shows that when $A = 1$, $B = 2$ and $C = 0$, $\arg\min_{\gamma\in\Gamma} \text{plim}\widetilde{Q}_n(\gamma) = 1/2$. Actually, if only m_1 is used

[20] This implies that in case (o) (i.e., $z = 1$), using only one moment condition cannot identify γ_0.

[21] In this example, $\widetilde{W}_{12}$ does not play any role, i.e., $\text{plim}\widetilde{Q}_n(\gamma)$ depends only on $\widetilde{W}_1$ and $\widetilde{W}_2$ because the last term of $\text{plim}\widetilde{Q}_n(\gamma)$ is zero.

(i.e., $A=1$, $B=0$ and $C=0$) or only m_2 is used (i.e., $A=0$, $B=1$ and $C=0$), $\arg\min_{\gamma\in\Gamma} \text{plim}\widetilde{Q}_n(\gamma) = 1/2$, where the parameter space Γ excludes the neighborhoods of 0 and 1. If $C \neq 0$, $\arg\min_{\gamma\in\Gamma} \text{plim}\widetilde{Q}_n(\gamma) = \gamma_0$ as long as $\widetilde{W} > 0$. For instance, Figure 2 shows that when $A=1$, $B=2$ and $C=1$, $\arg\min_{\gamma\in\Gamma} \text{plim}\widetilde{Q}_n(\gamma) = 1/2$.

We now check the behavior of the 2SLS estimators of this paper and CH for these three cases: (o) $\mathbb{E}[\varepsilon|q]=0$ and $z=1$; (i) $\mathbb{E}[\varepsilon|q]=0$ and $\mathbf{z}=(1,q)'$; and (ii) $\mathbb{E}[\varepsilon|q]\neq 0$ and $z=1$. For 2SLS, the moment conditions are $\mathbb{E}[y-1(q\leq\gamma)]=0$ in cases (o) and (ii) and $\mathbb{E}[\mathbf{z}(y-1(q\leq\gamma))]=\mathbf{0}$ in case (i). In case (o),

$$\text{plim}\widehat{Q}_n(\gamma) = \left(\frac{1}{2}-\gamma\right)^2,$$

the same as $\text{plim}\widetilde{Q}_n(\gamma)$. In case (i) with $\widehat{W}=I_2$,

$$\text{plim}\widehat{Q}_n(\gamma) = \left(\frac{1}{2}-\gamma\right)^2 + \left(\frac{1}{8}-\frac{1}{2}\gamma^2\right)^2,$$

where for comparison with $\text{plim}\widetilde{Q}_n(\gamma)$, note that $\frac{1}{8}-\frac{1}{2}\gamma^2 = -\left(\frac{1}{2}\gamma^2-\frac{1}{8}\right)_+ + \left(\frac{1}{8}-\frac{1}{2}\gamma^2\right)_+$. In case (ii),

$$\text{plim}\widehat{Q}_n(\gamma) = \left(\frac{1}{2}-\gamma\right)^2,$$

the same as in case (o). For CH's 2SLS,

$$\text{plim}\widehat{S}_n(\gamma) = \mathbb{E}\left[(y-1(q\leq\gamma))^2\right]$$

in all three cases. In cases (o) and (i),

$$\text{plim}\widehat{S}_n(\gamma) = 1+|\gamma-1/2|;$$

in case (ii),

$$\text{plim}\widehat{S}_n(\gamma) = \begin{cases} \frac{7}{4}-2\gamma, & \text{if } 0\leq\gamma<\frac{1}{4}, \\ \frac{5}{4}, & \text{if } \frac{1}{4}\leq\gamma\leq\frac{3}{4}, \\ 2\gamma-\frac{1}{4}, & \text{if } \frac{3}{4}<\gamma\leq 1. \end{cases}$$

Figure 3 shows $\text{plim}\widetilde{Q}_n(\gamma)$ and $\text{plim}\widehat{S}_n(\gamma)$ in these three cases. For our 2SLS estimator, $\arg\min_{\gamma\in\Gamma} \text{plim}\widetilde{Q}_n(\gamma) = 1/2$ in all cases, giving the same identifying results as HHB's GMM using both m_1 and m_2. For CH's 2SLS, $\arg\min_{\gamma\in\Gamma} \text{plim}\widehat{S}_n(\gamma) = 1/2$ in cases (o) and (i), whereas $\arg\min_{\gamma\in\Gamma} \text{plim}\widehat{S}_n(\gamma) = [1/4, 3/4]$ in case (ii), which is unidentified.

We next check whether the expected moment conditions are differentiable. In case (i), for HHB's GMM,

$$\mathbb{E}[m_i(\beta_0,\delta_0,\gamma)] = \begin{pmatrix} -\left(\gamma-\frac{1}{2}\right)_+ \\ -\left(\frac{1}{2}\gamma^2-\frac{1}{8}\right)_+ \\ \left(\frac{1}{2}-\gamma\right)_+ \\ \left(\frac{1}{8}-\frac{1}{2}\gamma^2\right)_+ \end{pmatrix}$$

is not differentiable at $\gamma_0=1/2$, whereas for our 2SLS $\mathbb{E}[g_i(\beta_0,\delta_0,\gamma)] = \begin{pmatrix} \frac{1}{2}-\gamma \\ \frac{1}{8}-\frac{1}{2}\gamma^2 \end{pmatrix}$, which is differentiable at $\gamma_0=1/2$. In case (ii), $E[m_i(\beta_0,\delta_0,\gamma)]$ is differentiable at γ_0. This is due to the special

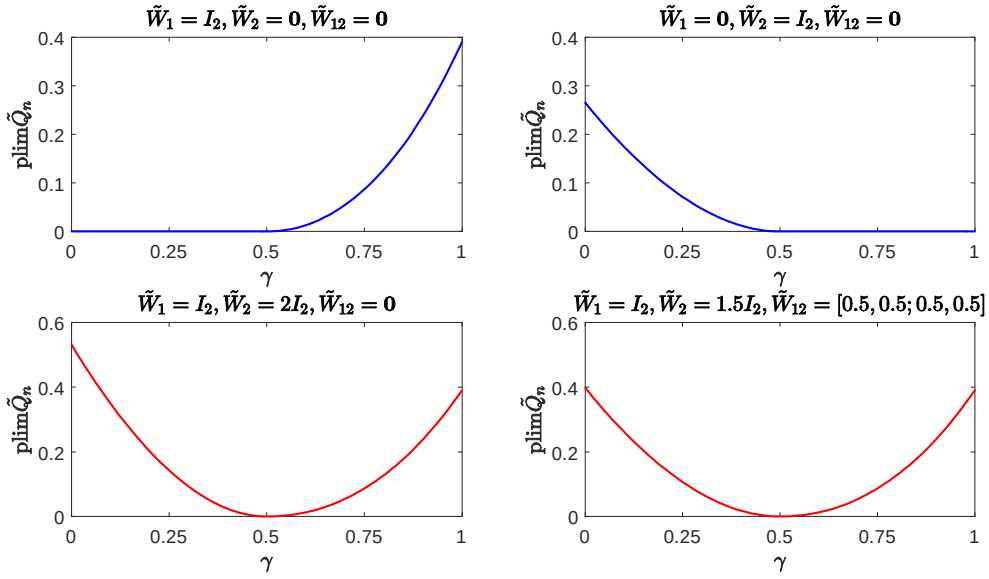

Figure 1: plim$\widetilde{Q}_n(\gamma)$ when $E[\varepsilon|q] = 0$ and $\mathbf{z} = (1, q)'$

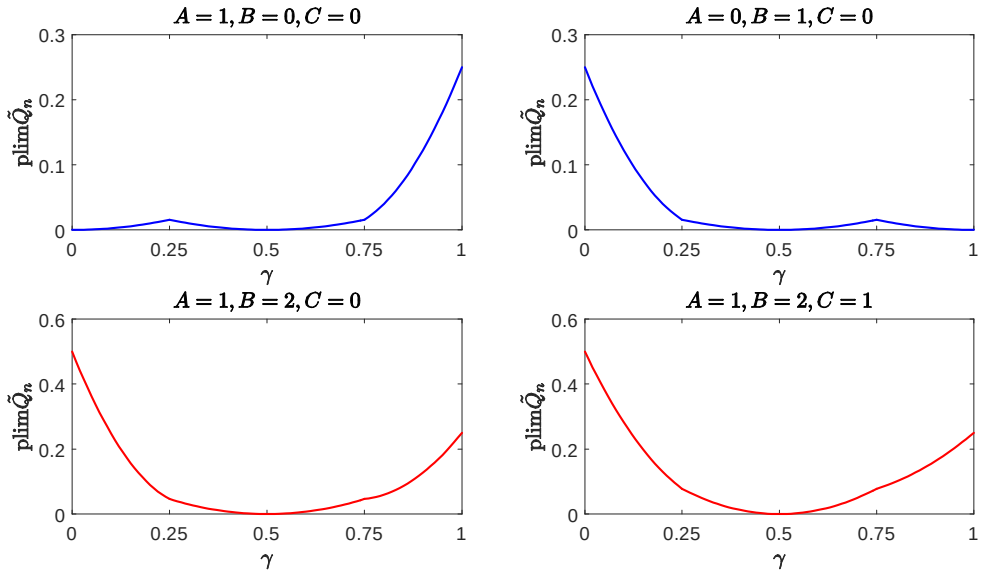

Figure 2: plim$\widetilde{Q}_n(\gamma)$ when $E[\varepsilon|q] \neq 0$

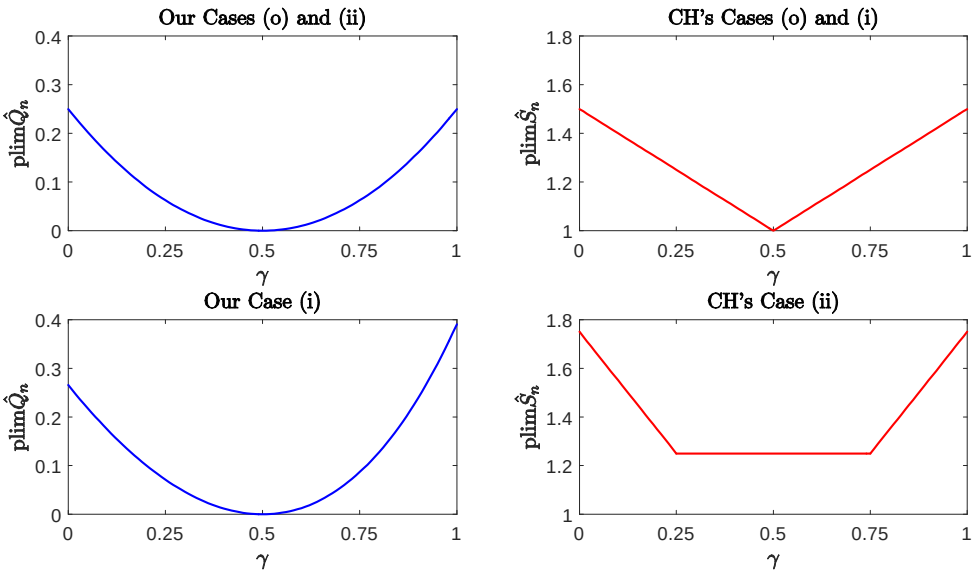

Figure 3: plim$\widehat{Q}_n(\gamma)$ and plim$\widehat{S}_n(\gamma)$

design of the DGP in this simple example. Specifically, using the formulae in (17) and (18), it turns out that $\frac{\partial \mathbb{E}[m_{1i}(\beta_0,\delta_0,\gamma)]}{\partial \gamma}\Big|_{\gamma=\gamma_0-} = -1/2 = \frac{\partial \mathbb{E}[m_{1i}(\beta_0,\delta_0,\gamma)]}{\partial \gamma}\Big|_{\gamma=\gamma_0+}$, and $\frac{\partial \mathbb{E}[m_{2i}(\beta_0,\delta_0,\gamma)]}{\partial \gamma}\Big|_{\gamma=\gamma_0-} = -1/2 = \frac{\partial \mathbb{E}[m_{2i}(\beta_0,\delta_0,\gamma)]}{\partial \gamma}\Big|_{\gamma=\gamma_0+}$. But, in general, $E[m_i(\beta_0,\delta_0,\gamma)]$ is not differentiable at γ_0 in case (ii). On the other hand, for our 2SLS, $\mathbb{E}[g_i(\beta_0,\delta_0,\gamma)] = \frac{1}{2} - \gamma$, which is always differentiable at γ_0.

3.4 Summary of Identification Results

Before summarizing the identification results for the existing estimators of γ_0, we provide a further comment on the distinction between "sum of squares" and "square of sums" criteria. Note that "sum of squares" criteria need not have more identification power than "square of sums" criteria. When q is endogenous, the example in Section 2.1 of Yu (2013a) shows that the 2SLS estimator of CH is not consistent, and the example in the previous subsection shows that the limit objective function of CH's 2SLS need not even have a unique minimizer. On the contrary, either the 2SLS estimator of this paper or the GMM estimator of HHB can generate a consistent estimator of γ_0.[22]

Table 1 summarizes the identification results for all possibly consistent estimators of γ_0 in various scenarios. The first four estimators require instruments and the last two do not. Among the last two, Perron and Yamamoto (2015) (PY in Table 1) use the LSE to estimate γ in a structural change model even when there is endogeneity. However, as shown in Yu (2015a), this strategy is valid only in the structural change context. From Table 1, it seems that the IDKE of YP has the most extensive identification power even though the method makes no use of instruments. Nevertheless, the IDKE cannot identify γ_0 in CTR models,[23] while HHB's GMM estimator and our 2SLS estimator can identify γ_0 even in such models (although inference needs further investigation). Table 1 also lists KST's STR estimator. As mentioned in the Introduction,

[22] Of course, we can claim that CH's 2SLS cannot be applied when q is endogenous; see YLP for modifications of CH's 2SLS to generate consistent estimators of γ_0.

[23] In such models, the IDKE can be extended also to take into account slope differences at each $\gamma \in \Gamma$ beyond level differences to identify γ_0.

when q is exogenous, their estimator is equivalent to CH's 2SLS estimator so it is consistent; but when q is endogenous, their estimator is not generally consistent unless the endogeneity is relatively small compared to the threshold effect. Taking Table 1 as a whole, we can see some interesting differences between structural change models and TR models – specifically case (o) vs. cases (i) and (ii). It is, however, becoming folklore in the literature that these two kinds of models are considered similar to each other (at least in terms of their asymptotic properties).[24] The present findings reveal that such folklore is misleading when there is endogeneity.

	(o): $\mathbb{E}[\varepsilon\|\mathbf{z},q]=0$ and $q \perp (\mathbf{z}',\mathbf{x}')'$ [25]		(i): $\mathbb{E}[\varepsilon\|\mathbf{z},q]=0$ but $q \not\perp (\mathbf{z}',\mathbf{x}')'$		(ii): $\mathbb{E}[\varepsilon\|\mathbf{z},q]\neq 0$ but $\mathbb{E}[\mathbf{z}\varepsilon_{\leq\gamma_0}]=\mathbb{E}[\mathbf{z}\varepsilon_{q>\gamma_0}]=\mathbf{0}$	
	Consistency	Literature	Consistency	Literature	Consistency	Literature
GMM of HHB	No	HHB	Yes[26]	This Paper	Yes[27]	This Paper
2SLS of This Paper	No	This Paper	Yes	This Paper	Yes	This Paper
2SLS of CH	Yes	HHB[28]	Yes	CH	No	Yu (2013a)
STR of KST	Yes	HHB	Yes	CH	No	YLP
LSE of PY and Yu (2015a)[29]	Yes	PY and Yu (2015a)[30]	No	Yu (2015a)	No	Yu (2015a)
IDKE of YP	Yes	YP	Yes	YP	Yes	YP

Table 1: Identification of γ_0 by Various (Possibly Valid) Estimators in Different Scenarios

4 Inference Based on the IDKE with $k_{\pm}(0)=0$

This section presents limit theory for the IDKE $\widehat{\gamma}$ in Method II where $k_{\pm}(0)=0$. To facilitate formulation of the limit distribution of $\widehat{\gamma}$, we define the following quantities,

$$\begin{aligned}\Delta_i &= \mathbb{E}[y_i|x_i,q_i=\gamma_0-]-\mathbb{E}[y_i|x_i,q_i=\gamma_0+] =: m_-(x_i)-m_+(x_i),\\ \Delta_f(x_i) &= \Delta_i\cdot f(x_i,\gamma_0),\end{aligned}$$

where $\Delta_i=\Delta(x_i,\gamma_0)$ and $\Delta_f(x_i)$ is the limit of $\widehat{\Delta}_i(\gamma_0)$ with $\widehat{\Delta}_i(\gamma)$ and $\Delta(x,\gamma_0)$ defined in (5) and (6) respectively. To derive the asymptotic distribution of $\widehat{\gamma}$ we use the following assumptions on $f(u|x,q)$, which is allowed to be discontinuous at $q=\gamma_0$.

Assumption U:

(a) *$f(u|x,q)$ is continuous in u for $(x',q)'\in X\times\Gamma_\epsilon^-$ and $(x',q)'\in X\times\Gamma_\epsilon^+$, where $\Gamma_\epsilon^-=(\underline{\gamma}-\epsilon,\gamma_0]$ and $\Gamma_\epsilon^+=(\gamma_0,\overline{\gamma}+\epsilon)$ for some $\epsilon>0$.*

[24] In structural change models, case (i) corresponds to the circumstance that the moments $E[(\mathbf{z}_t',\mathbf{x}_t')]$ are not equal for all t, i.e., that there is some nonstationarity in the mean of $E[(\mathbf{z}_t',\mathbf{x}_t')]$; case (ii) corresponds to $E[\varepsilon_t|\mathbf{z}_t]\neq 0$ for all t but with $\frac{1}{T_0}\sum_{t=1}^{T_0}\mathbb{E}[\mathbf{z}_t\varepsilon_t]=\frac{1}{T-T_0}\sum_{t=1}^{T-T_0}\mathbb{E}[\mathbf{z}_t\varepsilon_t]=\mathbf{0}$, where T_0 is the break point, i.e, $\mathbf{z}_t$ is not a valid instrument for all t but is valid when the information in each regime is integrated. These results echo the finding in YP that nonstationarity is often helpful in establishing identification.

[25] Here, we implicitly assume $\mathbf{z}$ and $\mathbf{x}$ do not include q.

[26] Both $m_1(\theta)$ and $m_2(\theta)$ are required to prove consistency.

[27] Either $m_1(\theta)$ or $m_2(\theta)$ is enough to prove consistency.

[28] Yu (2015a) strengthens this result a little. Specifically, let $\mathbf{z}=(\underline{\mathbf{z}}',q)'$ and $\mathbf{x}=(\underline{\mathbf{x}},q)'$. If $q\perp\underline{\mathbf{z}}$ and $E[\underline{\mathbf{x}}|\underline{\mathbf{z}},q]=g(\underline{\mathbf{z}})+q\lambda$, i.e., q need not be independent of $\underline{\mathbf{x}}$, then projecting $\underline{\mathbf{x}}$ only on $\underline{\mathbf{z}}$ in the first stage would generate a consistent estimator of γ_0.

[29] $\mathbf{z}$ is not necessary here.

[30] Yu (2015a)'s result is a little stronger. Specifically, if $q\perp\underline{\mathbf{x}}$ and $E[\varepsilon|\underline{\mathbf{x}},q]=g(\underline{\mathbf{x}})+q\lambda$, i.e., q need not be exogenous, then the LSE is consistent.

(b) $f(u|x,q)$ *is Lipschitz in* $(x',q)'$ *for* $(x',q)' \in X \times \Gamma_\epsilon^-$ *and* $(x',q)' \in \mathcal{X} \times \Gamma_\epsilon^+$.
(c) $\mathbb{E}[u^4|x,q]$ *is uniformly bounded on* $(x',q)' \in X \times \Gamma_\epsilon$, *where* $\Gamma_\epsilon = \Gamma_\epsilon^- \cup \Gamma_\epsilon^+$.

Given Assumption U, we impose the following conditions on the bandwidth h.

Assumption H: $h \to 0$, *and* $\sqrt{n}h^d/\ln n \to \infty$.

Observe that $nh^d = \sqrt{n}\ln n \frac{\sqrt{n}h^d}{\ln n} \to \infty$ when $\sqrt{n}h^d/\ln n \to \infty$. The limit theory for $\widehat{\gamma}$ is given in the next result.

Theorem 3 *Under Assumptions F, G, H, I, K, S and U,*

$$\sqrt{n/h}(\widehat{\gamma} - \gamma_0) \xrightarrow{d} N(0, \Sigma)$$

where

$$\Sigma = \frac{\mathbb{E}[\Delta_f^2(x_i) f^2(x_i)(\sigma_+^2(x_i) + \sigma_-^2(x_i))|q_i = \gamma_0]\xi_{(1)}}{f_q(\gamma_0)\left(\mathbb{E}[\Delta_f(x_i)\Delta_i f(x_i)|q_i = \gamma_0]\right)^2 k_+'(0)^2}$$

with $\xi_{(1)} = \int_0^1 k_+'(t)^2 dt$ *and* $\sigma_\pm^2(x) = \mathbb{E}\left[u^2|x, q = \gamma_0 \pm\right]$.

This result shows that $\widehat{\gamma}$ converges to γ_0 at the rate $\sqrt{n/h}$, a much faster rate than that of the DKE $\widetilde{\gamma}$ of DH because $\widehat{\gamma}$ utilizes more data information in estimation. Specifically, the convergence rate of DKE is $\sqrt{nh^{d-2}}$ and the relative rate $\sqrt{nh^{d-2}}/\sqrt{n/h} = \sqrt{h^{d-1}} \to 0$. Based on Theorem 2 of DH, the asymptotic variance of their estimator $\widetilde{\gamma}$ is

$$\Sigma_o = \frac{(\sigma_+^2(x_o) + \sigma_-^2(x_o))\kappa^2 \xi_{(1)}}{f(x_o, \gamma_0)\Delta_o^2 k_+'(0)^2}, \tag{21}$$

where $\kappa^2 = \int K(u_x)^2 du_x$ with $K(u_x) = \prod_{l=1}^{d-1} k(u_{x_l})$, and $\Delta_o = m_-(x_o) - m_+(x_o)$ which is equal to $(1, x_o', \gamma_0)\,\delta_0$ when $\mathbb{E}[\varepsilon|x,q]$ is continuous. This asymptotic variance is comparable to Σ, but critically relies on the choice of x_o. If $\Delta_i = \Delta$ and $\sigma_\pm^2(x) = \sigma^2$, then $\Sigma = O\left(\frac{\sigma^2}{f_q(\gamma_0)\Delta^2 k_+'(0)^2}\right)$. As expected, Σ is decreasing in $f_q(\gamma_0)$, $|\Delta|$ and $k_+'(0)$ and increasing in σ^2.

The convergence rate $\sqrt{n/h}$ of $\widehat{\gamma}$ exceeds the usual parametric rate $\sqrt{n}$. To understand this increase over the parametric rate, some heuristic analysis is helpful. For this purpose, we use the simple case where $d = 1$, so that q is the only covariate. The convergence rate is then determined by the balance between an empirical process and a deterministic centering process. Recall that

$$\widehat{\gamma} = \arg\max_{\gamma \in \Gamma} \widehat{Q}_n(\gamma) = \arg\max_{\gamma \in \Gamma} \left\{\widehat{Q}_n(\gamma) - \widehat{Q}_n(\gamma_0)\right\}.$$

Because $\widehat{\gamma}$ maximizes $\widehat{Q}_n(\gamma) - \widehat{Q}_n(\gamma_0)$ on Γ and $\gamma_0 \in \Gamma$, we have the decomposition

$$0 \le \widehat{Q}_n(\widehat{\gamma}) - \widehat{Q}_n(\gamma_0) = [Q_0(\widehat{\gamma}) - Q_0(\gamma_0)] + \left[\left(\widehat{Q}_n(\widehat{\gamma}) - Q_0(\widehat{\gamma})\right) - \left(\widehat{Q}_n(\gamma_0) - Q_0(\gamma_0)\right)\right],$$

where the first term on the extreme right side is the limit centering process and is less than zero because $\gamma_0 = \arg\max_{\gamma \in \Gamma} Q_0(\gamma)$, whereas the second term is the modulus of continuity of the empirical process, which exceeds zero. Hence, $Q_0(\widehat{\gamma}) - Q_0(\gamma_0)$ and

$$\sup_{|\gamma - \gamma_0| \le \delta} \left[\left(\widehat{Q}_n(\widehat{\gamma}) - Q_0(\widehat{\gamma})\right) - \left(\widehat{Q}_n(\gamma_0) - Q_0(\gamma_0)\right)\right] =: \frac{\phi_n(\delta)}{\sqrt{n}}$$

must balance out so their sum is greater than zero.

In the h neighborhood of γ_0, we can treat the model as a parametric one, so without loss of generality, assume

$$y_i = \Delta 1(q_i \leq \gamma_0) + u_i,$$

where $q_i = i/n$, $i = 1, \cdots, n$,[31] $u_i \sim \mathcal{N}(0,1)$ and $\Delta > 0$. Now, $\widehat{\gamma}$ tries to maximize $\widehat{\Delta}(\gamma) - \widehat{\Delta}(\gamma_0)$, where

$$\widehat{\Delta}(\gamma) = \frac{1}{nh} \sum_{i=n(\gamma-h)}^{n\gamma} k_- \left(\frac{i - n\gamma}{nh} \right) y_i - \frac{1}{nh} \sum_{i=n\gamma+1}^{n(\gamma+h)} k_+ \left(\frac{i - n\gamma}{nh} \right) y_i. \tag{22}$$

For $|\gamma - \gamma_0| \leq \delta$, and $\gamma < \gamma_0$, we have

$$\begin{aligned}
&\widehat{\Delta}(\gamma) - \widehat{\Delta}(\gamma_0) \\
&= \left[\tfrac{1}{nh} \sum_{i=n(\gamma-h)}^{n\gamma} k_- \left(\tfrac{i-n\gamma}{nh} \right) y_i - \tfrac{1}{nh} \sum_{i=n\gamma+1}^{n(\gamma+h)} k_+ \left(\tfrac{i-n\gamma}{nh} \right) y_i \right] - \left[\tfrac{1}{nh} \sum_{i=n(\gamma_0-h)}^{n\gamma_0} k_- \left(\tfrac{i-n\gamma_0}{nh} \right) y_i - \tfrac{1}{nh} \sum_{i=n\gamma_0+1}^{n(\gamma_0+h)} k_+ \left(\tfrac{i-n\gamma_0}{nh} \right) y_i \right] \\
&= \left[\Delta + \tfrac{1}{nh} \sum_{i=n(\gamma-h)}^{n\gamma} k_- \left(\tfrac{i-n\gamma}{nh} \right) u_i - \tfrac{\Delta}{nh} \sum_{i=n\gamma+1}^{n\gamma_0} k_+ \left(\tfrac{i-n\gamma}{nh} \right) - \tfrac{1}{nh} \sum_{i=n\gamma+1}^{n(\gamma+h)} k_+ \left(\tfrac{i-n\gamma}{nh} \right) u_i \right] \\
&\quad - \left[\Delta + \tfrac{1}{nh} \sum_{i=n(\gamma_0-h)}^{n\gamma_0} k_- \left(\tfrac{i-n\gamma_0}{nh} \right) u_i - \tfrac{1}{nh} \sum_{i=n\gamma_0+1}^{n(\gamma_0+h)} k_+ \left(\tfrac{i-n\gamma_0}{nh} \right) u_i \right] \\
&\approx -\Delta \left(\tfrac{1}{nh} \sum_{i=n\gamma+1}^{n\gamma_0} k_+ \left(\tfrac{i-n\gamma}{nh} \right) \right) + \tfrac{1}{nh} \sum_{i=n(\gamma-h)}^{n\gamma} \left[k_- \left(\tfrac{i-n\gamma}{nh} \right) - k_- \left(\tfrac{i-n\gamma_0}{nh} \right) \right] u_i + \tfrac{1}{nh} \sum_{i=n\gamma_0+1}^{n(\gamma_0+h)} \left[k_+ \left(\tfrac{i-n\gamma_0}{nh} \right) - k_+ \left(\tfrac{i-n\gamma}{nh} \right) \right] u_i \\
&\quad - \tfrac{1}{nh} \sum_{i=n\gamma+1}^{n\gamma_0} \left[k_+ \left(\tfrac{i-n\gamma}{nh} \right) + k_- \left(\tfrac{i-n\gamma}{nh} \right) \right] u_i \\
&= -O\left(\Delta \textstyle\int_0^{\frac{\gamma_0-\gamma}{h}} k_+(v) dv \right) + O_p \left(k'_-(0) \sqrt{\tfrac{1}{nh} \textstyle\int_{-1}^0 \left(\tfrac{\gamma_0-\gamma}{h} \right)^2 dv} + k'_+(0) \sqrt{\tfrac{1}{nh} \textstyle\int_0^1 \left(\tfrac{\gamma-\gamma_0}{h} \right)^2 dv} \right) \\
&\quad - O_p \left(\sqrt{\tfrac{1}{nh} \textstyle\int_0^{\frac{\gamma_0-\gamma}{h}} \left(k_+(v) + k_- \left(\tfrac{\gamma_0-\gamma}{h} - v \right) \right)^2 dv} \right).
\end{aligned}$$

If Δ is fixed, $k_+(0) = 0$ and $k'_+(0) > 0$, then $\int_0^{\frac{\gamma_0-\gamma}{h}} k_+(v) dv = O\left(\left(\frac{\gamma_0-\gamma}{h} \right)^2 \right)$, $\int_{-1}^0 \left(\frac{\gamma_0-\gamma}{h} \right)^2 dv = \int_0^1 \left(\frac{\gamma-\gamma_0}{h} \right)^2 dv = \left(\frac{\gamma_0-\gamma}{h} \right)^2$ and $\int_0^{\frac{\gamma_0-\gamma}{h}} \left(k_+(v) + k_- \left(\frac{\gamma_0-\gamma}{h} - v \right) \right)^2 dv = O\left(\left(\frac{\gamma_0-\gamma}{h} \right)^3 \right)$. As a result, $Q_0(\gamma) - Q_0(\gamma_0) = O\left(\delta^2/h^2 \right)$ and $\phi_n(\delta) = \sqrt{\delta^2/h^3}$ since $\left(\frac{\gamma_0-\gamma}{h} \right)^3 = o\left(\left(\frac{\gamma_0-\gamma}{h} \right)^2 \right)$. Suppose $\widehat{\gamma} - \gamma_0 = O_p\left(r_n^{-1} \right)$; solving $\frac{1}{r_n^2 h^2} \approx \frac{\sqrt{1/r_n^2}}{\sqrt{nh^3}}$, we get $r_n = \sqrt{n/h}$.

For comparison, consider the IDKE in YP and in Method III. In the former, $\int_0^{\frac{\gamma_0-\gamma}{h}} k_+(v) dv = \frac{\gamma_0-\gamma}{h}$, $\int_{-1}^0 \left(\frac{\gamma_0-\gamma}{h} \right)^2 dv = \int_0^1 \left(\frac{\gamma-\gamma_0}{h} \right)^2 dv = \left(\frac{\gamma_0-\gamma}{h} \right)^2$, and $\int_0^{\frac{\gamma_0-\gamma}{h}} \left(k_+(v) + k_- \left(\frac{\gamma_0-\gamma}{h} - v \right) \right)^2 dv = \frac{\gamma_0-\gamma}{h}$ since $k_\pm(0) > 0$. As a result, $Q_0(\gamma) - Q_0(\gamma_0) = O(\delta/h)$ and $\phi_n(\delta) = \sqrt{\delta/h^2}$ since $\left(\frac{\gamma_0-\gamma}{h} \right)^2 = o\left(\frac{\gamma_0-\gamma}{h} \right)$. Solving $\frac{1}{r_n h} \approx \frac{\sqrt{1/r_n}}{\sqrt{nh^2}}$, we get $r_n = n$. In the latter (as will be detailed in the next section), $\Delta \to 0$, so $Q_0(\gamma) - Q_0(\gamma_0) = O(\Delta\delta/h)$ and $\phi_n(\delta) = \sqrt{\delta/h^2}$. Solving $\frac{\Delta}{r_n h} \approx \frac{\sqrt{1/r_n}}{\sqrt{n/h^2}}$, we get $r_n = n\Delta^2$.

Figure 4 illustrates these heuristics. For example, in Method II, because $Q_0(\gamma) - Q_0(\gamma_0)$ is quadratic (in the neighborhood of γ_0) as in the regular parameter case, we expect $\widehat{\gamma}$ to have an asymptotic normal distribution. The extra h in the convergence rate $\sqrt{n/h}$ arises because the variations in $Q_0(\gamma) - Q_0(\gamma_0)$ and $\phi_n(\delta)$ are both in the scale of h in this nonparametric setup. In YP, $Q_0(\gamma) - Q_0(\gamma_0)$ is a nonsmooth function of γ (in the neighborhood of γ_0) such that γ can be more easily identified than in Method II, so

[31] Locally, q_i follows a uniform distribution on $[\gamma_0 - h, \gamma_0 + h]$, so we assume it follows the discrete form of $U[0,1]$ here.

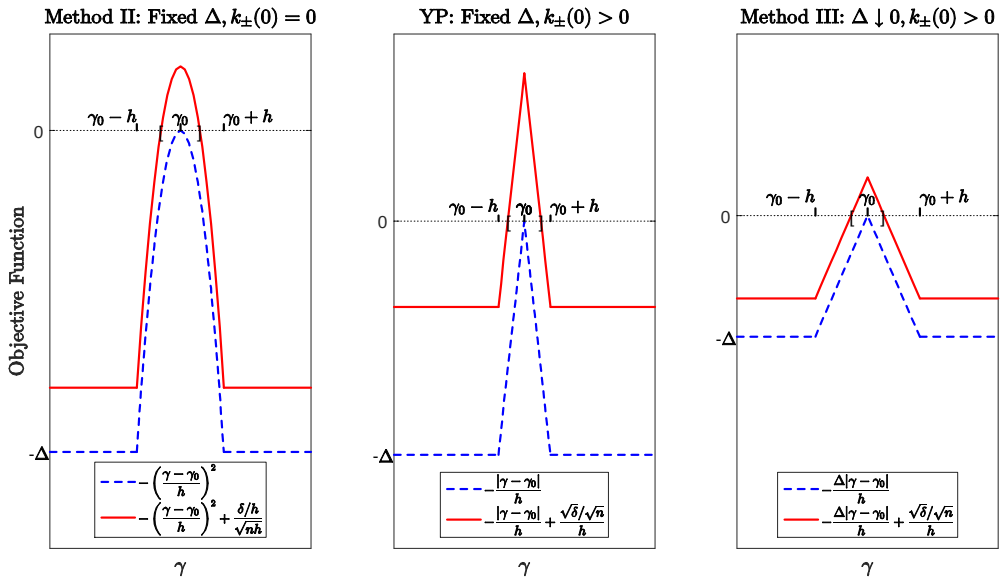

Figure 4: Balancing $Q_0(\gamma) - Q_0(\gamma_0)$ and $\frac{\phi_n(\delta)}{\sqrt{n}}$ in Method II, YP and Method III

the convergence rate n is faster than $\sqrt{n/h}$. In Method III, $Q_0(\gamma) - Q_0(\gamma_0)$ is still nonsmooth but the nonsmoothness is less severe (the left and right derivatives of $Q_0(\gamma) - Q_0(\gamma_0)$ at γ_0 are $O(\Delta/h)$, less than $O(1/h)$ order in YP), so the convergence rate is slower than that in YP.[32] In YP and Method III, the convergence rates of $\widehat{\gamma}$ are actually the same as in the parametric cases.

For inference of γ based on inverting the t statistic, we need to estimate Σ in Theorem 3. A straightforward approach is to use the sample analog. Specifically, we can estimate Σ by

$$\widehat{\Sigma} = \frac{\frac{1}{n}\sum_{i=1}^{n} k_h(q_i - \widehat{\gamma})\widehat{\Delta}_i^2(\widehat{\gamma})\widehat{f}^2(x_i)2\widehat{u}_i^2\xi_{(1)}}{\left(\frac{1}{n}\sum_{i=1}^{n} k_h(q_i - \widehat{\gamma})\widehat{\Delta}_i^2(\widehat{\gamma})\widehat{f}^{-1}(x_i,\widehat{\gamma})\widehat{f}(x_i)\right)^2 k_+'(0)^2},$$

where

$$\begin{aligned}
\widehat{\Delta}_i^2(\gamma) &= \left(\widehat{m}_-(x_i,\gamma)\widehat{f}_-(x_i,\gamma) - \widehat{m}_+(x_i,\gamma)\widehat{f}_+(x_i,\gamma)\right), \\
\widehat{f}(x_i) &= \frac{1}{n-1}\sum_{j=1,j\neq i}^{n} K_{h,ij}^x, \quad \widehat{f}(x_i,\gamma) = \frac{1}{n-1}\sum_{j=1,j\neq i}^{n} K_{h,ij}^x k_h(q_j - \gamma), \\
\widehat{u}_i(\gamma) &= y_i - \widehat{m}_-(x_i,\gamma)1(q_i \leq \gamma) - \widehat{m}_+(x_i,\gamma)1(q_i > \gamma), \widehat{u}_i = \widehat{u}_i(\widehat{\gamma})
\end{aligned}$$

[32] To understand the convergence rates of $\widehat{\gamma}$ in Methods I and II, we can compare these rates with those in Seo and Linton (2007) where a parametric TR model is considered. In the SLSE of Seo and Linton (2007), $1(q_i > \gamma)$ in the objective function of the LSE is changed to $K\left(\frac{q_i-\gamma}{h}\right)$ with $K(\cdot)$ being a cdf, which results in a convergence rate of $\sqrt{n/h}$, which is exactly the same as in Method II; when h is fixed, the convergence rate reduces to $\sqrt{n}$, the same as in Method I. On the other hand, although the convergence rate of Method II and that of Seo and Linton are the same, the asymptotic distributions are still different.

with

$$\begin{aligned}\widehat{m}_{\pm}(x_i,\gamma) &= \frac{\frac{1}{n-1}\sum_{j=1,j\neq i}^{n} K_{h,ij}^{x} k_h^{\pm}(q_j-\gamma)y_j}{\frac{1}{n-1}\sum_{j=1,j\neq i}^{n} K_{h,ij}^{x} k_h^{\pm}(q_j-\gamma)}, \\ \widehat{f}_{\pm}(x_i,\gamma) &= \frac{1}{n-1}\sum_{j=1,j\neq i}^{n} K_{h,ij}^{x} k_h^{\pm}(q_j-\gamma).\end{aligned}$$

The next result establishes that $\widehat{\Sigma}$ is consistent.

Theorem 4 *Under the assumptions of Theorem 3,* $\widehat{\Sigma} \xrightarrow{p} \Sigma$.

Another method of inference is based on inverting the LR statistic. Although this method has been proposed in the small-threshold-effect framework by Hansen (2000), it seems new in the current setting. Our LR statistic can be used to test whether $\gamma = \gamma_0$ and is defined as

$$LR_n^{(1)}(\gamma) = nh\frac{k_+'(0)}{\xi_{(1)}}\frac{\mathbb{E}[\Delta_f(x_i)\Delta_i f(x_i)|q_i=\gamma_0]}{\mathbb{E}[\Delta_f^2(x_i)f^2(x_i)(\sigma_+^2(x_i)+\sigma_-^2(x_i))|q_i=\gamma_0]}\left(\widehat{Q}_n(\widehat{\gamma})-\widehat{Q}_n(\gamma)\right).$$

Corollary 3 *Under the assumptions of Theorem 3,*

$$LR_n^{(1)}(\gamma_0) \xrightarrow{d} \chi_1^2.$$

To construct a CI for γ based on $LR_n^{(1)}$ we need to estimate $\frac{\mathbb{E}[\Delta_f(x_i)\Delta_i f(x_i)|q_i=\gamma_0]|}{\mathbb{E}[\Delta_f^2(x_i)f^2(x_i)(\sigma_+^2(x_i)+\sigma_-^2(x_i))|q_i=\gamma_0]}$. The natural estimator is $\frac{\frac{1}{n}\sum_{i=1}^{n} k_h(q_i-\widehat{\gamma})\widehat{\Delta}_i^2(\widehat{\gamma})\widehat{f}^{-1}(x_i,\widehat{\gamma})\widehat{f}(x_i)}{\frac{1}{n}\sum_{i=1}^{n} k_h(q_i-\widehat{\gamma})\widehat{\Delta}_i^2(\widehat{\gamma})\widehat{f}^2(x_i)2\widehat{u}_i^2}$, which is consistent from Theorem 4. Hence, the $(1-\alpha)100\%$ LR-CI for γ is

$$\left\{\gamma : \widehat{LR}_n^{(1)}(\gamma) \leq \mathrm{cv}_\alpha\right\},$$

where $\widehat{LR}_n^{(1)}(\gamma)$ replaces $\frac{\mathbb{E}[\Delta_f(x_i)\Delta_i f(x_i)|q_i=\gamma_0]|}{\mathbb{E}[\Delta_f^2(x_i)f^2(x_i)(\sigma_+^2(x_i)+\sigma_-^2(x_i))|q_i=\gamma_0]}$ in $LR_n^{(1)}(\gamma)$ by its estimate, and cv_α is the $(1-\alpha)100\%$ quantile of χ_1^2.

5 Inference Based on the IDKE with Shrinking Threshold Effects

In the previous section, the IDKE was adjusted by letting $k_\pm(0) = 0$ to construct a CI for γ and the threshold effect was taken as fixed. In this section, the IDKE is adjusted from a different perspective by allowing for the threshold effect to shrink to zero with the sample size but requiring that $k_\pm(0) > 0$.

5.1 Optimal Rate of Convergence for γ

First, we discuss the interpretation of a shrinking threshold effect. As argued in Section 2.4 of YP, the local shifter $1(q > \gamma)$ plays the role of an instrument. When q shifts from the left side of γ to its right side, the shift in the mean of y shrinks to zero. This behavior can be interpreted as the manifestation of a weak IV problem in the threshold regression context. A natural question that then arises is the identifiability of γ as δ shrinks to zero. To put this question a different way, we can ask what is the minimum magnitude of δ that ensures identification of γ. For this purpose, we cast the model in the following framework.

Suppose $\mathcal{P}$ is a family of probability models on some fixed measurable space $(\Omega, \mathcal{A})$. Let γ be a functional defined on $\mathcal{P}$. Given an estimator $\widehat{\gamma}$ of γ and a loss function $L(\widehat{\gamma}, \gamma)$, the maximum expected loss over $P \in \mathcal{P}$

is defined to be

$$R\left(\widehat{\gamma},\mathcal{P}\right)=\sup_{P\in\mathcal{P}}\mathbb{E}_P\left[L\left(\widehat{\gamma},\gamma(P)\right)\right],$$

where $\mathbb{E}_P$ is the expectation operator under the probability measure P. A popular loss function (e.g., Stone (1980)) is the 0-1 loss

$$L\left(\widehat{\gamma},\gamma\right)=1\left\{\left|\widehat{\gamma}-\gamma\right|>\frac{\epsilon}{2}\right\}$$

for some fixed $\epsilon>0$, which will be used in this paper. Under this loss, $R\left(\widehat{\gamma},\mathcal{P}\right)$ is the maximum probability that $\widehat{\gamma}$ is not in the $\epsilon/2$ neighborhood of γ. The goal is to find an achievable lower bound for the minimax risk defined by

$$\inf_{\widehat{\gamma}}R\left(\widehat{\gamma},\mathcal{P}\right)=\inf_{\widehat{\gamma}}\sup_{P\in\mathcal{P}}\mathbb{E}_P\left[L\left(\widehat{\gamma},\gamma(P)\right)\right]. \tag{23}$$

Only if δ is large enough, will the right side converge to zero. The best rate of convergence of $R\left(\widehat{\gamma},\mathcal{P}\right)$ to zero is then called the *optimal rate of convergence* or the *minimax rate of convergence*. Now $P\in\mathcal{P}$ in our model is characterized by $m_{\pm}(x,q)$ and γ as follows:

$$\begin{aligned}\mathcal{P}(s,B) \quad = \quad & \left\{P_{m_{\pm},\gamma}:\frac{dP_{m_{\pm},\gamma}}{d\mu}=f(x,q)\varphi_{x,q}\left(y-m_-(x,q)1(q\leq\gamma)-m_+(x,q)1(q>\gamma)\right),\right.\\ & \left.m_-(x,q)\in\mathcal{C}_s\left(B,\mathcal{X}\times\Gamma_\epsilon^-\right),m_+(x,q)\in\mathcal{C}_s\left(B,\mathcal{X}\times\Gamma_\epsilon^+\right),\int u\varphi_{x,q}(u)du=0,\gamma\in\Gamma\right\},\end{aligned}$$

where μ is Lebesgue measure on $\mathbb{R}^{\overline{d}}$, $\varphi_{x,q}(u)$ is the conditional density of u given $(x',q)'$, and $\mathcal{C}_s\left(\cdot,\cdot\right)$ is defined in Section 2.2.

To formulate a precise statement of our next result, let

$$\rho_n=\sqrt{\int_{\mathcal{X}}\left(m_-(x,\gamma)-m_+(x,\gamma)\right)^2f(x|\gamma)dx},$$

where $\gamma=\gamma(P)$, and $f(x|\gamma)$ can be replaced by any weight function $w(x)$ with $0<c\leq w(x)\leq C<\infty$ and $\int_{\mathcal{X}}w(x)=1$. If $\mathbb{E}[\varepsilon|x,q]$ is continuous, then $\rho_n=\left|\mathbb{E}\left[\mathbf{x}|q=\gamma\right]'\delta_n\right|=O\left(\|\delta_n\|\right)$, similar to the $\|\delta_n\|$ in Section 3.

Theorem 5 *Suppose Assumptions F, S and U hold, and $P\in\mathcal{P}(s,B)$ with $s\geq1$. If $n^{\frac{s}{2s+1}}\rho_n\to\infty$, then*

$$\underline{\lim}_{n\to\infty}\inf_{\widehat{\gamma}}\sup_{P\in\mathcal{P}(s,B)}P\left(n\rho_n^2\left|\widehat{\gamma}-\gamma(P)\right|>\frac{\epsilon}{2}\right)\geq C,$$

and if $n^{\frac{s}{2s+1}}\rho_n=O\left(1\right)$, then

$$\underline{\lim}_{n\to\infty}\inf_{\widehat{\gamma}}\sup_{P\in\mathcal{P}(s,B)}P\left(\left|\widehat{\gamma}-\gamma(P)\right|>\frac{\epsilon}{2}\right)\geq C,$$

for some positive constant C and small $\epsilon>0$.

We begin our discussion of this result by clarifying a key difference between the parametric and nonparametric threshold models with shrinking threshold effects. In the former, as long as the jump size is $n^{-\alpha}$ with $0<\alpha<1/2$ (i.e., larger than $n^{-1/2}$), γ can be identified; in the latter, however, we require a jump size larger than $n^{-\frac{s}{2s+1}}$ to identify γ. In other words, the minimum rate of convergence for γ in the nonparametric model must be larger than $n^{\frac{1}{2s+1}}$ rather than *any* rate diverging to infinity as in the parametric model. In the parametric model, $s=\infty$, so $n^{-\frac{s}{2s+1}}=n^{-1/2}$ and $n^{-\frac{1}{2s+1}}=0$, i.e., the parametric result is a limiting

special case of Theorem 5 as $s \to \infty$. Such a difference between the parametric model and the nonparametric model does not seem to have been explicitly recognized in the literature. For example, Müller and Song (1997) show that the convergence rate of the DKE is $n\rho_n^2$ when q is the only regressor by implicitly assuming a trade off in rates under which ρ_n is taken to be larger than $n^{-\frac{s}{2s+1}}$. In fact, when γ can be identified, the optimal rate of convergence for γ is the same as in the parametric case. This rate is achieved by the IDKE, as shown in the next section.

5.2 Asymptotics for $\widehat{\gamma}$

To facilitate finding an expression for the limit distribution of $\widehat{\gamma}$, we define the following quantities

$$\begin{aligned} D_n &= \mathbb{E}[\Delta_f(x_i)\Delta_i f(x_i)|q_i = \gamma_0]/\rho_n^2, \\ V_{1n} &= \mathbb{E}[\Delta_f^2(x_i) f^2(x_i)\sigma_-^2(x_i)|q_i = \gamma_0]/\rho_n^2, \\ V_{2n} &= \mathbb{E}[\Delta_f^2(x_i) f^2(x_i)\sigma_+^2(x_i)|q_i = \gamma_0]/\rho_n^2, \end{aligned}$$

where ρ_n is evaluated at γ_0. We also impose the following conditions on the bandwidth h.

Assumption H′: $h \to 0$, $\sqrt{n}h^d/\ln n \to \infty$, $\rho_n \to 0$, $\rho_n/h^s \to \infty$, $nh\rho_n^2 \to \infty$.

If we employ the optimal bandwidth $h = O\left(n^{-\frac{1}{2s+1}}\right)$, then $\sqrt{n}h^d/\ln n = n^{\frac{2(s-d)+1}{2(2s+1)}}/\ln n \to \infty$ under Assumption G′ ($s \geq d$). Also, $\rho_n/h^s \to \infty$ and $nh\rho_n^2 \to \infty$ hold when $n^{\frac{s}{2s+1}}\rho_n \to \infty$. The limit distribution of $\widehat{\gamma}$ is given in the following theorem.

Theorem 6 *Under Assumptions F, G′, H′, I, K′, S and U, if $D_n \to D$ and $V_{\ell n} \to V_\ell$ as $n \to \infty$, then*

$$n\rho_n^2(\widehat{\gamma} - \gamma_0) \xrightarrow{d} \omega \cdot \Lambda(\lambda),$$

where $\omega = \frac{1}{f_q(\gamma_0)}\frac{V_1}{D^2}$, and

$$\Lambda(\lambda) = \underset{r}{\operatorname{argmax}} \begin{cases} W_1(-r) - \frac{|r|}{2}, & \text{if } r \leq 0, \\ \sqrt{\lambda}W_2(r) - \frac{|r|}{2}, & \text{if } r > 0, \end{cases}$$

with $\lambda = V_2/V_1$, and $W_\ell(r)$, $\ell = 1, 2$, being two independent standard Wiener processes on $[0, \infty)$.

In some special cases, the asymptotic distribution of $\widehat{\gamma}$ can be simplified. For example, if $\sigma_-^2(x_i) = \sigma_{10}^2$ and $\sigma_+^2(x_i) = \sigma_{20}^2$, so that the model is locally homoskedastic within each regime, then $\lambda = \sigma_{20}^2/\sigma_{10}^2$; if $\sigma_-^2(x_i) = \sigma_+^2(x_i) = \sigma_0^2$, so that the model is homoskedastic locally around γ_0, then $\lambda = 1$. To compare with the asymptotic distribution of $\widehat{\gamma}$ in Method II, note that by Slutsky's theorem,

$$n\rho_n^2 \frac{f_q(\gamma_0)D_n^2}{V_{1n}}(\widehat{\gamma} - \gamma_0) \xrightarrow{d} \Lambda(\lambda)$$

in Method III, whereas

$$\sqrt{\frac{n\rho_n^2}{h}\frac{f_q(\gamma_0)D_n^2}{V_{1n}(1+\lambda)}\frac{k_+'(0)^2}{\xi_{(1)}}}(\widehat{\gamma} - \gamma_0) \xrightarrow{d} \mathcal{N}(0,1) \tag{24}$$

in Method II, so we use a different normalization on $(\widehat{\gamma} - \gamma_0)$ to achieve a nondegenerate limit distribution. We can show that when $\rho_n \to 0$, the result in (24) still holds and the CI based on $LR_n^{(1)}(\gamma)$ remains valid. The convergence rate of $\widehat{\gamma}$ in Method II is $\sqrt{n\rho_n^2/h}$. Since $\sqrt{n\rho_n^2/h}/n\rho_n^2 = \sqrt{1/(nh\rho_n^2)} \to 0$, the $\widehat{\gamma}$

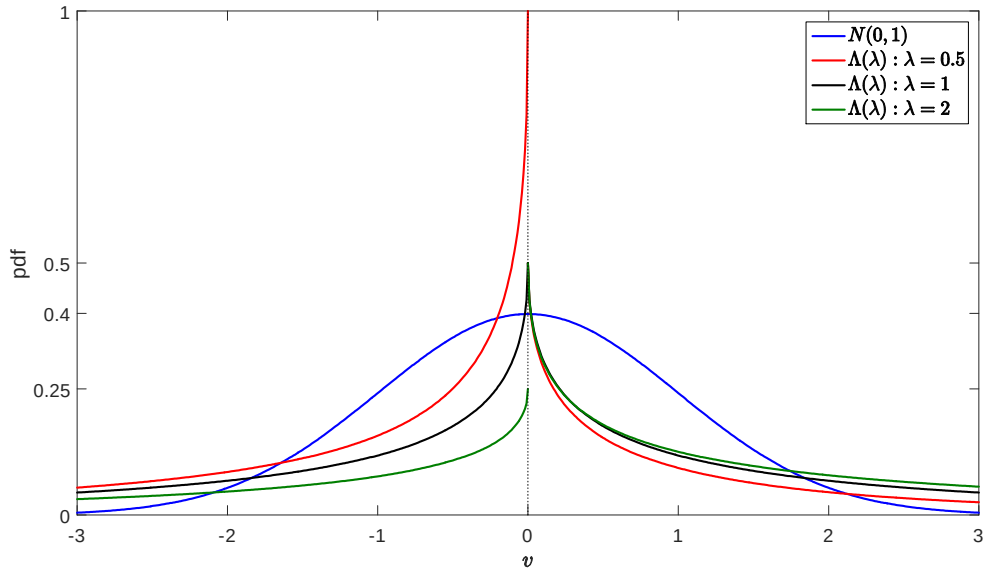

Figure 5: Comparison Between the PDFs of $\mathcal{N}(0,1)$ and $\Lambda(\lambda)$: $\lambda = 0.5, 1, 2$

estimator in Method II has a slower convergence rate. But its convergence rate is still faster than that of the 2SLS estimator in Section 3 because $\sqrt{n\rho_n^2/h}/\sqrt{n\left\|\delta_n\right\|^2} = O\left(\sqrt{1/h}\right) \to \infty$. Figure 5 shows the difference between the $\mathcal{N}(0,1)$ and $\Lambda(\lambda)$ limit densities, where the analytic form of the density of $\Lambda(\lambda)$ is reported in Appendix B of Bai (1997), viz.,

$$p(x) = \begin{cases} -\frac{1}{2}\Phi\left(-\frac{\sqrt{|x|}}{2}\right) + \frac{1}{2}\left(1+\frac{2}{\lambda}\right)\exp\left(\frac{1}{2}\frac{1}{\lambda}\left(1+\frac{1}{\lambda}\right)|x|\right)\Phi\left(-\frac{\left(1+\frac{2}{\lambda}\right)\sqrt{|x|}}{2}\right), & \text{if } x<0, \\ -\frac{1}{2\lambda}\Phi\left(-\frac{1}{2}\sqrt{\frac{x}{\lambda}}\right) + \left(1+\frac{1}{2\lambda}\right)\exp\left(\frac{1+\lambda}{2}x\right)\Phi\left(-\left(\sqrt{\lambda}+\frac{1}{2\sqrt{\lambda}}\right)\sqrt{x}\right), & \text{if } x>0, \end{cases}$$

with $\Phi(\cdot)$ being the cdf of $\mathcal{N}(0,1)$. Intuitively, when the heteroskedasticity measure $\lambda > 1$, it is more likely that $\Lambda(\lambda)$ achieves the maximum at $r > 0$. This intuition explains why the right tail of $\Lambda(\lambda)$ is heavier than the left tail. Interestingly, the effects of λ on the limit distributions of the two γ estimators are different: its effect on the estimator $\widehat{\gamma}$ in Method II is to increase variance (but maintain symmetry), whereas its effect on the estimator $\widehat{\gamma}$ in Method III is to introduce skewness.

For comparison, we state the limit distribution of the DKE in the following corollary. For this purpose, we adjust Assumption H′ as follows.

Assumption H″: $h \to 0$, $\sqrt{n}h^d/\ln n \to \infty$, $\Delta_o \to 0$, $\Delta_o/h^s \to \infty$, $nh^d\Delta_o^2 \to \infty$, *where* Δ_o *is defined in (21) and is equal to* $(1, x_o', \gamma_0)\,\delta_n$ *when* $E[\varepsilon|x,q]$ *is continuous.*

The optimal bandwidth $h = O\left(n^{-\frac{1}{2s+d}}\right)$ satisfies Assumption H″.

Corollary 4 *Under Assumptions F, G′, H″, I, K′, S and U,*

$$nh^{d-1}\Delta_o^2(\widetilde{\gamma} - \gamma_0) \xrightarrow{d} \omega_o \cdot \Lambda(\lambda_o, \kappa),$$

where $\omega_o = \frac{\sigma_-^2(x_o)}{f(x_o,\gamma_0)}$, *and*

$$\Lambda(\lambda_o,\kappa) = \underset{r}{\text{argmax}} \begin{cases} W_1(-r) - \frac{|r|}{2\kappa}, & \text{if } r \le 0, \\ \sqrt{\lambda_o} W_2(r) - \frac{|r|}{2\kappa}, & \text{if } r > 0, \end{cases}$$

with $\lambda_o = \frac{\sigma_+^2(x_o)}{\sigma_-^2(x_o)}$, κ^2 *being defined in (21), and standard Brownian motions* $W_\ell(r)$, $\ell = 1, 2$, *as in Theorem 6.*

The distribution of $\Lambda(\lambda_o,\kappa)$ is derived in Proposition 1 of Stryhn (1996). Since it will not be used for inference, it is omitted here. Compared with the convergence rate of $\widehat{\gamma}$ (viz., $n\rho_n^2$), the convergence rate of $\widetilde{\gamma}$ (viz., $nh^{d-1}\Delta_o^2$) is much slower especially when d is large. But it is still faster than the convergence rate of the DKE in Method II because the ratio $nh^{d-1}\Delta_o^2/\sqrt{nh^{d-2}\Delta_o^2} = \sqrt{nh^d\Delta_o^2} \to \infty$. To compare with the limit distribution of $\widetilde{\gamma}$ in Method II, note that by Slutsky's theorem,

$$nh^{d-1}\Delta_o^2 \frac{f(x_o,\gamma_0)}{\sigma_-^2(x_o)}(\widetilde{\gamma} - \gamma_0) \xrightarrow{d} \Lambda(\lambda_o,\kappa)$$

in Method III, whereas

$$\sqrt{\frac{nh^{d-1}\Delta_o^2}{h} \frac{f(x_o,\gamma_0)}{\sigma_-^2(x_o)(1+\lambda_o)} \frac{k_+'(0)^2}{\kappa^2 \xi_{(1)}}}(\widetilde{\gamma} - \gamma_0) \xrightarrow{d} \mathcal{N}(0,1)$$

in Method II. The limit distributions of $\widetilde{\gamma}$ in both methods involve only information local to x_o. Similar to λ in the limit distributions of $\widehat{\gamma}$, λ_o affects only the variance of $\widetilde{\gamma}$ in Method II, but affects symmetry in Method III. A new factor κ^2 also appears in the limit distributions of $\widetilde{\gamma}$; different from λ_o, the factor κ^2 increases variance but does not affect symmetry in either case.

It is also interesting to notice that the limit distribution of $\widehat{\gamma}$ in Method III does not depend on the kernel choice whereas the limit distribution of $\widetilde{\gamma}$ in Method III does depend on the kernel choice on x (although not on q). These results echo Theorem 1 and Corollary 1 of YP where ρ_n is fixed and $k_\pm(0) > 0$. But when $k_\pm(0) = 0$, from Theorem 3, the asymptotic distribution of $\widehat{\gamma}$ depends on the kernel choice on q, and from (21), the asymptotic distribution of $\widetilde{\gamma}$ depends on the kernel choice on both x and q. In other words, whether $k_\pm(0) = 0$ or not does indeed affect the role of the kernel on q with respect to data usage (and hence efficiency) of the estimators.

We next discuss inference concerning the threshold parameter γ based on our IDKE approach. Although we can construct a CI for γ by inverting the asymptotic distribution of $\widehat{\gamma}$ in Theorem 6, Hansen (2000) shows that such CIs perform poorly due to the identification failure when $\rho_n = 0$. He suggests constructing CIs for γ by inverting the LR statistic instead, which in our case is defined as

$$LR_n^{(2)}(\gamma) = nh \frac{1}{4k_+(0)} \frac{D_n}{V_{1n}} \left(\widehat{Q}_n(\widehat{\gamma}) - \widehat{Q}_n(\gamma) \right).$$

To do so, we make use of the following result.

Corollary 5 *Under the assumptions of Theorem 6,*

$$LR_n^{(2)}(\gamma_0) \xrightarrow{d} M(\lambda),$$

where $M(\lambda)$ *follows the distribution* $P(M(\lambda) \le z) = (1 - e^{-z})(1 - e^{-z/\lambda})$ *with* λ *defined in Theorem 6.*

To construct CIs for γ, we need to estimate D_n/V_{1n} and λ. By similar procedures to those of the last

section, we can show that

$$\begin{aligned}\frac{\widehat{D}_n}{\widehat{V}_{1n}} &= \frac{\frac{1}{n}\sum_{i=1}^n k_h(q_i-\widehat{\gamma})\widehat{\Delta}_i^2(\widehat{\gamma})\widehat{f}^{-1}(x_i,\widehat{\gamma})\widehat{f}(x_i)}{\frac{1}{n}\sum_{i=1}^n k_h^-(q_i-\widehat{\gamma})\widehat{\Delta}_i^2(\widehat{\gamma})\widehat{f}^2(x_i)\widehat{u}_i^2},\\ \widehat{\lambda} &= \frac{\frac{1}{n}\sum_{i=1}^n k_h^+(q_i-\widehat{\gamma})\widehat{\Delta}_i^2(\widehat{\gamma})\widehat{f}^2(x_i)\widehat{u}_i^2}{\frac{1}{n}\sum_{i=1}^n k_h^-(q_i-\widehat{\gamma})\widehat{\Delta}_i^2(\widehat{\gamma})\widehat{f}^2(x_i)\widehat{u}_i^2}\end{aligned}$$

are the required consistent estimators, where $\widehat{\Delta}_i(\widehat{\gamma})$, $\widehat{f}(x_i,\widehat{\gamma})$, $\widehat{f}(x_i)$ and $\widehat{u}_i$ are defined in the last section. If $\sigma_-^2(x_i)=\sigma_{10}^2$ and $\sigma_+^2(x_i)=\sigma_{20}^2$, then λ can be simply estimated by the ratio

$$\widehat{\lambda}=\frac{\frac{1}{n}\sum_{i=1}^n k_h^+(q_i-\widehat{\gamma})\widehat{u}_i^2}{\frac{1}{n}\sum_{i=1}^n k_h^-(q_i-\widehat{\gamma})\widehat{u}_i^2}.$$

Given all these components, the $(1-\alpha)$ LR-CI for γ is

$$\left\{\gamma : \widehat{LR}_n^{(2)}(\gamma) \le \widehat{\mathrm{cv}}_\alpha\right\},$$

where $\widehat{LR}_n^{(2)}(\gamma)$ replaces D_n/V_{1n} in $LR_n^{(2)}(\gamma)$ by its estimates, and $\widehat{\mathrm{cv}}_\alpha$ is the $(1-\alpha)$ quantile of M obtained by replacing λ by its estimate.

To compare the LR statistic $LR_n^{(2)}(\gamma)$ with $LR_n^{(1)}(\gamma)$ in the last section, note that $LR_n^{(1)}(\gamma)$ can be expressed as

$$LR_n^{(1)}(\gamma)=nh\frac{k_+'(0)}{\xi_{(1)}}\frac{D_n}{V_{1n}+V_{2n}}\left(\widehat{Q}_n(\widehat{\gamma})-\widehat{Q}_n(\gamma)\right).$$

If $\widehat{Q}_n(\widehat{\gamma})-\widehat{Q}_n(\gamma)$ are the same in these two LR statistics, then

$$\frac{LR_n^{(1)}(\gamma)}{LR_n^{(2)}(\gamma)}=4\frac{k_+(0)k_+'(0)}{\xi_{(1)}}\frac{V_{1n}}{V_{1n}+V_{2n}}=:R_n. \tag{25}$$

If the model is locally homoskedastic in each regime, then $R_n=4\frac{\sigma_{10}^2}{\sigma_{10}^2+\sigma_{20}^2}\frac{k_+(0)k_+'(0)}{\xi_{(1)}}$, which is further simplified to $2\frac{k_+(0)k_+'(0)}{\xi_{(1)}}$ when the model is locally homoskedastic in both regimes. However, $\widehat{Q}_n(\widehat{\gamma})-\widehat{Q}_n(\gamma)$ are not the same in $LR_n^{(1)}(\gamma)$ and $LR_n^{(2)}(\gamma)$ because the employed kernels are different and the $\widehat{\gamma}$'s are different. To compare the asymptotic distributions of these two LR statistics, we plot the asymptotic pdfs in Figure 6 and report their 95% critical values in Table 2.

Test Stat.	$LR_n^{(1)}$	$LR_n^{(2)}(\lambda=0.5)$	$LR_n^{(2)}(\lambda=1)$	$LR_n^{(2)}(\lambda=2)$
95% crit	3.841	3.040	3.676	6.081

Table 2: 95% Critical Values of $LR_n^{(1)}$ and $LR_n^{(2)}$ for $\lambda=0.5,1,2$

5.3 Comparison With the Parametric LSE

We close this section by comparing the IDKE with the LSE in the parametric case (see, e.g., Hansen (2000)). From YP, the LSE $\widehat{\gamma}_{LSE}$ obtained by minimizing

$$\min_{\beta,\delta}(Y-X\beta-X_{\le\gamma}\delta)'(Y-X\beta-X_{\le\gamma}\delta)=Y'Y-Y'P_\gamma Y \tag{26}$$

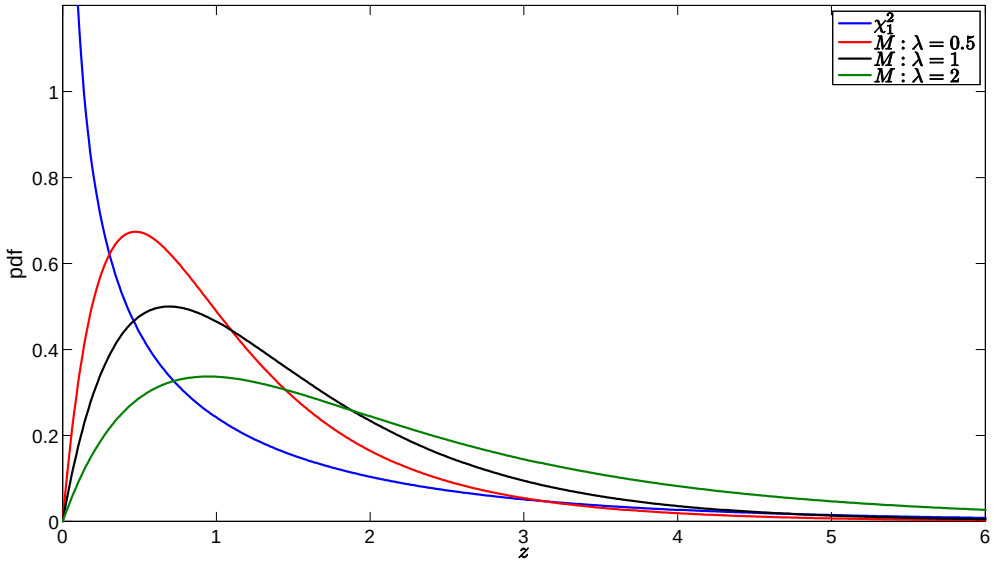

Figure 6: Comparison Between the PDFs of χ_1^2 and M for $\lambda = 0.5, 1, 2$

is equivalent to estimation by maximizing

$$\left(\widehat{\delta}' X'\right)\left[X\left(X'X\right)^{-1}X'_{>\gamma}X_{>\gamma}\left(X'X\right)^{-1}X'_{\leq\gamma}X_{\leq\gamma}\left(X'X\right)^{-1}X'\right]\left(X\widehat{\delta}\right), \tag{27}$$

where P_γ is the projection matrix onto $span\left(X, X_{\leq\gamma}\right)$, $\widehat{\delta}\left(\gamma\right)$ is the LSE of δ based on splitting according to the threshold γ, and X and $X_{\leq\gamma}$ are defined in (2) with a similar definition for $X_{>\gamma}$. Here, $X\widehat{\delta}\left(\gamma\right) = \left\{\mathbf{x}_i'\widehat{\delta}\left(\gamma\right)\right\}_{i=1}^n$, and $\mathbf{x}_i'\widehat{\delta}\left(\gamma\right)$ is an estimator of the conditional mean differential $\Delta\left(x_i,\gamma\right)$. Since $\widehat{\Delta}_i\left(\gamma\right)$ is estimating $\Delta\left(x_i,\gamma\right)f(x_i,\gamma)$, $X\widehat{\delta}\left(\gamma\right)$ is mimicing $\left\{\widehat{\Delta}_i\left(\gamma\right)/f\left(x_i,\gamma\right)\right\}_{i=1}^n$. In the parametric case, $f(x_i,\gamma_0)$ and $f(x_i)$ do not appear in D_n, V_{1n} and V_{2n}, so $\rho_n^2 D_n^2/V_{1n}$ reduces to $\frac{\mathbb{E}[\Delta_i^2|q_i=\gamma_0]^2}{\mathbb{E}[\Delta_i^2 u_i^2|q_i=\gamma_0-]}$ and λ reduces to $\frac{\mathbb{E}[\Delta_i^2 u_i^2|q_i=\gamma_0+]}{\mathbb{E}[\Delta_i^2 u_i^2|q_i=\gamma_0-]}$; if $\sigma_-^2(x_i) = \sigma_{10}^2$ and $\sigma_+^2(x_i) = \sigma_{20}^2$, then $\rho_n^2 D_n^2/V_{1n}$ further reduces to $\frac{\mathbb{E}[\Delta_i^2|q_i=\gamma_0]}{\sigma_{10}^2}$ and λ further reduces to $\frac{\sigma_{20}^2}{\sigma_{10}^2}$. A natural question is how to generate the same asymptotic distribution as in the parametric case when $\mathbb{E}[\varepsilon|x,q]$ is continuous. By careful inspection of the derivations in the proofs we can show that, if the objective function of the IDKE changes to

$$\frac{1}{n}\sum_{i=1}^n \frac{\widehat{f}\left(x_i,\gamma\right)}{\widehat{f}\left(x_i\right)}\left[\frac{\frac{1}{n-1}\sum_{j=1,j\neq i}^n y_j K_{h,ij}^{\gamma-}}{\frac{1}{n-1}\sum_{j=1,j\neq i}^n K_{h,ij}^{\gamma-}} - \frac{\frac{1}{n-1}\sum_{j=1,j\neq i}^n y_j K_{h,ij}^{\gamma+}}{\frac{1}{n-1}\sum_{j=1,j\neq i}^n K_{h,ij}^{\gamma+}}\right]^2,$$

then the asymptotic distribution of the IDKE is the same as that of the parametric LSE, where $\widehat{f}\left(x_i,\gamma\right)$ and $\widehat{f}\left(x_i\right)$ are consistent estimators of $f\left(x_i,\gamma\right)$ and $f\left(x_i\right)$, respectively. Asymptotically, we impose a weight $f_{q|x}(\gamma_0|x_i)$ on Δ_i^2. This weight is intuitive in the sense that when there are more data points in the neighborhood of $q=\gamma_0$ at x_i, we impose a larger weight on Δ_i^2. In fact, we can also show that the IDKE using this

objective function has the same asymptotic distribution as the LSE even in the framework of YP.[33] In other words, the complex-looking weights in the square bracket of the objective function (27) are asymptotically equivalent to $\left\{f_{q|x}(\gamma_0|x_i)\right\}_{i=1}^n$. Such an equivalence result is not at all obvious from the original least squares objective function (26).

6 Two Specification Tests

In this section, we study limit theory of the two specification tests in Section 2.2. We first specify some regularity conditions which are modifications of Assumptions F, G and U given earlier.

Assumption F′: $f(x,q) \in \mathcal{C}_1(B, \mathcal{X} \times \mathcal{Q})$.
Assumption F″: $f(x,q) \in \mathcal{C}_\lambda(B, \mathcal{X} \times \Gamma_\epsilon)$ *with* $\lambda \geq 1$, *and* $0 < \underline{f} \leq f(x,q) \leq \overline{f} < \infty$ *for* $(x',q)' \in \mathcal{X} \times \Gamma_\epsilon$.
Assumption G″: (i) $g(x,q) \in \mathcal{C}_s(B, \mathcal{X} \times \mathcal{Q})$ *with* $s \geq 2$; *(ii)* $g(x,q) \in C_s(B, \mathcal{X} \times \Gamma_\epsilon)$ *with* $s \geq 2$;
Assumption U′:

(a) $f(u|x,q)$ *is continuous in* u *for* $(x',q)' \in X \times Q^-$ *and* $(x',q)' \in X \times Q^+$, *where* $Q^- = [\underline{q}, \gamma_0]$ *and* $\mathcal{Q}^+ = (\gamma_0, \overline{q}]$.

(b) $f(u|x,q)$ *is Lipschitz in* $(x',q)'$ *for* $(x',q)' \in X \times Q^-$ *and* $(x',q)' \in \mathcal{X} \times \mathcal{Q}^+$.

(c) $\mathbb{E}[u^4|x,q]$ *is uniformly bounded on* $(x',q)' \in \mathcal{X} \times \mathcal{Q}$.

The following Assumptions B1 and B2 are made on the bandwidths used in the first and second tests.

Assumption B1: $nh^d \to \infty$, $h \to 0$.
Assumption B2: $nh^d \to \infty$, $b \to 0$, $h/b \to 0$, $nh^{d/2}b^{2\eta} \to 0$, *where* $\eta = \min(\lambda + 1, s)$.

Given $d > 1$, $h/b \to 0$ implies $h^{d/2}/b \to 0$, so $nh^d \to \infty$ implies that $nh^{d/2}b \to \infty$, where $nh^{d/2}b$ is the magnitude of $I_n^{(2)}$ under $H_1^{(2)}$. The quantity $nh^{d/2}b^{2\eta}$ is the bias of $I_n^{(2)}$ under $H_0^{(2)}$, so the assumption $nh^{d/2}b^{2\eta} \to 0$ guarantees that $I_n^{(2)}$ is centered at the origin. Under $H_0^{(1)}$, the bias of $I_n^{(1)}$ is $h^{d/2}$, so $h \to 0$ ensures that $I_n^{(1)}$ is also centered at the origin. The condition $h/b \to 0$ requires that h is smaller than b, which helps to generate power under $H_1^{(2)}$ and shrink the bias under $H_0^{(2)}$ to zero. Intuitively, if $h/b \to 0$, then the term $K_{h,ij}$ in $I_n^{(2)}$ makes the product $\widehat{e}_i\widehat{e}_j$ behave like a squared term, producing an effect that generates power. In the first test, $m(x,q)$ under $H_0^{(1)}$ is parametric, so the corresponding bandwidth of b is a constant so that $h \to 0$ necessarily implies $h/b \to 0$. In testing $H_0^{(2)}$ versus $H_1^{(2)}$, our test statistic $I_n^{(2)}$ still applies when $\mathbb{E}[\varepsilon|x,q]$ is not smooth at $q = \gamma_0$. But in this case the null hypothesis is better modified to the equivalence $m_-(x) = m_+(x)$ for all $x \in \mathcal{X}$ and g in Assumption G″ does not need to be smooth at $q = \gamma_0$. Also, we need to add the requirement $nh^{d/2}b^3 \to 0$ to Assumption B2, where $nh^{d/2}b^3$ is the bias of $I_n^{(2)}$ attributed to the cusp of $m(x,q)$ at $q = \gamma_0$.

In the second test, we impose the following assumption on the kernel $l_b(\cdot,t)$.

Assumption L: $l_b(\cdot,t)$ *takes the form of (3) with order* $p = s + \lambda - 1$.

Thus, $l_b(\cdot,t)$ may be a higher order kernel to reduce the bias in $\widehat{y}_i$.

[33] Using such an objective function, the asymptotic distribution of the IDKE with $k_\pm(0) = 0$ in Section 4 would also change. For example, Σ would change to $\frac{\left(E[\Delta_i^2 u_i^2|q_i=\gamma_0^-]+E[\Delta_i^2 u_i^2|q_i=\gamma_0^+]\right)\xi_{(1)}}{f_q(\gamma_0)\left(E[\Delta_i^2|q_i=\gamma_0]\right)^2 k'_+(0)^2}$, and $LR_n^{(1)}(\gamma)$ changes to $nh\frac{k'_+(0)}{\xi_{(1)}}\frac{E[\Delta_i^2|q_i=\gamma_0]}{E[\Delta_i^2 u_i^2|q_i=\gamma_0^-]+E[\Delta_i^2 u_i^2|q_i=\gamma_0^+]}\left(\widehat{Q}_n(\widehat{\gamma}) - \widehat{Q}_n(\gamma)\right)$.

6.1 Limit Theory for the Two Tests

The following two theorems give the limit distribution of $I_n^{(\ell)}$ under the null $H_0^{(\ell)}$ and local power under $H_1^{(\ell)}$. Note that the main component of $I_n^{(\ell)}$ under $H_0^{(\ell)}$ is a degenerate U-statistic, so the asymptotic distribution is normal rather than a functional of a chi-square process, as in the usual structural change literature.

Theorem 7 *Under Assumptions B1, F′, G″(i), K, S, and U′, the following hold:*

(i)

$$I_n^{(1)} \xrightarrow{d} \mathcal{N}\left(0, \Sigma^{(1)}\right)$$

uniformly over $\mathcal{H}_0^{(1)}$*, where*

$$\Sigma^{(1)} = 2\int k^{2d}(u)du\mathbb{E}\left[f\left(x,q\right)\sigma^4\left(x,q\right)\right], \text{ with } \sigma^2\left(x,q\right) = \mathbb{E}[u^2|x,q],$$

which can be consistently estimated by

$$v_n^{(1)2} = \frac{2h^d}{n(n-1)}\sum_i\sum_{j\neq i} K_{h,ij}^2 \widehat{e}_i^2\widehat{e}_j^2.$$

Hence, a test based on the studentized test statistic $T_n^{(1)} = I_n^{(1)}/v_n^{(1)}$

$$t_n^{(1)} = 1\left(T_n^{(1)} > z_\alpha\right),$$

has significance level α*, where* z_α *is the* $1-\alpha$ *quantile of* $N(0,1)$.[34]

(ii) *If under* $H_1^{(1)}$*,* $m(x,q) - \overline{m}(x,q) = n^{-1/2}h^{-d/4}\Delta_n(x,q)$ *such that* $\int \Delta_n(x,q)^2 f(x,q)^2 dxdq \to \Delta$*, then*

$$I_n^{(1)} \xrightarrow{d} N\left(\Delta, \Sigma^{(1)}\right) \text{ and } T_n^{(1)} \xrightarrow{d} \mathcal{N}\left(\Delta/\sqrt{\Sigma^{(1)}}, 1\right),$$

so that the test $t_n^{(1)}$ *is consistent and* $P_m\left(T_n^{(1)} > z_\alpha\right) \to 1$ *for any* $m(\cdot)$ *such that* $\int \left(m(x,q) - \overline{m}(x,q)\right)^2 f(x,q)^2 dxdq \neq 0$*. Furthermore, the result continues to hold when* z_α *is replaced by any nonstochastic constant* $C_n = o\left(nh^{d/2}\right)$.

According to this result, $I_n^{(1)}$ has only trivial power if $\mathbb{E}\left[\left(m(x,q) - \overline{m}(x,q)\right)^2 f(x,q)\right] = 0$. Consider the following special example to illustrate. Suppose $m(x,q)$ under $H_0^{(1)}$ is $\mathbf{x}'\beta + \mathbf{x}'\delta 1\left(q \leq \gamma\right)$, and the alternative is $m(x,q) = \mathbf{x}'\beta + \mathbf{x}'\delta 1\left(q \leq \gamma\right) + \mathbf{x}'\xi + \mathbf{x}'\zeta 1\left(q \leq \gamma\right)$, then obviously, $\mathbb{E}\left[\left(m(x,q) - \overline{m}(x,q)\right)^2 f(x,q)\right] = 0$ under $H_1^{(1)}$ and $I_n^{(1)}$ has no discriminatory power against such $m(x,q)$. This point was observed for classical specification testing without threshold effects – see, e.g., Bierens and Ploberger (1997, p. 1135). Possible cases that do generate non-trivial power include models where (i) $m(x,q)$ takes the same parametric form but has a different threshold point from $\overline{m}(x,q)$, and (ii) $m(x,q)$ takes a nonparametric form for which $\int \left(m(x,q) - \overline{m}(x,q)\right)^2 f(x,q)^2 dxdq > 0$.

Theorem 8 *Under Assumptions B2, F″, G″(ii), K, L, S, and U, the following hold:*

(i)

$$I_n^{(2)} \xrightarrow{d} \mathcal{N}\left(0, \Sigma^{(2)}\right)$$

[34] The test is one-sided because $I_n^{(1)}$ is based on the L^2-distance between $m(\cdot)$ and $\overline{m}(\cdot)$.

uniformly over $\mathcal{H}_0^{(2)}$*, where*

$$\Sigma^{(2)} = 2\int k^{2d}(u)du\mathbb{E}\left[1_q^\Gamma f\left(x,q\right)\sigma^4\left(x,q\right)\right],$$

which can be consistently estimated by

$$v_n^{(2)2} = \frac{2h^d}{n(n-1)}\sum_i\sum_{j\neq i} 1_i^\Gamma 1_j^\Gamma K_{h,ij}^2\widehat{e}_i^2\widehat{e}_j^2.$$

As a result, the test based on the studentized test statistic $T_n^{(2)} = I_n^{(2)}/v_n^{(2)}$

$$t_n^{(2)} = 1\left(T_n^{(2)} > z_\alpha\right),$$

has significance level α*, where* z_α *is the* $1-\alpha$ *quantile of* $\mathcal{N}(0,1)$*.*

(ii) *If under* $H_1^{(2)}$*,* $m_-(x) - m_+(x) = n^{-1/2}h^{-d/4}b^{-1/2}\Delta_n(x)$ *such that* $\int \Delta_n(x)^2 f(x,\gamma_0)^2 dx \to \Delta$*, then*

$$I_n^{(2)} \xrightarrow{d} N\left(\zeta\Delta, \Sigma^{(2)}\right) \text{ and } T_n^{(2)} \xrightarrow{d} N\left(\zeta\Delta/\sqrt{\Sigma^{(2)}}, 1\right),$$

where $\zeta = 2\int_0^1\left(\int_v^1 l(u)du\right)^2 dv$*, and the test* $t_n^{(2)}$ *is consistent with* $P_m\left(T_n^{(2)} > z_\alpha\right) \to 1$ *for any* m *such that* $\int\left(m_-(x) - m_+(x)\right)^2 f(x,\gamma_0)^2 dx \neq 0$*. The result continues to hold when* z_α *is replaced by any nonstochastic constant* $C_n = o\left(nh^{d/2}b\right)$*.*

These two theorems show that $I_n^{(1)}$ and $I_n^{(2)}$ have power against different deviations of $m(x,q)$ from H_0. For $I_n^{(1)}$, power is generated from global deviations of $m(x,q)$ from H_0, just as in classical specification testing (see, e.g., Theorem 3 of Zheng (1996) and Theorem 3.1 of Fan and Li (2000)). For $I_n^{(2)}$, power is generated only from local deviations in the neighborhood of $q = \gamma_0$. In consequence, we need a larger deviation for $I_n^{(2)}$ than for $I_n^{(1)}$ to generate non-trivial power – specifically, $n^{-1/2}h^{-d/4}b^{-1/2}/n^{-1/2}h^{-d/4} = b^{-1/2} \to \infty$.

6.2 Bootstrapping Critical Values

As is evident from the proofs of Theorems 7 and 8, the convergence rates of $T_n^{(1)}$ and $T_n^{(2)}$ to the standard normal is slow. The bias under $H_0^{(1)}$ is $h^{d/2}$ and under $H_0^{(2)}$ is $nh^{d/2}b^{2\eta}$. Both these rates are low for some standard choices of bandwidth. As argued in the literature of classical specification testing (see, e.g., Härdle and Mammen (1993), Li and Wang (1998), Stute et al. (1998), Delgado and Manteiga (2001), and Gu et al. (2007)), an improved approximation of the finite-sample distribution of $T_n^{(\ell)}$ can be obtained using the wild bootstrap (Wu, 1986; Liu, 1988). We therefore suggest that the following algorithm WB be used in both tests, with $\widehat{e}_i$ and $\widehat{y}_i$ having different definitions in the two tests.

Algorithm WB:

Step 1: For $i = 1, \cdots, n$, generate the two-point wild bootstrap residual $u_i^* = \widehat{e}_i\left(1-\sqrt{5}\right)/2$ with probability $\left(1+\sqrt{5}\right)/\left(2\sqrt{5}\right)$, and $u_i^* = \widehat{e}_i\left(1+\sqrt{5}\right)/2$ with probability $\left(\sqrt{5}-1\right)/\left(2\sqrt{5}\right)$, then $\mathbb{E}^*\left[u_i^*\right] = 0$, $\mathbb{E}^*\left[u_i^{*2}\right] = \widehat{e}_i^2$ and $\mathbb{E}^*\left[u_i^{*3}\right] = \widehat{e}_i^3$, where $\mathbb{E}^*\left[\cdot\right] = \mathbb{E}\left[\cdot|\mathcal{F}_n\right]$ and $\mathcal{F}_n = \left\{(x_i', q_i, y_i)\right\}_{i=1}^n$.

Step 2: Generate the bootstrap resample $\{y_i^*, x_i, q_i\}_{i=1}^n$ by[35]

$$y_i^* = \widehat{y}_i + u_i^*.$$

Then obtain the bootstrap residuals $\widehat{e}_i^* = y_i^* - \widehat{y}_i^*$, where $\widehat{y}_i^*$ is defined similarly to $\widehat{y}_i$ except that y_i in the construction of $\widehat{y}_i$ is replaced by y_i^* .

Step 3: Use the bootstrap samples to compute the statistics

$$\begin{aligned} I_n^{(1)*} &= \frac{nh^{d/2}}{n(n-1)} \sum_i \sum_{j \neq i} K_{h,ij} \widehat{e}_i^* \widehat{e}_j^*, \\ I_n^{(2)*} &= \frac{nh^{d/2}}{n(n-1)} \sum_i \sum_{j \neq i} 1_i^\Gamma 1_j^\Gamma K_{h,ij} \widehat{e}_i^* \widehat{e}_j^*, \end{aligned}$$

and the estimated asymptotic variances

$$\begin{aligned} v_n^{(1)*2} &= \frac{2h^d}{n(n-1)} \sum_i \sum_{j \neq i} K_{h,ij}^2 \widehat{e}_i^{*2} \widehat{e}_j^{*2}, \\ v_n^{(2)*2} &= \frac{2h^d}{n(n-1)} \sum_i \sum_{j \neq i} 1_i^\Gamma 1_j^\Gamma K_{h,ij}^2 \widehat{e}_i^{*2} \widehat{e}_j^{*2}. \end{aligned}$$

The studentized bootstrap statistics are $T_n^{(\ell)*} = I_n^{(\ell)*}/v_n^{(\ell)*}$. Here, the same b and h are used as in $I_n^{(\ell)}$ and $v_n^{(\ell)2}$ in Theorems 7 and 8.[36]

Step 4: Repeat steps 1-3 B times, and use the empirical distribution of $\left\{T_{n,k}^{(\ell)*}\right\}_{k=1}^B$ to approximate the null distribution of $T_n^{(\ell)}$. We reject $H_0^{(\ell)}$ if $T_n^{(\ell)} > T_{n(\alpha B)}^{(\ell)*}$, where $T_{n(\alpha B)}^{(\ell)*}$ is the upper α-percentile of $\left\{T_{n,k}^{(\ell)*}\right\}_{k=1}^B$.

In Step 1, a popular way to simulate u_i^* in the second test is based on $\widehat{e}_i$'s centralized counterpart $\overline{\widehat{e}}_i = \widehat{e}_i - \overline{\widehat{e}}$ rather than $\widehat{e}_i$ itself, where $\overline{\widehat{e}} = \sum_{i=1}^n \widehat{e}_i 1_i^{\Gamma_b} \Big/ \sum_{i=1}^n 1_i^{\Gamma_b}$, $\Gamma_b = \left(\underline{\gamma} - b, \overline{\gamma} + b\right)$; see, e.g., Gijbels and Goderniaux (2004) and Su and Xiao (2008). Such a formulation can lead to $\sum_{i=1}^n \overline{\widehat{e}}_i 1_i^{\Gamma_b} \Big/ \sum_{i=1}^n 1_i^{\Gamma_b} = 0$,[37] which will not affect the asymptotic results but may affect the finite sample performance of Algorithm WB especially under $H_1^{(2)}$.

The bootstrap sample is generated by imposing the null hypothesis. Therefore, the bootstrap statistic $T_n^{(\ell)*}$ will mimic the null distribution of $T_n^{(\ell)}$ even when the null hypothesis is false. When the null is false, $\widehat{e}_i$ is not a consistent estimate of ε_i or u_i. Nevertheless, the following theorem shows that the above bootstrap procedure is valid. This is because our studentized test statistic $T_n^{(\ell)}$ is invariant to the variance of e. But the wild bootstrap procedure is not valid if the test statistic $I_n^{(\ell)}$ is used instead of $T_n^{(\ell)}$.[38]

[35] To construct $I_n^{(2)*}$, we need only the data with $q_i \in \left[\underline{\gamma} - b, \overline{\gamma} + b\right]$.

[36] If we use a data-adaptive bandwidth such as cross-validation based on each bootstrap sample, then the algorithm is extremely time-consuming. See Chapter 3 of Mammen (1992) for related discussions.

[37] In the first test, $\frac{1}{n}\sum_{i=1}^n \widehat{e}_i = \frac{1}{n}\sum_{i=1}^n \widehat{e}_i 1(q_i \leq \widehat{\gamma}) + \frac{1}{n}\sum_{i=1}^n \widehat{e}_i 1(q_i > \widehat{\gamma}) = 0$ since the covariates include a constant term.

[38] The wild bootstrap for $I_n^{(2)}$ should be valid because the misspecification in the variance of e happens only in a b neighborhood of γ_0.

Theorem 9 *Under the assumptions of Theorem 7 and 8,*

$$\sup_{z\in\mathbb{R}}\left|P\left(T_n^{(\ell)*}\leq z|\mathcal{F}_n\right)-\Phi(z)\right|=o_p\left(1\right),$$

where $\Phi(\cdot)$ *is the cdf of* $\mathcal{N}(0,1)$.

7 Simulations

We report simulations designed to assess the performance of our CIs and tests. We will concentrate on procedures whose performance is unclear in the literature. For inference, we will check only the CIs that invert the two LR statistics in Sections 4 and 5. It is unnecessary to compare the performance of the IDKE and the DKE because YP have already shown that the former performs much better than the latter in finite samples. Similarly, we do not check the performance of estimators and CIs based on our 2SLS or efficient GMM procedure because these are compared in YLP with other estimators and CIs that employ instruments. Further, we do not report the performance of CIs based on inverting the t statistics because in Method II its performance is similar to the LR-based CI and in Method III its performance is worse. For the two specification tests, we investigate only the two nonparametric tests developed in the main text because the performances of parametric tests developed in Supplement D are widely available in the literature.[39] Another reason for focusing on these two CI constructions and two specification tests is that neither involves instruments. As mentioned in the Introduction, good instruments are hard to find and justify in practice, so these methods have appeal in applied work.

We use a similar DGP as in YP for the simulation designs. Specifically, $y=\delta_1 1(q\leq\gamma)+\varepsilon$, i.e., the threshold effect does not depend on x, where $\gamma=0$ and $\Gamma=[-0.1,0.1]$, x and q are independent and each is uniformly distributed over $[-0.5,0.5]$, and $\varepsilon|\left(x,q\right)\sim N(-\delta_2 q^3,0.1^2)$. In CI construction, we let $\delta_1=0.1$ and 0.2, indicating small and large threshold effects, respectively, and $\delta_2=1$, indicating severe endogeneity. In testing endogeneity, we let $\delta_1=0.2$ and $\delta_2=0,0.2,0.5$ and 1, where $\delta_2=0$ corresponds to the null. In testing threshold effects, $\delta_1=0,0.1,0.2$ and 0.5, and $\delta_2=1$, where $\delta_1=0$ corresponds to the null. For the IDKE with $k_{\pm}(0)=0$,

$$k_{-}(x,r)=-x(1+x)1(-1\leq x\leq r)\Big/\left(\frac{1}{6}-\frac{1}{2}r^2-\frac{1}{3}r^3\right),0\leq r\leq 1,$$

and for the IDKE with shrinking threshold effects,

$$k_{-}(x,r)=\frac{3}{4}(1-x^2)1(-1\leq x\leq r)\Big/\left(\frac{1}{2}+\frac{3}{4}r-\frac{1}{4}r^3\right),0\leq r\leq 1, \tag{28}$$

which degenerates to the Epanechnikov kernel when $r=1$; $k_{+}(x,r)=k_{-}(-x,r)$.[40] In both tests, the kernel in $K_{h,ij}$ is specified in (28), and in the second test, $\widehat{y}_i$ is estimated by the local linear smoother which implies a second-order boundary kernel in $L_{b,ij}$ as required in Assumption L. Following DH, three bandwidths h are used based on the formula $Cn^{-1/2}$ with proportionality constants $C=2$, 3 and 4; in the second test, $b=\frac{1}{2}h^{1/2}$ to guarantee $h/b\to 0$ and $nh^{d/2}b^{2\eta}\to 0$ with $\eta=2$.[41] The simulation study in Müller (1991)

[39] Although the literature (e.g., Zheng, 1996; Li and Wang, 1998) provides simulation results when the approximation function $\widetilde{m}(x,q)$ in (7) is smooth, there are no corresponding results when $\widetilde{m}(x,q)$ is discontinuous. Also, although Porter and Yu (2015) investigate the finite sample performance of a similar structural change test as $I_n^{(2)}$, no covariates x are included in that work.

[40] These kernel functions imply $\xi_{(1)}=12$ and $k_{+}'(0)=6$ in Corollary 3 and $k_{+}(0)=1.5$ in Corollary 5. So $R_n=2\frac{k_{+}(0)k_{+}'(0)}{\xi_{(1)}}=1.5$ in our DGP.

[41] Notice that the range of bandwidths chosen is quite large, since the ratio of the proportionality constants between the

shows that a bandwidth without boundary adjustment works well, and we therefore use the same bandwidth for both interior and boundary points. $N = 500$ replications with sample size 500 and 1000 are used. In Algorithm WB, $B = 399$ when $n = 500$ and $B = 199$ when $n = 1000$. For CI construction the confidence level used is 95%, and for testing the level of significance is 5%.

7.1 Two LR-Based CIs

The coverage and average length of the two LR-based CIs are reported in Table 3. From Table 3, both methods perform well in coverage. Reductions in bandwidth and expansion of sample size both marginally improve coverage. On the other hand, different bandwidths and sample sizes have a big impact on CI lengths. Specifically, under our DGP, the medium bandwidth seems to perform satisfactorily for CI length among various scenarios and a larger sample size shrinks the length significantly. Another phenomenon deserving of mention is that CI length decreases sharply when the jump size doubles in both methods. This outcome is expected because larger threshold effects make the threshold point easier to identify. Comparing Method II to Method III, the latter behaves a little better in coverage. This may stem from the fact that the latter makes full use of the data information around γ_0 ($k_{\pm}(0) > 0$) while the former makes only marginal use of such information ($k_{\pm}(0) = 0$). This improvement comes at the cost that the CIs in Method III are generally longer than those in Method II.

7.2 Two Nonparametric Specification Tests

The size and power of the two nonparametric specification tests are reported in Tables 4 and 5, respectively. From these two tables, all tests have size close to the nominal 5% except in the second test with a large h where the test is undersized. A large h implies a large bias in $I_n^{(2)}$ so the rejection probability is adversely affected. The power of the endogeneity test is very good - even when $\delta_2 = 1$ and $n = 500$, the power is 100%. The power of the second test is also very good - even when $\delta_1 = 0.5$ and $n = 500$, the power is close to 100%. As a benchmark, $\delta_1 = 0.2$ corresponds to two standard deviations of the error term u which follows $\mathcal{N}\left(0, 0.1^2\right)$.

	$k_{\pm}(0) = 0$							
	Coverage				Length ($\times 10^{-2}$)			
n	500		1000		500		1000	
δ_1	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2
$C = 2$	0.998	0.962	0.998	0.964	17.6	7.11	15.23	4.27
$C = 3$	0.994	0.958	1	0.964	16.12	6.85	11.49	3.71
$C = 4$	0.984	0.954	0.992	0.970	15.98	7.52	10.62	4.07
	$\delta_n \to 0$							
	Coverage				Length ($\times 10^{-2}$)			
n	500		1000		500		1000	
δ_1	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2
$C = 2$	1	0.980	1	0.972	18.52	7.83	17.94	5.46
$C = 3$	1	0.986	1	0.984	17.23	5.11	14.59	2.47
$C = 4$	0.998	0.984	1	0.898	16.27	4.45	11.72	15.78

first and the last is 2. In the estimation of γ, the bandwidth is smaller than the usual optimal bandwidth (which is of rate $n^{-\frac{1}{2s+d}} = n^{-\frac{1}{6}}$), just as suggested in Porter and Yu (2015). In the second test, $N = n \times (2 \times \frac{1}{2} C^{1/2} n^{-1/4})^2 = C n^{1/2}$ data points are used to obtain $\widehat{y}_i$. When $C = 2$ and $n = 500$, $N \approx 45$. When $C = 4$ and $n = 1000$, $N \approx 126$.

Table 3: Comparison of Inferential Methods: Coverage and average length
for nominal 95% confidence for γ with $\delta_2 = 1$ and bandwidth proportionality constant C.

n	500				1000			
δ_2	0	0.2	0.5	1	0	0.2	0.5	1
$C = 2$	5.2	8	52.8	100	5	11.8	78.2	100
$C = 3$	5.8	10.2	68.8	100	3.6	15.8	94	100
$C = 4$	5.2	10.8	78.8	100	3.8	21	97.2	100

Table 4: Size and Power of $T_n^{(1)}$ (%): Nominal significance level $5\%, \delta_1 = 0.2$

n	500				1000			
δ_1	0	0.1	0.2	0.5	0	0.1	0.2	0.5
$C = 2$	4.4	17.4	78.2	99.4	3	29	93	100
$C = 3$	4.0	22.2	78.4	100	2.8	46.2	98.6	100
$C = 4$	3.8	17.2	68.6	99.4	1.8	49.8	98.8	100

Table 5: Size and Power of $T_n^{(2)}$ (%): Nominal significance level $5\%, \delta_2 = 1$

8 Conclusion

All three methods of estimation presented here for threshold point regression remain valid and invariant to endogeneity of the threshold variable. To the best of our knowledge these methods are the only ones in the literature offering such robustness. The first method is a nonlinear 2SLS method and requires instruments, while the other two methods are based on smoothing the objective function of the IDKE and do not require any instrumentation in their implementation. These are important advantages in empirical work where valid instruments are often scarce.

Our development and discussion of the 2SLS method clarifies some puzzles in the current literature about the properties of threshold regression estimation. We draw attention in particular to the following matters that are resolved in the paper: (i) why the usual GMM method cannot identify the threshold point in structural change models; (ii) why two groups of moments are required to identify the threshold point when the threshold variable is exogenous and correlated with the covariates and instruments, whereas only one group of moments is sufficient when the threshold variable is endogenous; and (iii) why the bootstrap is valid for our 2SLS method while it is generally invalid for the usual GMM approach.

In discussing the two IDKE-smoothing methods, we show that these IDKEs use different normalizations to obtain operable asymptotic distributions under different assumptions and we explain why their convergence rates are different. We further show how to construct confidence intervals by inverting the LR statistics in both methods. Our three inferential methods provide considerable flexibility to practitioners. When instruments are available, the 2SLS method of estimation can be used, coupled with use of the bootstrap for inference. When instruments are absent, the other two methods can be used.

Two specification tests are suggested, one designed to check for the presence of endogeneity and the other to check for threshold effects. Our results show that it is possible to test for threshold effects in the absence of instrumentation even if endogeneity is present. An important implication of the test for endogeneity in empirical work is that it helps to assess whether instruments are required to achieve consistent estimation of the structural coefficients. Both tests are similar to score tests and have convenient asymptotic normal distributions, although a wild bootstrap procedure is suggested to determine critical values for improved finite sample performance.

References

Acemoglu, D., S. Johnson and J.A. Robinson, 2001, The Colonial Origins of Comparative Development: An Empirical Investigation, *American Economic Review*, 91, 1369-1401.

Amemiya, T., 1974, The Nonlinear Two-Stage Least-Squares Estimator, *Journal of Econometrics*, 2, 105-110.

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

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

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

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

Bierens, H.J. and W. Ploberger, 1997, Asymptotic Theory of Integrated Conditional Moment Tests, *Econometrica*, 65, 1129-1151.

Blundell R. and J.L. Powell, 2003, Endogeneity in Nonparametric and Semiparametric Regression Models, in Dewatripont, M., L.P. Hansen, and S.J. Turnovsky, eds., *Advances in Economics and Econometrics: Theory and Applications*, Eighth World Congress, Vol. II (Cambridge University Press), 312-357.

Brown, B.W. and W.K. Newey, 2002, Generalized Method of Moments, Efficient Bootstrapping, and Improved Inference, *Journal of Business & Economic Statistics*, 20, 507-517.

Caner, M. and B.E. Hansen, 2004, Instrumental Variable Estimation of a Threshold Model, *Econometric Theory*, 20, 813-843.

Chan, K.S., 1993, Consistency and Limiting Distribution of the Least Squares Estimator of a Threshold Autoregressive Model, *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, B. and Y. Hong, 2012, Testing for Smooth Structural Changes in Time Series Models via Nonparametric Regression, *Econometrica*, 80, 1157-1183.

Chiou, Y.-Y., M.-Y. Chen and J. Chen, 2018, Nonparametric Regression with Multiple Thresholds: Estimation and Inference, *Journal of Econometrics*, 206, 472–514.

Cleveland, W.S., E. Grosse and W.M. Shyu, 1992, Local Regression Models, *Statistical Models in S*, J.M. Chambers, and T.J. Hastie, eds., 309–376, New York: Chapman and Hall.

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.

De Jong, R.M., 1996, The Bierens Test Under Data Dependence, *Journal of Econometrics*, 72, 1-32.

Delgado, M.A. and J. Hidalgo, 2000, Nonparametric Inference on Structural Breaks, *Journal of Econometrics*, 96, 113-144.

Delgado, M.A. and W.G. Manteiga, 2001, Significance Testing in Nonparametric Regression Based on the Bootstrap, *Annals of Statistics*, 29, 1469-1507.

Fan, Y. and Q. Li, 1996, Consistent Model Specification Tests: Omitted Variables and Semiparametric Functional Forms, *Econometrica*, 64, 865-890.

Fan, Y. and Q. Li, 2000, Consistent Model Specification Tests: Kernel-Based Tests versus Bierens' ICM Tests, *Econometric Theory*, 16, 1016-1041.

Fan, J. and W. Zhang, 2008, Statistical Methods with Varying Coefficient Models, *Statistics and Its Interface*, 1, 179-195.

Frankel, J. and D. Romer, 1999, Does Trade Cause Growth?, *American Economic Review*, 89, 379-399.

Gao, J., I. Gijbels and S. Van Bellegem, 2008, Nonparametric Simultaneous Testing for Structural Breaks, *Journal of Econometrics*, 143, 123-142.

Gijbels, I. and A. Goderniaux, 2004, Bandwidth Selection for Changepoint Estimation in Nonparametric Regression, *Technometrics*, 46, 76-86.

Gu, J., D. Li and D. Liu, 2007, Bootstrap Non-parametric Significance Test, *Nonparametric Statistics*, 19, 215-230.

Hall, A.R., S. Han and O. Boldea, 2012, Inference Regarding Multiple Structural Changes in Linear Models with Endogenous Regressors, *Journal of Econometrics*, 170, 281-302.

Hall, P. and J.L. Horowitz, 1996, Bootstrap Critical Values for Tests Based on Generalized Method of Moments Estimators, *Econometrica*, 64, 891-916.

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

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

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

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

Härdle, W. and E. Mammen, 1993, Comparing Nonparametric versus Parametric Regression Fits, *Annals of Statistics*, 21, 1926-1947.

Hastie, T.J. and R.J. Tibshirani, 1993, Varying-Coefficient Models, *Journal of the Royal Statistical Society, Series B*, 55, 757–796.

Henderson, D.J., C.F. Parmeter and L. Su, 2017, M-Estimation of a Nonparametric Threshold Regression Model, mimeo.

Hidalgo, J., 1995, A Nonparametric Conditional Moment Test for Structural Stability, *Econometric Theory*, 11, 671-698.

Horowitz, J.L., 1992, A Smoothed Maximum Score Estimator for the Binary Response Model, *Econometrica*, 60, 505-531.

Kapetanios, G., 2010, Testing for Exogeneity in Threshold Models, *Econometric Theory*, 26, 231-259.

Kourtellos, A., T. Stengos and C.-M. Tan, 2016, Structural Threshold Regression, *Econometric Theory*, 32, 827-860.

Kourtellos, A., T. Stengos and Y. Sun, 2017, Endogeneity in Semiparametric Threshold Regression, mimeo.

Kristensen, D., 2012, Non-Parametric Detection and Estimation of Structural Change, *Econometrics Journal*, 15, 420-461.

Lee, S., 2014, Asymptotic Refinements of a Misspecification-Robust Bootstrap for Generalized Method of Moments Estimators, *Journal of Econometrics*, 178, 398-413.

Li, Q. and S. Wang, 1998, A Simple Consistent Bootstrap Test for a Parametric Regression Function, *Journal of Econometrics*, 87, 145-165.

Liu, R.Y., 1988, Bootstrap Procedures under Some Non i.i.d. Models, *Annals of Statistics*, 16, 1696-1708.

Mammen, E., 1992, *When Does Bootstrap Work?: Asymptotic Results and Simulations*, New York: Springer.

Manski, C.F., 1975, Maximum Score Estimation of the Stochastic Utility Model of Choice, *Journal of Econometrics*, 3, 205-228.

Manski, C.F., 1985, Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator, *Journal of Econometrics*, 27, 313-333.

Müller, H.-G., 1991, Smooth Optimum Kernel Estimators Near Endpoints, *Biometrika*, 78, 521-530.

Müller, H.-G. and K.S. Song, 1997, Two-stage Change-point Estimators in Smooth Regression Models, *Statistics and Probability Letters*, 34, 323-335.

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, 2111-2245.

Papageorgiou, C., 2002, Trade as a Threshold Variable for Multiple Regimes, *Economics Letters*, 77, 85-91.

Perron, P. and Y. Yamamoto, 2015, Using OLS to Estimate and Test for Structural Changes in Models with Endogenous Regressors, *Journal of Applied Econometrics*, 30, 119-144.

Porter, J., 2003, Estimation in the Regression Discontinuity Model, Mimeo, Department of Economics, University of Wisconsin at Madison.

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 US GNP, *Journal of Applied Econometrics*, 10, 109-125.

Powell, J.L., J.H. Stock and T.M. Stoker, 1989, Semiparametric Estimation of Index Coefficients, *Econometrica*, 57, 1403-1430.

Robinson, P.M., 1988, Root-N-Consistent Semiparametric regression, *Econometrica*, 56, 931-954.

Robinson, P.M., 1989, Nonparametric Estimation of Time-Varying Parameters, *Statistical Analysis and Forecasting of Economic Structural Change*, P. Hackl, eds., Berlin: Springer, Ch. 15, 253-264.

Robinson, P.M., 1991, Time-Varying Nonlinear Regression, *Economic Structural Change Analysis and Forecasting*, P. Hackl, eds., Berlin: Springer, Ch. 13, 179-190.

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

Seo, M.H. and Y. Shin, 2016, Dynamic Panels with Threshold Effect and Endogeneity, *Journal of Econometrics*, 195, 169-186.

Stone, C.J., 1980, Optimal Rates of Convergence for Nonparametric Estimators, *Annals of Statistics*, 8, 1348-1360.

Stryhn, H., 1996, The Location of the Maximum of Asymmetric Two-Sided Brownian Motion with Triangular Drift, *Statistics & Probability Letters*, 29, 279-284.

Stute, W., W.G. Manteiga and M.P. Quindimil, 1998, Bootstrap Approximation in Model Checks for Regression, *Journal of the American Statistical Association*, 93, 141–149.

Su, L. and Z. Xiao, 2008, Testing Structural Change in Time-series Nonparametric Regression Models, *Statistics and Its Interface*, 347-366.

Tan, C.M., 2010, No One True Path: Uncovering the Interplay Between Geography, Institutions, and Fractionalization in Economic Development, *Journal of Applied Econometrics*, 25, 7, 1100-1127.

Wu, C.F.J., 1986, Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis, *Annals of Statistics*, 14, 1261-1295.

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

Yu, P., 2013a, Inconsistency of 2SLS Estimators in Threshold Regression with Endogeneity, *Economics Letters*, 120, 532-536.

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

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

Yu, P., 2015a, Consistency of the Least Squares Estimator in Threshold Regression with Endogeneity, *Economics Letters*, 131, 41-46.

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

Yu, P., 2016, Understanding Estimators of Treatment Effects in Regression Discontinuity Designs, *Econometric Reviews*, 35, 586-637.

Yu, P., 2017, Threshold Regression Under Misspecification, mimeo.

Yu, P. and X. Fan, 2019, Threshold Regression With A Threshold Boundary, mimeo.

Yu, P., Q. Liao and P.C.B. Phillips, 2018, Control Function Approaches in Threshold Regression with Endogeneity, mimeo.

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

Yu, P., Q. Liao and P.C.B. Phillips, 2019, Online Supplement for 'Inference and Specification Testing in Threshold Regression with Endogeneity', Working Paper Supplement.

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

Zheng, J.X., 1996, A Consistent Test of Functional Form via Nonparametric Estimation Techniques, *Journal of Econometrics*, 75, 263–289.