Inferences and Specification Testing in Threshold Regression with Endogeneity

Qin Liao*
University of Hong Kong

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

Ping Yu‡
University of Hong Kong

First Version: June 2017
This Version: February 2018

Abstract

This paper proposes three inference methods for the threshold point in endogenous threshold regression and two specification tests to test for the presence of endogeneity and threshold effects without relying on instrumentation of the covariates. The first inference method is a parametric 2SLS method and is suitable when instruments are available. The second and third methods are based on smoothing the objective function of the IDKE in different ways and do not require instrumentation. Both specification tests are score-type tests; especially, 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.

Keywords: threshold regression, endogeneity, confidence interval, 2SLS, IDKE, specification testing, bootstrap, U-statistic

JEL-Classification: C21, C24, C26,

*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, P.O. Box 208281, New Haven, CT, USA; email: peter.phillips@yale.edu
‡Faculty of Business and Economics, The University of Hong Kong, Pokfulam Road, Hong Kong; corresponding author email: pingyu@hku.hk.
1 Introduction

In recognition of potential shifts in economic relationships, threshold models have become increasingly popular in recent econometric practice; see Hansen (2011) for a recent overview of their 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, see, e.g., Potter (1995). Threshold models are also useful in cross sectional applications. For example, Hansen (2000) applies the threshold model to show that depending on the starting point, rich countries and poor countries have different growth patterns. All these literature assumes exogenous regressors and threshold variable, however, in practical works the threshold model is commonly plagued by 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 usual setup of endogenous threshold regression (TR) is

\[
y = x'\beta_1 1(q \leq \gamma) + x'\beta_2 1(q > \gamma) + \varepsilon
\]

\[
= x'\beta + x'\delta 1(q \leq \gamma) + \varepsilon,
\]

\[
\mathbb{E}[\varepsilon|x] \neq 0,
\]

where \( x = (1, x', q)' \in \mathbb{R}^{d+1}, \) \( d \) is the dimension of the nonconstant covariates \( (x, q) \), and the parameter of interest is \( \theta = (\beta_1', \beta_2', \gamma)' \) or equivalently, \( \theta := (\beta', \delta', \gamma)' \) 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 \( \gamma \). The literature on estimation of this model includes the following three main contributions. First, Caner and Hansen (2004) use two-stage least squares (2SLS) to estimate \( \gamma \) in the small-threshold-effect framework of Hansen (2000), but they assume \( q \) is exogenous in the sense of \( \mathbb{E}[\varepsilon|x] = \mathbb{E}[\varepsilon|x] \). Second, also in Hansen (2000)’s framework, Kourtelllos et al. (2016) use a control function approach to deal with the case where \( q \) is also endogenous. But their setup is parametric (see Kourtelllos et al. (2017) for a semiparametric extension) and the asymptotic theory is not fully resolved. More specifically, in a companion paper, Liao et al. (2017), we show that the consistency, convergence rate and asymptotic distribution of Kourtelllos et al. (2016)’s estimator are correct only in very special cases. Third, Yu and Phillips (2017) use an integrated difference kernel estimator (IDKE) to estimate \( \gamma \) in the fixed-threshold-effect framework of Chan (1993). Their estimator can be applied regardless of whether \( q \) is endogenous and whether any instruments exist (as required in the previous two methods). Even if no instruments are available such that the model reduces to a nonparametric threshold regression, their estimator is still \( n \)-consistent, just as in the parametric setup. The endogeneity problem is also considered in related structural change literature, where the threshold variable is the time index and is always exogenous; see Yu and Phillips (2017) for a detailed literature review.

In spite of the theoretical developments in the estimation of \( \gamma \) in endogenous threshold regression, inferences on \( \gamma \) are still difficult in practice especially when \( q \) is endogenous. First, Caner and Hansen (2004)’s method can be applied only if \( q \) is exogenous. As shown in Yu (2013a), Caner and Hansen (2004)’s estimator is generally inconsistent when \( q \) is endogenous. Second, as mentioned above, Kourtelllos et al. (2016)’s approach is not generically applicable. Third, the asymptotic distribution of the IDKE in Yu and Phillips (2017) is too complicated to be applicable. This paper seeks to alleviate this difficulty by proposing three new methods of confidence interval (CI) construction for \( \gamma \). All three methods can be applied regardless of

---

1 If \( q \) is exogenous, then Kourtelllos et al. (2016)’s estimator is the same as Caner and Hansen (2004)’s estimator.

2 We won’t discuss inferences on regular parameters such as \( \beta \) and \( \delta \) because such inferences are standard in the literature; see, e.g., Caner and Hansen (2004) and Yu and Phillips (2017).
whether \( q \) is endogenous or not; to our knowledge, they are the only valid and applicable inference methods which are robust to the endogeneity of \( q \). The first method is a parametric 2SLS method and requires instruments, while the other two methods are based on smoothing the objective function of the IDKE in different ways so instruments are not necessary. 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 fully uses the data around the threshold point. Both these IDKE-smoothing methods are nonparametric and require kernel and bandwidth selection. Practitioners can select an inferential method among these three based on its suitability to the data and data availability. For example, if instruments are available, then the first method can be used; otherwise, the second and third methods may be preferable. Furthermore, we put forward two specification tests; the first one tests for endogeneity and the second one tests for the presence of threshold effects both with and without instruments. Both kinds of tests are score-type tests, i.e., we construct the test statistics under the null, so their asymptotic properties are easy to develop. More importantly, our tests of structural shifts are easier to implement in practice than the popular Wald test especially when instruments are unavailable. Practitioners should pay more attention to the inference methods and specification tests without instrumentation because good instruments are often hard to find and justify in practical work.

The rest of the paper is organized as follows. Section 2 provides an overview of the three inference methods and the two specification tests; Sections 3 to 5 discuss the three inference methods one by one in details 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. Proofs of theorems with supporting propositions and lemmas are given in Supplements A, B and C, respectively. Additional discussions on parametric tests for the presence of threshold effects when instruments are present are given in Supplement D because such tests are well understood in the literature.

A word on notation. The three inference methods we are going to discuss in the paper are labeled as Method I, II and III, respectively. We refer YP for Yu and Phillips (2017), DH for Delgado and Hidalgo (2000), CH for Caner and Hansen (2004), and HHB for Hall et al. (2012) hereinafter. The symbol \( \ell \) is used to indicate the two regimes in (1) or the two specification tests, and is not written out explicitly as ‘\( \ell = 1, 2 \)’.

For a matrix \( A \), \( A > 0 \) means that it is positive definite, and \( \text{span}(A) \) means its column space. For any random vector \( x \), \( x_{\leq \gamma} := x^1(q \leq \gamma) \) and \( x_{> \gamma} \) is similarly defined. For any two random vectors \( x \) and \( y \), \( x \perp y \) means \( x \) and \( y \) are independent. A parameter with a subscript 0 means its true value.

## 2 Overview of Inferences and Specification Testing

In this section, we briefly overview the three inference methods and the two specification tests. Another purpose of this section is to introduce necessary notations for future discussion.

### 2.1 Overview of Inference Methods for the Threshold Point

If we write the model (1) as \( y = G(x, q; \theta) + \varepsilon, \mathbb{E}[\varepsilon|x, q] \neq 0 \), where \( G(x, q; \theta) = x^0/\beta + x^11(q \leq \gamma) \) is a nonlinear function of \((x, q)\), then the model can be treated as a nonlinear endogenous problem. 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; this can explain why the 2SLS estimators in Yu (2013a) are not consistent. To restore the consistency of 2SLS, we must maintain the linear structure of the model. In other words, instead of projecting \((x, q)\) on available instruments \( z \), we first project \((x, x_{\leq \gamma})\) for a fixed \( \gamma \) on \( z \) to get the projections
Based on this intuition, we will show that: (i) \( \hat{\beta}(\gamma) \) and \( \hat{\delta}(\gamma) \) by regressing \( y \) on \( \hat{x} \) and \( \hat{x}_{\leq \gamma} \); finally, \( \hat{\gamma} \) is obtained by minimizing \( \sum_{i=1}^{n}(y_i - \hat{x}'i\hat{\beta}(\gamma) + \hat{x}'i\hat{\delta}(\gamma))^2 \), and set \( \hat{\beta} = \hat{\beta}(\hat{\gamma}), \hat{\delta} = \hat{\delta}(\hat{\gamma}) \). It is easy to see that this procedure is equivalent to
\[
\left( \hat{\beta}, \hat{\delta}, \hat{\gamma} \right) = \arg\min_{\beta, \delta, \gamma} (Y - X\beta - X_{\leq \gamma}\delta)'P_Z(Y - X\beta - X_{\leq \gamma}\delta),
\]
where \( Y, X, X_{\leq \gamma} \) and \( Z \) are stacking \( y_i, x_i', x_{\leq \gamma,i}' \) and \( z \) respectively, and \( P_Z \) is the projection matrix on \( \text{span} (Z) \). This method, labelled as Method I, treats \( \gamma \) as a regular parameter and is exactly the nonlinear 2SLS of Amemiya (1974). We will also show that this 2SLS estimator is a special case of the GMM estimator considered in HHB (see also Andrews (1993)) but has more desirable asymptotic properties.

To understand the estimator \( \hat{\gamma} \), we explore a simple case where \( x = 1, \beta_0 = 0 \) 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 \( \hat{\gamma} = \hat{F}_q^{-1}(\hat{\mathbb{E}}[y]) \).

Based on this intuition, we will show that: (i) \( \hat{\gamma} \) is \( \sqrt{n} \)-consistent, asymptotically normal, and the asymptotic variance involves the density of \( q \) at \( \gamma_0 \) (i.e., \( f_q(\gamma_0) \)) as in quantile regression; and (ii) different from the usual threshold regression estimators where \( \hat{\gamma} \) is asymptotically independent of \( (\hat{\beta}, \hat{\delta}) \), this new estimator \( \hat{\gamma} \) correlates with \( (\hat{\beta}, \hat{\delta}) \) asymptotically. Given the asymptotic normality of \( \hat{\gamma} \), the confidence interval for \( \gamma \) can be constructed by the bootstrap as in quantile regression to avoid nonparametric estimation of \( f_q(\gamma_0) \) in the asymptotic distribution.

Note that different from CH, our 2SLS estimator can be applied no matter \( q \) is endogenous or not. Yu (2013a) shows that when \( q \) is exogenous, CH's estimator is inconsistent if the first stage predictor is a projection rather than a conditional mean. On the contrary, our 2SLS estimator requires only a linear projection in the first stage, so is more robust in this aspect.

Away from this IV-based estimation, we next introduce IV-free estimators in Methods II and III by extending the IDKE of \( Y_P \) in different directions. Without instruments, the model reduces to a nonparametric threshold regression,
\[
y = m(x, q) + u = g(x, q) + x'\delta 1(q \leq \gamma) + u,
\]
where \( g(x, q) = x'\beta + \mathbb{E}[\varepsilon|x, q] \) is any smooth function, and \( u = \varepsilon - \mathbb{E}[\varepsilon|x, q] \) satisfies \( \mathbb{E}[u|x, q] = 0 \). To construct the IDKE of \( \gamma \), 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 a generalized kernel function is obtained as follows. Define the space
\[
\mathcal{M}_p([a, b]) = \left\{ g \in \text{Lip}([a, b]), \int_a^b x^j g(x)dx = \begin{cases} 
1, & \text{if } j = 0, \\
0, & \text{if } 1 \leq j \leq p - 1
\end{cases} \right\},
\]
where \( \text{Lip}([a, b]) \) 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) \( k_-(\cdot, r) \in \mathcal{M}_p([-1, r]) \) and \( k_+(\cdot, r) \in \mathcal{M}_p([-r, 1]) \).

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

3
(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 \( \gamma \); see Appendix A of Porter and Yu (2015) for some 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 the parameter space.

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

We do not restrict \( \delta_{0} \) to be fixed or shrinking to zero; actually, \( \delta_{0} \) is 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 dealt with, at least in a conceptually straightforward manner by using the continuous covariate estimator within samples homogeneous in the discrete covariates, at the expense of extra additional notations. Requiring the support of \( x \) to be \([0, 1]^{d-1}\) is not restrictive and can be achieved by the use of some 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

\[
\begin{align*}
    k(\cdot) &= k_{+}(\cdot, 1) = k_{-}(\cdot, 1) \in \mathcal{M}_{p}([-1, 1]),
    k_{h}(u) &= k(u/h)/h, \\
    k_{+}(\cdot) &= k_{+}(\cdot, 0) \in \mathcal{M}_{p}([0, 1]),
    k_{h}^{+}(u) &= k_{+}(u/h)/h, \\
    k_{-}(\cdot) &= k_{-}(\cdot, 0) \in \mathcal{M}_{p}([-1, 0]),
    k_{h}^{-}(u) &= k_{-}(u/h)/h,
\end{align*}
\]

and

\[
k_{h}(u, t) = \begin{cases} 
    \frac{1}{h} k \left( \frac{u}{h} \right), & \text{if } h \leq t \leq 1 - h, \\
    \frac{1}{h} k_{+} \left( \frac{u}{h}, \frac{t}{h} \right), & \text{if } 0 \leq t \leq h, \\
    \frac{1}{h} k_{-} \left( \frac{u}{h}, \frac{1-t}{h} \right), & \text{if } 1 - h \leq t \leq 1.
\end{cases}
\]  

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 will only need \( k_{h}(u, t) \) to be a first order kernel function (but with different specifications) to estimate \( \gamma \) in our Methods II and III. Formally, we require in Method II that

**Assumption K:** \( k_{h}(u, t) \) takes the form of (3) with \( p = 1, k_{+}(0) = k_{-}(0) = 0, \) and \( k'_{+}(0) > 0, k'_{-}(0) < 0. \)

and in Method III that

**Assumption K':** \( k_{h}(u, t) \) takes the form of (3) with \( p = 1 \) and \( k_{+}(0) = k_{-}(0) > 0; \)

Assumption K mimics assumptions B2 and B3 of DH and Assumption K' is Assumption K in YP.

Given \( k_{h}(u, t) \), the IDKE of \( \gamma \) is constructed as an extremum estimator

\[
\hat{\gamma} = \arg \max_{\gamma} \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{1}{n-1} \sum_{j=1,j \neq i}^{n} y_{ij} k'_{h,i,j} - \frac{1}{n-1} \sum_{j=1,j \neq i}^{n} y_{ij} K'_{h,i,j} \right]^{2}.
\]

\[
= \arg \max_{\gamma} \frac{1}{n} \sum_{i=1}^{n} \tilde{\Delta}^{2}_{i} (\gamma) =: \arg \max_{\gamma} \tilde{Q}_{n} (\gamma),
\]

\( ^{3}\)Note here that the usual symmetric kernel is a second order kernel, but the boundary kernel is only a first order kernel because \( \int u k_{h}(u, t) \neq 0. \)
where

\begin{align*}
K_{h,ij}^{\gamma+} &= \prod_{t=1}^{d-1} k_h(x_{ij} - x_{it}, x_{it}) \cdot k_h^+(q_j - \gamma) =: K_{h,ij}^x k_h^+(q_j - \gamma), \\
K_{h,ij}^{\gamma-} &= \prod_{t=1}^{d-1} k_h(x_{ij} - x_{it}, x_{it}) \cdot k_h^-(q_j - \gamma) =: K_{h,ij}^x k_h^-(q_j - \gamma).
\end{align*}

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 middle points of the contiguous \(q_i\)'s in the optimization process of (4). In other words, the argmax operator is a middle-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 \(\sum_{j=1,j\neq i}^n K_{h,ij}^{\gamma+}\) does not appear in the construction of \(\hat{\gamma}\), thereby avoiding random denominator issues in conditional mean estimation and simplifying the derivation of the limit distribution of \(\hat{\gamma}\), a technique that dates back at least to Powell et al. (1989). This form of \(\hat{\gamma}\) 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 the local polynomial estimator (LPE) might be considered to ameliorate bias. However, when \(d\) is large, there are not many data points in a \(h\) neighborhood of \((x', \gamma)\). As a result, not only does the LLS lose degrees of freedom (by estimating more parameters) but its denominator matrix tends to be close to singular. Furthermore, different from the regular parameter (such as the conditional mean) estimation, use of the LLS does not affect the first-order asymptotic distribution of \(\hat{\gamma}\).

The CI construction based on the asymptotic distribution of \(\hat{\gamma}\) under Assumption K' and fixed threshold effects (i.e., in the framework of YP) is challenging because the asymptotic distribution involves a compound Poisson process which is hard to simulate. Methods II and III use different smoothing schemes to achieve operable asymptotic distributions. Method II assumes fixed threshold effects but uses the data in the neighborhood of \(\gamma_0\) only marginally. The resulting asymptotic distribution is normal and the CI construction of \(\hat{\gamma}\) excludes \(j = i\), which is a standard strategy in converting a V-statistic to a U-statistic. Also, the normalization factor \(\sum_{j=1,j\neq i}^n K_{h,ij}^{\gamma+}\) does not appear in the construction of \(\hat{\gamma}\), thereby avoiding random denominator issues in conditional mean estimation and simplifying the derivation of the limit distribution of \(\hat{\gamma}\), a technique that dates back at least to Powell et al. (1989). This form of \(\hat{\gamma}\) 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 the local polynomial estimator (LPE) might be considered to ameliorate bias. However, when \(d\) is large, there are not many data points in a \(h\) neighborhood of \((x', \gamma)\). As a result, not only does the LLS lose degrees of freedom (by estimating more parameters) but its denominator matrix tends to be close to singular. Furthermore, different from the regular parameter (such as the conditional mean) estimation, use of the LLS does not affect the first-order asymptotic distribution of \(\hat{\gamma}\).

We provide some intuitions here on the validity of \(\hat{\gamma}\). For this purpose, we impose the following assumptions on the distribution of \((x', q)\) and on \(g(x, q)\).

**Assumption F:** The density \(f(x, q)\) of \((x, q)\) is second order continuously differentiable and satisfies \(0 < f \leq f(x, q) \leq F < \infty\) for \((x, q) \in \mathcal{X} \times \Gamma\), where \(\Gamma := (\gamma - \epsilon, \gamma + \epsilon)\) for some \(\epsilon > 0\) and some fixed quantities \((\ell, \bar{f}, \bar{F})\).

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

Assumption F implies that \(f_{q}(\gamma)\) is continuous, and \(0 < \ell_{\gamma} \leq f_{q}(\gamma) \leq \bar{F}_{\gamma} < \infty\) for \(\gamma \in \Gamma\) and fixed \((\ell_{\gamma}, \bar{F}_{\gamma})\), and the conditional density \(f_{q|x}(x|q)\) is bounded below and above for \((x, q) \in \mathcal{X} \times \Gamma\); see Yu and Zhao (2013) for relaxing 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 part of
Assumption F implies that \( \gamma_0 \) is not on the boundary of \( Q \). Under these two assumptions, we expect the objective function \( \tilde{Q}_n (\gamma) \) to converge to

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

Since \( f(x) \) and \( f(x, \gamma) \) are continuous in \( x \) and \( \gamma \), there will be a jump in the limit only if \( \gamma = \gamma_0 \) which provides identifying information. As a result, the threshold point can be identified and consistently estimated by maximizing \( \tilde{Q}_n (\gamma) \). Given that \( \mathbb{E}[y|x, q = \gamma_0-] - \mathbb{E}[y|x, q = \gamma_0+] = (1, x', \gamma_0) \delta_0 \), we need the following assumption to identify \( \gamma_0 \).

**Assumption I**: \((1, x', \gamma_0) \delta_0 \neq 0 \) for \( x \) in some set of positive Lebesgue measure in \( X \).

For comparison, we state the following Assumption I’.

**Assumption I’**: \( \delta_0 \neq 0 \), where \( \neq \) here means that at least one element is unequal.

Note that Assumption I is stronger than Assumption I’. For example, \( \delta_0 = \left\{ \begin{array}{ll} (1, 0, -\frac{1}{\gamma_0})' & \text{if } \gamma_0 \neq 0, \\ (0, 0, 1)' & \text{if } \gamma_0 = 0, \end{array} \right. \) is nonzero but does not satisfy Assumption I. The stated condition implies that \( P'((1, x', \gamma_0) \delta_0 \neq 0) > 0 \), which excludes the continuous threshold regression of Chan and Tsay (1998).

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

\[
\tilde{\gamma} = \arg \max_{\gamma} \tilde{\Delta}^2_0 (\gamma),
\]

where \( \tilde{\Delta}_0 (\gamma) = \frac{1}{n} \sum_{j=1}^{n} y_j K_{h,j}^{-} - \frac{1}{n} \sum_{j=1}^{n} y_j K_{h,j}^{+} \) with

\[
K_{h,j}^{-} = \prod_{l=1}^{d-1} k_h(x_{lj} - x_{ol}, x_{ol}) \cdot k_h^{-} (q_j - \gamma) , \quad K_{h,j}^{+} = \prod_{l=1}^{d-1} k_h(x_{lj} - x_{ol}, x_{ol}) \cdot k_h^{+} (q_j - \gamma),
\]

and where \( x_o \) is some fixed point in the interior of \( X \).\(^5\) 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 \( x_i \)'s, thereby removing the problem of choosing \( x_o \). Further, the usage of all the data ensures that the IDKE has greater identifying capability than the DKE in both Methods II and III.

### 2.2 Overview of Two Specification Tests

The first specification test addresses potential endogeneity and the corresponding hypotheses \( H^{(1)} \) are

\[
H_0^{(1)} : \mathbb{E}[e|x, q] = 0, \\
H_1^{(1)} : \mathbb{E}[e|x, q] \neq 0.
\]

\(^5\)Strictly speaking, DH normalize the first term of \( \tilde{\Delta}_0 (\gamma) \) by \( \tilde{f}_n^{-} =: \frac{1}{n} \sum_{j=1}^{n} K_{h,j}^{-} \) and the second term by \( \tilde{f}_n^{+} =: \frac{1}{n} \sum_{j=1}^{n} K_{h,j}^{+} \). However, their estimator is asymptotically equivalent to \( \arg \max_{\gamma} \tilde{\Delta}^2_0 (\gamma) / f(x_o, \gamma_0)^2 \) and has the same asymptotic distribution as \( \tilde{\gamma} \).
This exogeneity test can be conducted prior to model estimation and the techniques developed in Fan and Li (1996) and Zheng (1996) can be used to test the null $H_0^{(1)}$. In the second test, the hypotheses $H^{(2)}$ are

- $H_0^{(2)} : \beta_1 = \beta_2$ or $\delta = 0$,
- $H_1^{(2)} : \beta_1 \neq \beta_2$ or $\delta \neq 0$.

If $H^{(1)}_0$ is not rejected, i.e., there is no evidence of endogeneity, then $H^{(2)}$ involves 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 $H^{(1)}_0$ is rejected, the ensuing situation is more complex. When there are instruments, Wald-type test statistics can be used, such as the sup-statistic in Section 5 of Caner and Hansen (2004), or score-type statistics as in Yu (2013b). Since the asymptotic distributions of both these types of test statistics are not pivotal, the simulation method of Hansen (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 standard nowadays. When there are no instruments, the Wald-type statistic is hard to implement since its asymptotic distribution is hard to derive given that $\hat{\delta}$ can only be estimated at a nonparametric rate – see Section 3.3 of Porter and Yu (2015) for discussion\footnote{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.}. 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 $H^{(2)}_0$ encompasses more data generating processes than the null hypothesis in the usual structural change literature where $m(x,q)$ has a simple parametric form. In other words, the usual tests have power against alternatives in which $m(x,q)$ does not take the form $x'\beta + x'\delta 1(q \leq \gamma)$ (see, e.g., Section 5.4 of Andrews (1993))\footnote{In this framework and assuming $m(x,q) = x'\beta(q)$, 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 in Section 6\footnote{We can also imagine cases where the parametric test does not have power although there is a nonparametric threshold effect; see Example 1 of Hidalgo (1995).}, e.g., we can construct residuals $\hat{\epsilon}_i$ in $I^{(2)}_t$ 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).}, but our test will not have any power against such cases as long as $m(x,q)$ is continuous. A simple example may clarify the point. Suppose $m(x,q) = \alpha + \beta q$ while the specification in $H^{(1)}$ is $y = \alpha + \delta 1(q \leq \gamma) + \varepsilon$. It is easy to see that the usual tests have power against $m(x,q)$ although it is very smooth. In summary, the usual tests have power against both misspecification and structural change, while our test has power only against structural change, which might be more relevant in practical work\footnote{However, the $H^{(1)}_0$ involves 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 $H^{(1)}_0$ is rejected, the ensuing situation is more complex. When there are instruments, Wald-type test statistics can be used, such as the sup-statistic in Section 5 of Caner and Hansen (2004), or score-type statistics as in Yu (2013b). Since the asymptotic distributions of both these types of test statistics are not pivotal, the simulation method of Hansen (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 standard nowadays. When there are no instruments, the Wald-type statistic is hard to implement since its asymptotic distribution is hard to derive given that $\hat{\delta}$ can only be estimated at a nonparametric rate – see Section 3.3 of Porter and Yu (2015) for discussion}. But this 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 power. Understandably so, because our test is essentially nonparametric whereas the usual tests are parametric.

In the discussion of the two specification tests, $H_0$ indicates both $H^{(1)}_0$ and $H^{(2)}_0$, and $H_1$ indicates both $H^{(1)}_1$ and $H^{(2)}_1$, $I^{(1)}_q = 1(q \in \Gamma)$, $I^{(2)}_q = 1(q_i \in \Gamma)$, $m_i = m(x_i, q_i) = E[y_i(x_i, q_i)]$, $f_i = f(x_i, q_i)$, $K_{h,ij} = K_{h}^{(i)} \cdot k_h(q_j - q_i)$, and $L_{b,ij} = L_{b}^{(i)} \cdot l_b(q_j - q_i)$ with $l_b(\cdot)$ similarly defined as $k_h(\cdot)$. Denote the class of probability measures under $H^{(l)}_0$ as $\mathcal{H}^{(l)}_0$ and under $H^{(l)}_1$ as $\mathcal{H}^{(l)}_1$. Both $\mathcal{H}^{(l)}_0$ and $\mathcal{H}^{(l)}_1$ 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 expectations as $P_m$ and $E_m$, respectively. To unify notation, we define $u_i = y_i - E[y_i | x_i, q_i] = y_i - m_i$ under both the null and alternative in both tests.
For the first test, we use the test statistic
\[ I_n^{(1)} = \frac{nh^{d/2}}{n(n-1)} \sum_{i,j \neq i} K_{h,ij} \hat{e}_i \hat{e}_j, \]
and, for the second, we use
\[ I_n^{(2)} = \frac{nh^{d/2}}{n(n-1)} \sum_{i,j \neq i} 1_i^T 1_j^T K_{h,ij} \hat{e}_i \hat{e}_j. \]

The exact forms of \( \hat{e}_i \) in these two tests are defined later. To motivate the statistics, let \( e = y - \bar{m}(x) \), where
\[ \bar{m}(\cdot) = \arg \inf_{\tilde{m}(x,q) = x'\beta + x'\delta 1(q \leq \gamma)} \mathbb{E} [ (y - \tilde{m}(x,q))^2 ] \]
\[ = \arg \inf_{\tilde{m}(x,q) = x'\beta + x'\delta 1(q \leq \gamma)} \mathbb{E} [ (m(x,q) - \tilde{m}(x,q))^2 ] \]
in the first test, and
\[ \bar{m}(\cdot) = \arg \inf_{\tilde{m} \in C_s(B, \mathcal{X} \times \mathcal{Q})} \mathbb{E} [ (y - \tilde{m}(x,q))^2 1_q^T ] \]
\[ = \arg \inf_{\tilde{m} \in C_s(B, \mathcal{X} \times \mathcal{Q})} \mathbb{E} [ (m(x,q) - \tilde{m}(x,q))^2 1_q^T ] , \]
in the second test, where \( C_s(B, \mathcal{X} \times \mathcal{Q}) \) 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 \( \tilde{m}(x,q) = x'\beta + x'\delta 1(q \leq \gamma) \) to approximate \( m(x,q) \) in the first test and use \( \tilde{m} \in C_s(B, \mathcal{X} \times \mathcal{Q}) \) to approximate \( m(x,q) \) in the second test. Note that in the first test the model need not have a threshold effect. The class of functions \( \{x'\beta + x'\delta 1(q \leq \gamma)\} \) include the linear function where \( \delta = 0 \), the continuous threshold regression of Chan and Tsay (1998) where \( \delta \neq 0 \) but \( \delta_x = 0 \) and \( \delta_\alpha + \delta_\gamma = 0 \), and the usual threshold regression where \( \delta_x \neq 0 \) or \( \delta_\alpha \neq \delta_\gamma \neq 0 \); see Yu (2017) for more discussions on misspecified threshold regression. Here, \( \delta \) is partitioned according to the partition of \( x = (1, x', q)' \) as \( (\delta_\alpha, \delta_x', \delta_\gamma' )' \).

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

To construct a feasible test statistic, we need the sample analogue of \( e \) and \( \mathbb{E}[e|x,q] f(x,q) \). For the first test, the sample counterpart of \( e \) is
\[ \hat{e}_i = y_i - \hat{y}_i = y_i - [x'\hat{\beta} + x'\hat{\delta} 1(q_i \leq \hat{\gamma})], \]
where \( (\hat{\beta}', \hat{\delta}', \hat{\gamma})' \) is the least squares estimator. For the second test, let
\[ \hat{e}_i = y_i - \hat{y}_i = (m_i - \hat{m}_i) + (u_i - \hat{u}_i), \]
where
\[ \hat{y}_i = \frac{1}{n-1} \sum_{j \neq i} y_j L_{0,ij} / \hat{f}_i \]
and \( \hat{f}_i \) is the corresponding kernel estimator of \( f_i \) given by

\[
\hat{f}_i = \frac{1}{n-1} \sum_{j \neq i} L_{bi,ij},
\]

and \( \hat{m}_i \) and \( \hat{u}_i \) are defined in the same way as \( \hat{g}_i \) in (9) with \( y_j \) replaced by \( m_j \) and \( u_j \), respectively. Under \( H_0 \), \( \hat{e}_i \) is a good estimate of \( u_i \), while under \( H_1 \), \( \hat{e}_i \) includes a bias term which generates power. Now, \( \mathbb{E}[e|x,q]f(x,q) \) at \((x',q')^T \) is estimated by \( \frac{1}{n-1} \sum_{j \neq i} \hat{e}_j K_{h,ij}^F \) in the first test and \( \frac{1}{n-1} \sum_{j \neq i} \hat{e}_j K_{h,ij}^T \) in the second test. Hence, we may regard \( I_n^{(1)} \) and \( I_n^{(2)} \) as the sample analogs of \( \mathbb{E}[e|\mathbb{E}[e|x,q]f(x,q)] \) and \( \mathbb{E}[e|\mathbb{E}[e|x,q]f(x,q)1^T] \), respectively. 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 \( x \)), while 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 \( \theta \) and discuss approaches in constructing CIs for \( \gamma \). First, note that the 2SLS estimator of \( \gamma \) can be rewritten as a GMM estimator:

\[
\hat{\gamma} = \arg \min_{\gamma} \hat{Q}_n (\gamma),
\]

where

\[
\hat{Q}_n (\gamma) = \min_{\beta, \delta} \hat{Q}_n (\theta) := \min_{\theta} \hat{g}_n (\theta)' \hat{W} \hat{g}_n (\theta)
\]  

(10)

and

\[
\hat{g}_n (\theta) = \frac{1}{n} \sum_{i=1}^{n} g_i (\theta) = \frac{1}{n} \sum_{i=1}^{n} z_i (y_i - x_i' \beta - x_i' \gamma' \delta)
\]

with \( \hat{W} = \left( \frac{1}{n} Z'Z \right)^{-1} \). To study the asymptotic properties of \( \hat{\theta} \), we impose the following assumption.

**Assumption IV:** \( \mathbb{E}[z\varepsilon] = 0, \dim(z) = l \geq 2(d + 1) + 1 = 2d + 3 \),

\[
G = \left( \mathbb{E}[z'x'], \mathbb{E}[zz' \leq \gamma_0], \mathbb{E}[zz' | q = \gamma_0 \delta_0 \varepsilon_0 (\gamma_0) \right)
\]

is of full column rank, \( G_\gamma = \mathbb{E}[z (x', x' \leq \gamma)] =: (G_{1,2, \gamma}) \) is of full column rank for any \( \gamma \in \Gamma \), \( W := \mathbb{E}[zz'] \) > 0, and \( \Omega := \mathbb{E}[zz' \varepsilon^2] > 0 \). Also, there does not exist a vector \( a \in \mathbb{R}^{2(d + 1)} \) such that \( \mathbb{E}[z' \varepsilon] = 0 \) for any \( \gamma_0 \).

The full column rankness of \( G \) excludes continuous TR models where \( x' \delta_0 q = \gamma_0 \) is always zero so that the third part of \( G \) is a zero matrix.\(^9\) Anyway, this assumption is weaker than the full column rankness of \( \mathbb{E}[zz' | q = \gamma_0] \). This is because if \( \mathbb{E}[zz' | q = \gamma_0] \) is of full column rank, then 1 and \( q \) cannot be elements of \( x \) simultaneously; otherwise, the first and the last columns of \( \mathbb{E}[zz' | q = \gamma_0] \) are collinear. All other conditions in Assumption IV are standard except the last condition. This condition is required for the identification of \( \gamma_0 \). Note that if \( \mathbb{E}[z\varepsilon] = 0 \) then \( \mathbb{E}[zy] = G_{\gamma_0} \theta_0 \) with \( \theta = (\beta', \delta)' \). If there exists such an \( a := (a_1', a_2')' \) such that \( G_{\gamma} a = G_{2, \gamma_0} \delta_0 \) for some \( \gamma \neq \gamma_0 \), then under this \( \gamma \), we can still let \( \theta = (\beta_0 + a_1, a_2) \) to satisfy the moment conditions, i.e., the model is not identified by the moment conditions. This condition requires that

\[^9\]But zeroeness of \( x' \delta_0 q = \gamma_0 \) does not imply \( E[zz' \leq \gamma_0] \delta_0 = 0 \) or \( E[zz' > \gamma_0] \delta_0 = 0 \).
an \(l\)-dimensional vector \(G_{\gamma, \gamma_0} \delta_0 \notin \bigcup_{\gamma \neq \gamma_0} \text{span} \ (G_1, G_2, \gamma)\), where \(\text{span} \ (G_1, G_2, \gamma)\) is a \(2(d+1) < l\) dimensional space. There is an important case where this condition is violated. If \(q\) is independent of \((z', x')'\) as in the structural change model where \(q\) is the time index, then \(G_{\gamma, \gamma_0} = G_1 F_q (\gamma_0) \) and \(a = (F_q (\gamma_0) \delta_0, 0)'\) satisfy \(G_{\gamma, a} = G_{\gamma, \gamma_0} \delta_0\). In the TR context, if \(q\) is independent of the rest of the system, then \(q\) should be included in \(z\), and can not be independent of \((z', x')'\). This condition also implies the usual assumption that \(z\) cannot be independent of the endogenous variables \((x', q)'\)\(^{10}\). If this is the case, then \(G_{\gamma, \gamma_0} = \mathbb{E} [z] \mathbb{E} [x'_{\leq \gamma_0}]\) spans a one-dimensional space, which can obviously be spanned by \(G_{1} = \mathbb{E} [z] \mathbb{E} [x']\) and \(G_{2, \gamma} = \mathbb{E} [z] \mathbb{E} [x'_{\leq \gamma}]\).

**Theorem 1** Under Assumptions F, I, IV and \(S\), especially \(\tilde{\gamma}\), is consistent and

\[
\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N (0, V)
\]

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

Note that we need only Assumption \(I'\) to show the consistency of \(\tilde{\gamma}\), but to derive the asymptotic distribution, we need Assumption I; otherwise, \(G\) is not of full column rank. Also, as predicted in Section 2.1, \(f_q (\gamma_0)\) appears in \(V\), and \(\hat{\gamma}\) is not asymptotically independent of \((\hat{\beta}, \hat{\delta})\). \(\hat{W}\) can be any positive definite matrix, and we still use \(\hat{W}\) to denote such a matrix and use \(W\) to denote its limit. To get some intuition on the asymptotic variance of \(\hat{\gamma}\), consider the simple example in Section 2.1 again. In this example, \(G = f_q (\gamma_0), \ \Omega = \text{Var} (\varepsilon)\) and \(W\) is irrelevant, so \(V = \text{Var} (\varepsilon) / f_q (\gamma_0)^2\). In fact, \(\hat{\gamma} = \hat{\gamma}_q^{-1}(\tilde{y})\), so

\[
\sqrt{n}(\hat{\gamma} - \gamma_0) = \sqrt{n} \left[ \hat{F}_q^{-1}(\tilde{y}) - F_q^{-1}(\gamma_0) \right] + \sqrt{n} \left[ \hat{F}_q^{-1}(\mathbb{E} [y]) - F_q^{-1}(\mathbb{E} [y]) \right] + \sqrt{n} \left( F_q^{-1}(\gamma_0) - F_q^{-1}(\mathbb{E} [y]) \right).
\]

The first term is roughly \(\sqrt{n} \left( \sum_{i=1}^{n} (\psi (\tilde{y}) - \psi (\mathbb{E} [y])) \right)\) with \(\psi (\gamma) = \frac{\tau^{-1}(q \leq \gamma)}{f_q (\gamma_0)}\). By a stochastic equicontinuity result, it is \(o_p (1)\), so the asymptotic distribution of \(\sqrt{n}(\hat{\gamma} - \gamma_0)\) is the same as \(\sqrt{n} \left( \hat{F}_q^{-1}(\mathbb{E} [y]) - F_q^{-1}(\mathbb{E} [y]) \right) + \sqrt{n} \left( F_q^{-1}(\gamma_0) - F_q^{-1}(\mathbb{E} [y]) \right)\), where the first term represents the randomness in \(\hat{F}_q^{-1}\) and the second term represents the randomness in \(\tilde{y}\) (recall that \(\hat{\gamma} = \hat{\gamma}_q^{-1}(\tilde{y})\)). By the Bahadur representation, \(\sqrt{n} \left( \hat{F}_q^{-1}(\mathbb{E} [y]) - F_q^{-1}(\mathbb{E} [y]) \right) \approx \frac{E_q (\gamma_0) - 1(q \leq \gamma_0)}{f_q (\gamma_0)}\), and by the Delta method, \(\sqrt{n} \left( F_q^{-1}(\gamma_0) - F_q^{-1}(\mathbb{E} [y]) \right) \approx \frac{y - \mathbb{E} [y]}{f_q (\gamma_0)}\). So by the continuous mapping theorem, \(\sqrt{n}(\hat{\gamma} - \gamma_0) \approx \frac{E_q (\gamma_0) - 1(q \leq \gamma_0)}{f_q (\gamma_0)} + \frac{y - \mathbb{E} [y]}{f_q (\gamma_0)}\), and thus we have the asymptotic variance \(V = \text{Var} (\varepsilon) / f_q (\gamma_0)^2\).

By choosing \(\hat{W} = \Omega^{-1}\) with \(\hat{\Omega} = n^{-1} \sum_{i=1}^{n} z_i \xi_i^2\) and \(\xi_i = y_i - x_i' \beta - x_{i, \leq \gamma} \delta_i\), we get the efficient estimator of \(\theta_0\).

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

\[
\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N (0, (G'\Omega^{-1}G)^{-1})\]

When the model is homoskedastic, \(\text{Var} [z' x] = \sigma^2\), then our 2SLS estimator is efficient. As to the inference of \(\gamma\), we suggest to use the bootstrap methods such as in Hall and Horowitz (1996), Brown and Newey (2002) or Lee (2014) to avoid the estimation of \(\mathbb{E} [z x'] q = \gamma_0\) and \(f_q (\gamma_0)\) (e.g., by the numerical derivative in Section 7.3 of Newey and McFadden (1994) or some typical kernel or series estimators).

\(^{10}\)In the nonlinear scenario, uncorrelatedness in the linear scenario should be strengthened to independence. Also, all elements of \((x', q)'\) should be endogenous; otherwise, \(z\) should include the exogenous elements of \((x', q)'\) and cannot be independent of \((x', q)'\).
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:

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

where

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

with

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

As commented by HHB, the inconsistency stems from the fact that the minimand is a quadratic form in the sample moment, i.e., a square of sums. This "square of sums" structure allows the opportunity for the effects of the misspecification associated with the selection of the wrong threshold point to offset in the minimand. In contrast, the objective function of the 2SLS estimator in CH takes a "sum of squares" form, which generates a consistent estimator of $\gamma$. Specifically, the objective function of the 2SLS of CH is

$$\bar{S}_n (\theta) = \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \beta_1 (\tilde{W} \tilde{z}_i) 1(q_i \leq \gamma) - \beta_2 (\tilde{W} \tilde{z}_i) 1(q_i > \gamma) \right)^2,$$  \hspace{1cm} (11)

where $\tilde{W} \tilde{z}_i$ is the first-stage prediction of $x_i$.\[^{11}\] Conditional on the comments by HHB, it seems surprising that our GMM estimator is consistent even if the minimand is also a sum of squares. The key point here is not the distinction between "square of sums" and "sum of squares", but the fact that the threshold variable in the structural change model is the time index that is independent of the other components of the system. Also, whether "sum of squares" has more identification power than "square of sums" depends on specific contexts.

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

$$\tilde{W} = \begin{pmatrix} I_l \\ I_l \end{pmatrix} \bar{W} \begin{pmatrix} I_l & I_l \end{pmatrix},$$  \hspace{1cm} (12)

$\tilde{Q}_n (\theta) = \bar{Q}_n (\theta)$, where $I_l$ is an $l \times l$ identity matrix. This $\tilde{W}$ is only positive semidefinite, but not positive definite. In other words, our 2SLS estimator does not fully explore the information in $\tilde{m}_n (\theta)$. This is why we need $l > 2(d + 1)$ instruments, while HHB’s GMM estimator need only $l > (d + 1)$ instruments. This is also why our 2SLS estimator is not consistent when $q$ is like a time index (because the general GMM estimator is not consistent). The moment conditions in $\tilde{m}_n (\theta)$ explore the special structure of threshold regression - $\beta_1$ and $\beta_2$ are involved only in one regime of the system, while the moment conditions in $\tilde{g}_n (\theta)$ are designed for any nonlinear system $y = G (x, q; \theta) + \varepsilon$ with $\mathbb{E} [\varepsilon | x, q] \neq 0$. Essentially, the moment conditions $\tilde{m}_n (\theta)$ explores $\mathbb{E} [z \mathbb{I}_{z \leq \gamma_0}] = 0$ and $\mathbb{E} [z \mathbb{I}_{z > \gamma_0}] = 0$, while $\tilde{g}_n (\theta)$ explores only $\mathbb{E} [z \varepsilon] = 0$.

We now formally state the consistency of $\hat{\gamma}$ when $q$ is independent of $(z', x')'$. First, we impose the following assumption.

**Assumption IV**: $\dim (z) = l \geq (d + 1) + 1 = d + 2$, $\tilde{W} \overset{p}{\rightarrow} W > 0$\[^{12}\]$ and $\mathbb{E} [z x'_{\leq \gamma}]$ and $\mathbb{E} [z x'_{> \gamma}]$ are of

\[^{11}\]Following the general setup of this paper, we do not assume a threshold effect in the first stage.
\[^{12}\]To save notation, we still use $W$ to denote the limit of $\tilde{W}$. This should not introduce any confusion.
full column rank for any $\gamma \in \Gamma$. (i) If $q$ is exogenous (i.e., $q$ is included in $z$, and $E[z|x] = 0$), then there does not exist $a = (a'_1, a'_2)' \in \mathbb{R}^{(d+1)}$ such that $E[zz'_{\gamma}] a_2 = E[zz'_{\gamma_0}]\delta_0$ for any $\gamma < \gamma_0$ or $E[zz'_{\gamma}] a_1 = E[zz'_{\gamma_0}]\delta_0$ any $\gamma > \gamma_0$. (ii) If $q$ is endogenous, and satisfies only $E[ze_{\gamma}] = 0$ and $E[ze_{\gamma_0}] = 0$, then there does not exist $a = (a'_1, a'_2)' \in \mathbb{R}^{(d+1)}$ such that $E[zz'_{\gamma}] a_1 = E[ze_{\gamma}]$, $E[zz'_{\gamma_0}]\delta_0 + E[zz'_{\gamma}] a_2 = E[ze_{\gamma}]$ for any $\gamma < \gamma_0$ or $E[zz'_{\gamma}] a_1 - E[zz'_{\gamma_0}]\delta_0 = E[ze_{\gamma}]$ and $E[zz'_{\gamma}] a_2 = E[ze_{\gamma}]$ for any $\gamma > \gamma_0$.

**Theorem 2** Under Assumptions F, $\ell$, IV, and S, $\gamma$ is consistent.

HBB assume $\widetilde{W} = \text{diag}(\widetilde{W}_1, \widetilde{W}_2)$ 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 $\gamma$ or inconsistency of $\gamma$ in HBB’s setup. Similar to $\gamma$, we only require Assumption I’ rather than the stronger Assumption I to prove the consistency of $\gamma$. Different from Assumption IV, we need different assumptions for the identification of $\gamma_0$ depending on $q$ is exogenous or not. In this sense, reducing $m_n(\theta)$ to $\tilde{m}_n(\theta)$ makes the treatment of identification more uniform although loses some information. When $q$ is exogenous, Assumption IV’(i) requires some extra variation in $E[zz'_{q}] = \gamma$ when $\gamma$ moves away from $\gamma_0$. This condition implicitly precludes the possibility that $q$ is independent of $(z', x')'$ because if this is the case, then $E[zz'_{\gamma}] = E[zz'](1 - F_q(\gamma))$ and $E[zz'_{\gamma_0}] = E[zz'](1 - F_q(\gamma_0))$, so $a_2$ can be chosen as $\frac{1 - F_q(\gamma_0) - F_q(\gamma)}{1 - F_q(\gamma_0)}\delta_0$, and similarly, $a_1$ can be chosen as $\frac{F_q(\gamma) - F_q(\gamma_0)}{1 - F_q(\gamma_0)}\delta_0$. When $q$ is endogenous, we need also take into account of the variation in $E[ze_{\gamma}]$ and $E[ze_{\gamma}]$ as $\gamma$ moves away from $\gamma_0$. We will provide more intuitions on such identification information in the following discussion.

The proof of the theorem also shows the following results. First, if $q$ is indeed independent of $(z', x')'$ in case (i) or further assume $E[e|z, q] = 0$ in case (ii), then $\gamma_0$ cannot be identified by $\tilde{Q}_n(\gamma)$. This is essentially the case considered by HBB, and we label it case (o). Second, in case (i), both groups of moment conditions in $m_i(\theta)$ are required to identify $\gamma_0$; using only $m_1(\theta)$ or $m_2(\theta)$ are not enough. Third, in case (ii), either group of moment conditions in $m_i(\theta)$ can identify $\gamma_0$. For example, if there does not exist $a_1 \in \mathbb{R}^{d+1}$ such that $E[zz'_{\gamma}] a_1 = E[ze_{\gamma}]$ for any $\gamma < \gamma_0$ and $E[zz'_{\gamma}] a_1 - E[zz'_{\gamma_0}]\delta_0 = E[ze_{\gamma}]$ for any $\gamma > \gamma_0$, then $\gamma_0$ can be identified by only $m_i(\theta)$. By comparing with case (i), we can see that the identification power in either group of moment conditions in $m_i(\theta)$ comes solely from the correlation between $q$ and $e$. In other words, endogeneity is helpful in identifying $\gamma_0$ by moment conditions.

It seems that the correlation of $q$ with the rest of the system is critical for the identification of $\gamma_0$. When $q$ is independent of $(z', x', e)'$, then the combination of $m_1$ and $m_2$ cannot identify $\gamma_0$; if $q$ is independent of $e$ but not $(z', x')'$, then combination of $m_1$ and $m_2$ can (but $m_1$ or $m_2$ individually cannot) identify $\gamma_0$; if $q$ is correlated with all of $(z', x', e)'$, then either $m_1$ or $m_2$ can identify $\gamma_0$. What is the intuition here? We can understand these results using Lemma 2.3 of Newey and McFadden (1994) which states that as long as $W E[m_i(\theta)] \neq 0$ for $\theta \neq \theta_0$, then $\theta_0$ is identified. In case (o), for any $W$, $W E[m_i(\theta)] \neq 0$ for $\theta \neq \theta_0$ cannot hold. In case (i), when $W > 0$ or $W = \left( \begin{array}{c} I \hline I \end{array} \right) W_0 \left( \begin{array}{c} I \hline I \end{array} \right)$ for some $W_0 > 0$, $W E[m_i(\theta)] \neq 0$ for $\theta \neq \theta_0$.

---

13Assume $\widetilde{W} = \text{diag}(\widetilde{W}_1, \widetilde{W}_2)$. Because there is no restriction on $\widetilde{W}_1$ and $\widetilde{W}_2$ to obtain the consistency of $\gamma$ in both case (i) and case (ii), when $\widetilde{W}_1 > \widetilde{W}_2$, we are essentially using only $m_1(\theta)$ ($m_2(\theta)$) in $\tilde{Q}_n(\gamma)$. In this sense, it is surprising to see that case (i) requires both $m_1(\theta)$ and $m_2(\theta)$, while case (ii) requires only $m_1(\theta)$ or $m_2(\theta)$. Essentially, the limiting behaviors of $\tilde{Q}_n(\gamma)$ in these two cases are quite different; see the following discussion and example for more intuitions on this point. Anyway, note that either $\widetilde{W}_1 = 0$ or $\widetilde{W}_2 = 0$ violates $\widetilde{W} > 0$, so Assumption IV’ does not imply the identification of $\gamma_0$ 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., $\widetilde{W}$ in (12)).

14In cases (o) and (i), we require only $E[e|z, q] = 0$, and in case (ii), $q$ can be independent of $(z', x')'$. Here, we use three sequentially stronger assumptions to distinguish these three cases.
In case (ii), when \( W > 0 \) or \( W = \begin{pmatrix} I \ 
 I \end{pmatrix} W_0 \begin{pmatrix} I \ 
 I \end{pmatrix} \) for some \( W_0 > 0 \) or \( W = \begin{pmatrix} W_1 & 0 \\
 0 & 0 \end{pmatrix} \) with \( W_1 > 0 \) or \( W = \begin{pmatrix} 0 & 0 \\
 0 & W_2 \end{pmatrix} \) with \( W_2 > 0 \), \( \mathbb{W} \mathbb{E} \left[ z \begin{pmatrix} (y - x' \beta_1) 1(q \leq \gamma) \\
 (y - x' \beta_2) 1(q > \gamma) \end{pmatrix} \right] = 0 \) only if \( \gamma = \gamma_0 \) for any \( \beta_1 \) and \( \beta_2 \), or equivalently,

\[
\mathbb{W} \mathbb{E} \left[ \begin{bmatrix} z y_{\leq \gamma} \\
 z y_{> \gamma} \end{bmatrix} \right] = \mathbb{W} \mathbb{E} \left[ \begin{bmatrix} z x' \gamma_{\leq \gamma} \beta_1 \\
 z x' \gamma_{> \gamma} \beta_2 \end{bmatrix} \right] = 0
\]

when \( \gamma \neq \gamma_0 \) for any \( \beta_1 \) and \( \beta_2 \). Note that

\[
\mathbb{E} \left[ \begin{bmatrix} z y_{\leq \gamma} \\
 z y_{> \gamma} \end{bmatrix} \right] = \mathbb{E} \left[ \begin{bmatrix} z \left( x' \gamma_{\leq \gamma} \beta_{10} + x' \gamma_{0} \beta_{20} + \epsilon \right) 1(q \leq \gamma) \\
 z \left( x' \gamma_{\leq \gamma} \beta_{10} + x' \gamma_{0} \beta_{20} + \epsilon \right) 1(q > \gamma) \end{bmatrix} \right]
\]

which is equal to

\[
\mathbb{E} \left[ z x' \gamma_{\leq \gamma} \right] \beta_{10} + \mathbb{E} \left[ z \eps_{\leq \gamma} \right] + \mathbb{E} \left[ z x' \gamma_{> \gamma} \right] \beta_{20} + \mathbb{E} \left[ z \eps_{> \gamma} \right]
\]

when \( \gamma < \gamma_0 \), and equal to

\[
\mathbb{E} \left[ z x' \gamma_{\leq \gamma} \right] \beta_{10} + \mathbb{E} \left[ z x' \gamma_{< \gamma_{2}} \right] \beta_{20} + \mathbb{E} \left[ z \eps_{\leq \gamma} \right]
\]

when \( \gamma > \gamma_0 \). In case (i), \( \mathbb{E} \left[ z \eps_{\leq \gamma} \right] = \mathbb{E} \left[ z \eps_{> \gamma} \right] = 0 \), \( \mathbb{E} \left[ z x' \gamma_{\leq \gamma} \right] = \mathbb{E} \left[ z x' \gamma_{> \gamma} \right] F_q (\gamma) \), \( \mathbb{E} \left[ z x' \gamma_{> \gamma} \right] = \mathbb{E} \left[ z x' \gamma \right] (1 - F_q (\gamma)) \) and \( \mathbb{E} \left[ z x' \gamma_{< \gamma_{2}} \right] = \mathbb{E} \left[ z x' \gamma \right] (F_q (\gamma_{2}) - F_q (\gamma_{1})) \), where \( \gamma_{1} < \gamma_{2} \). So we can choose \( \beta_1 \) and \( \beta_2 \) such that

\[
\beta_1 = \beta_{10} \quad \text{and} \quad (1 - F_q (\gamma)) \beta_2 = (F_q (\gamma_{0}) - F_q (\gamma)) \beta_{10} + (1 - F_q (\gamma_{0})) \beta_{20}
\]

(13)

when \( \gamma < \gamma_0 \) and

\[
\beta_2 = \beta_{20} \quad \text{and} \quad F_q (\gamma) \beta_1 = F_q (\gamma_{0}) \beta_{10} + (F_q (\gamma) - F_q (\gamma_{0})) \beta_{20}
\]

(14)

when \( \gamma > \gamma_0 \) to make the equalities hold\(^{15}\). In other words, \( \text{plim} \bar{Q}_n (\gamma) = 0 \) for any \( \gamma \). In case (i), \( \mathbb{E} \left[ z \eps_{\leq \gamma} \right] = \mathbb{E} \left[ z \eps_{> \gamma} \right] = 0 \). So when \( \gamma < \gamma_0 \), we can choose \( \beta_1 = \beta_{10} \) to make \( \mathbb{E} \left[ z x' \gamma_{\leq \gamma} \right] \beta_1 = \mathbb{E} \left[ z y_{\leq \gamma} \right] \), but cannot choose \( \beta_2 \) such that \( \mathbb{E} \left[ z x' \gamma_{> \gamma} \right] \beta_2 = \mathbb{E} \left[ z y_{> \gamma} \right] \), and when \( \gamma > \gamma_0 \), we can choose \( \beta_2 = \beta_{20} \) to make \( \mathbb{E} \left[ z x' \gamma_{> \gamma} \right] \beta_2 = \mathbb{E} \left[ z y_{> \gamma} \right] \) but cannot choose \( \beta_1 \) such that \( \mathbb{E} \left[ z x' \gamma_{\leq \gamma} \right] \beta_1 = \mathbb{E} \left[ z y_{\leq \gamma} \right] \). In other words, if we use only \( m_1 \), then \( \text{plim} \bar{Q}_n (\gamma) = 0 \) for \( \gamma \in [\gamma_1, \gamma_2) \) and if we use only \( m_2 \), then \( \text{plim} \bar{Q}_n (\gamma) = 0 \) on \( [\gamma_0, \gamma_1] \), while if we use both \( m_1 \) and \( m_2 \), then \( \text{plim} \bar{Q}_n (\gamma) = 0 \) only if \( \gamma = \gamma_0 \). In case (ii), \( \mathbb{E} \left[ z \eps_{\leq \gamma} \right] \neq 0 \) and \( \mathbb{E} \left[ z \eps_{> \gamma} \right] \neq 0 \). 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 HHB’s GMM and CH’s 2SLS. In CH,

\[
\text{plim} \hat{S}_n (\theta) = \mathbb{E} \left[ \left( y - \beta' \left( \Pi' z \right) 1(q \leq \gamma) - \beta'_2 (\Pi' z) 1(q > \gamma) \right)^2 \right]
\]

\(^{15}\text{Note that we can choose } \beta_1 \text{ and } \beta_2 \text{ freely, so the choice of } \beta_1 \text{ and } \beta_2 \text{ depends on } \gamma.\)
so we are assume \( \text{E} [y | z, q] = \beta_1' (\Pi' z) 1(q \leq \gamma) + \beta_2' (\Pi' z) 1(q > \gamma) \) and use the conditional mean difference of \( y \) below \( \gamma_0 \) and above \( \gamma_0 \) to identify \( \gamma_0 \) (just as in the least squares estimation where \( \text{E}[\varepsilon | x] = 0 \)). Since 

\[
y = \beta_1' (\Pi' z + u) 1(q \leq \gamma) + \beta_2' (\Pi' z + u) 1(q > \gamma) + \varepsilon,
\]

where the first stage regression is assumed to be \( x = \Pi' z + u \), we must assume \( \text{E} [u | z, q] = 0 \) and \( \text{E} [\varepsilon | z, q] = 0 \) to make the conditional mean of \( y \) be 

\[
\beta_1' (\Pi' z) 1(q \leq \gamma) + \beta_2' (\Pi' z) 1(q > \gamma).
\]

To achieve such conditions, we must assume \( q \) is exogenous and so can be included in \( z \). Also, as argued in Yu (2013a), the first stage must be a regression rather than a projection, i.e., \( \text{E} [u | z] = 0 \) rather than \( \text{E} [zu] = 0 \). In a nonlinear environment, such a requirement seems not too stringent. On the contrary, the identification of \( \gamma_0 \) by HHB’s GMM is based on the matching of covariances just as in the usual linear GMM estimation. If \( \gamma_0 \) were known, we can identify \( \beta_1 \) by matching \( \text{E} [z y \leq \gamma_0] \) with \( \text{E} [z x \leq \gamma_0] \beta_1 \) and \( \beta_2 \) by matching \( \text{E} [z y > \gamma_0] \) with \( \text{E} [z x > \gamma_0] \beta_2 \). It is the nonlinear structure introduced by the unknown \( \gamma \) that makes the identification be divided into three different cases; in such a nonlinear system, the endogeneity of \( q \) is helpful rather than harmful as in CH’s 2SLS. Also, because the identification scheme of CH is different from that of HHB’s GMM, CH’s 2SLS requires only \( l \geq (d + 1) \) instruments, while HHB’s GMM requires at least \( d + 2 \) instruments. As to inference, the bootstrap is questionable for CH’s 2SLS given the negative results of Yu (2014).

The following example illustrates the identification results above intuitively.

**Example 1** Consider the simple example in Section 2.1. In this example, \( y = 1 (q \leq \gamma_0) + \varepsilon \); assume \( q \sim U [0, 1] \), and \( \gamma_0 = 1 / 2 \). \( \beta_0 = 0 \) and \( \delta_0 = 1 \) are known, \( x = 1 \), and \( \text{Var} (\varepsilon) = 1 \).

First assume \( \text{E} [\varepsilon | q] = 0 \); that is, there is actually no endogeneity. Let \( z = 1 \); then this is case (ii) - \( \text{E} [\varepsilon | q, z] = 0 \) and \( q \perp (z, x) \). The moment conditions used for identifying \( \gamma_0 \) are

\[
\text{E} \left( \begin{array}{c}
(y - 1) 1(q \leq \gamma) \\
y 1(q > \gamma)
\end{array} \right) = 0.
\]

(15)

Suppose \( \tilde{W} = I_2 \); then

\[
\text{plim} \tilde{Q}_n (\gamma) = \text{E} [(y - 1) 1(q \leq \gamma)]^2 + \text{E} [y 1(q > \gamma)]^2
\]

By some algebra,

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

where for \( a \in \mathbb{R}, a_+ = \max(a, 0) \). Obviously, \( \arg \min_{\gamma \in \mathbb{R}} \text{plim} \tilde{Q}_n (\gamma) = 1 / 2 \). This seems to contradict the nonidentification result of HHB. In fact, this is because \( \beta_0 \) and \( \delta_0 \) are known. In (13) and (14), \( \beta_1 \) and \( \beta_2 \) 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 \( [\gamma, \gamma_0] \). 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 \( [\gamma_0, \gamma] \). If they can be chosen freely, then it is obvious that the system cannot be identified - two equations and three unknowns.

Next, let \( z = (1, q)' \); then this is case (i) and the moment conditions used for identifying \( \gamma_0 \) are

\[
\text{E} \left( \begin{array}{c}
z (y - 1) 1(q \leq \gamma) \\
z y 1(q > \gamma)
\end{array} \right) = 0.
\]

\[\text{I} \geq (d + 1) \] is implied by CH’s assumption that \( E[\Pi' zz' | q = \gamma_0] = \Pi' E[zz' | q = \gamma_0] \Pi > 0.\]
Suppose \( \tilde{W} = \left( \begin{array}{cc} \tilde{W}_1 & \tilde{W}_{12} \\ \tilde{W}_{12}' & \tilde{W}_2 \end{array} \right) = \left( \begin{array}{cc} c_1 I_2 & \tilde{W}_{12} \\ \tilde{W}_{12}' & c_2 I_2 \end{array} \right) \); then

\[
\text{plim} \hat{Q}_n (\gamma) = c_1 \left[ - \left( \gamma - \frac{1}{2} \right)_+ \right]^2 + c_1 \left[ - \left( \frac{1}{2} \gamma^2 - \frac{1}{8} \right)_+ \right] + c_2 \left( \frac{1}{2} - \gamma \right)_+^2 + c_2 \left( \frac{1}{8} - \frac{1}{2} \gamma^2 \right)_+^2 + 2 \left( - \left( \gamma - \frac{1}{2} \right)_+ \right)^2 - \left( \frac{1}{2} \gamma^2 - \frac{1}{8} \right)_+ \right) \tilde{W}_{12} \left( \left( \frac{1}{2} - \gamma \right)_+ \left( \frac{1}{8} - \frac{1}{2} \gamma^2 \right)_+ \right).
\]

If \( \tilde{W}_{12} = 0 \), \( c_1 = 1 \) and \( c_2 = 0 \), then we use only \( m_1 (\gamma) \); Figure 1 shows that \( \arg \min_{\gamma \in \Gamma} \text{plim} \hat{Q}_n (\gamma) = [0, 1/2] \).
If \( \tilde{W}_{12} = 0 \), \( c_1 = 0 \) and \( c_2 = 1 \), then we use only \( m_2 (\gamma) \); Figure 1 shows that \( \arg \min_{\gamma \in \Gamma} \text{plim} \hat{Q}_n (\gamma) = [1/2, 1] \) \(^{17}\)
If \( c_1 \neq 0 \) and \( c_2 \neq 0 \), then we use both \( m_1 (\gamma) \) and \( m_2 (\gamma) \); Figure 1 shows that when either \( c_1 = 1 \), \( c_2 = 2 \) and \( \tilde{W}_{12} = 0 \) or \( c_1 = 1 \), \( c_2 = 1.5 \) and \( \tilde{W}_{12} = \left( \begin{array}{cc} 0.5 & 0.5 \\ 0.5 & 0.5 \end{array} \right) \), \( \arg \min_{\gamma \in \Gamma} \text{plim} \hat{Q}_n (\gamma) = 1/2 \) \(^{18}\)

Section 3.3 of Yu (2015b) considers a joint distribution of \( q \) and \( \varepsilon \) as follows:

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

where \( \phi(\cdot) \) is the density of the standard normal distribution. Obviously, \( E[\varepsilon | q] \neq 0 \), and \( E[\varepsilon | 1(q \leq \gamma_0)] = E[\varepsilon | 1(q > \gamma_0)] = 0 \), so this is case (ii). Suppose \( z = 1 \); then the moment conditions used for identifying \( \gamma_0 \) are \(^{15}\). Suppose \( \tilde{W} = \left( \begin{array}{cc} A & C \\ C & B \end{array} \right) \); then

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

By some algebra,

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

Figure 2 shows that when \( A = 1 \), \( B = 2 \) and \( C = 0 \), \( \arg \min_{\gamma \in \Gamma} \text{plim} \hat{Q}_n (\gamma) = 1/2 \). Actually, if only \( m_1 \) is used (i.e., \( A = 1 \), \( B = 0 \) and \( C = 0 \)) or only \( m_2 \) is used (i.e., \( A = 0 \), \( B = 1 \) and \( C = 0 \)), \( \arg \min_{\gamma \in \Gamma} \text{plim} \hat{Q}_n (\gamma) = 1/2 \), where \( \Gamma \) excludes the neighborhoods of 0 and 1. If \( C \neq 0 \), \( \arg \min_{\gamma \in \Gamma} \text{plim} \hat{Q}_n (\gamma) = \gamma_0 \) as long as \( \tilde{W} > 0 \); e.g., Figure 2 shows that when \( A = 1 \), \( B = 2 \) and \( C = 1 \), \( \arg \min_{\gamma \in \Gamma} \text{plim} \hat{Q}_n (\gamma) = 1/2 \).

We now check the 2SLS estimators of this paper and CH in these cases - (i) \( \mathbb{E}[\varepsilon | q] = 0 \) and \( z = 1 \); (i) \( \mathbb{E}[\varepsilon | q] = 0 \) and \( z = (1, q)' \); and (ii) \( \mathbb{E}[\varepsilon | q] \neq 0 \) and \( z = 1 \). For our 2SLS, the moment conditions are

\(^{17}\)This implies that in case (o) (i.e., \( z = 1 \)), using only one moment conditions cannot identify \( \gamma_0 \).

\(^{18}\)In this example, \( \tilde{W}_{12} \) does not play any role, i.e., \( \text{plim} \hat{Q}_n (\gamma) \) depends only on \( \tilde{W}_1 \) and \( \tilde{W}_2 \).
$\mathbb{E}[y - 1(q \leq \gamma)] = 0$ in cases (o) and (ii) and $\mathbb{E}[\mathbf{z}(y - 1(q \leq \gamma))] = \mathbf{0}$ in case (i). In case (o),

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

where $\frac{1}{2} - \gamma = - (\gamma - \frac{1}{2})_+ + (\frac{1}{2} - \gamma)_+$. In case (i) with $\hat{W} = I_2$,

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

where $\frac{1}{8} - \frac{1}{2} \gamma^2 = - \left( \frac{1}{2} \gamma^2 - \frac{1}{8} \right)_+ + \left( \frac{1}{2} - \frac{1}{4} \gamma^2 \right)_+$. In case (ii),

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

same as in case (o). For CH’s 2SLS,

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

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

$$\text{plim} \hat{\mathcal{S}}_n (\gamma) = 1 + |\gamma - 1/2|.$$

In case (ii),

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

Figure 3 shows $\text{plim} \tilde{\mathcal{Q}}_n (\gamma)$ and $\text{plim} \hat{\mathcal{S}}_n (\gamma)$ in these three cases. For our 2SLS, $\arg \min_{\gamma \in \Gamma} \text{plim} \tilde{\mathcal{Q}}_n (\gamma) = 1/2$ in all cases, the same identification results as HHB’s GMM using both $m_1$ and $m_2$. For CH’s 2SLS, $\arg \min_{\gamma \in \Gamma} \text{plim} \hat{\mathcal{S}}_n (\gamma) = 1/2$ in cases (o) and (i), and $\arg \min_{\gamma \in \Gamma} \text{plim} \hat{\mathcal{S}}_n (\gamma) = [1/4, 3/4]$ in case (ii) - un-identified!

We will not conduct inference on $\gamma$ based on $\gamma$. This is because $d\mathbb{E} [m_i (\theta_0)] / d\theta'$ does not exist as required in the usual GMM asymptotic distribution derivation. Especially,

$$\frac{\partial \mathbb{E} [m_i (\beta_0, \delta_0, \gamma)]}{\partial \gamma} \bigg|_{\gamma = \gamma_0^+} = \begin{pmatrix} \mathbb{E} [\mathbf{z}\epsilon | q = \gamma_0] f_q (\gamma_0) - \mathbb{E} [\mathbf{z}\mathbf{x} | q = \gamma_0] \delta_0 f_q (\gamma_0) \\ -\mathbb{E} [\mathbf{z}\epsilon | q = \gamma_0] f_q (\gamma_0) \end{pmatrix}$$

while

$$\frac{\partial \mathbb{E} [m_i (\beta_0, \delta_0, \gamma)]}{\partial \gamma} \bigg|_{\gamma = \gamma_0^-} = \begin{pmatrix} \mathbb{E} [\mathbf{z}\epsilon | q = \gamma_0] f_q (\gamma_0) \\ -\mathbb{E} [\mathbf{z}\epsilon | q = \gamma_0] f_q (\gamma_0) - \mathbb{E} [\mathbf{z}\mathbf{x} | q = \gamma_0] \delta_0 f_q (\gamma_0) \end{pmatrix}$$

Here, note that $\mathbb{E} [\mathbf{z}\epsilon 1(q \leq \gamma_0)] = \mathbf{0}$ and $\mathbb{E} [\mathbf{z}\epsilon 1(q > \gamma_0)] = \mathbf{0}$ does not imply $\mathbb{E} [\mathbf{z}\epsilon | q = \gamma_0] = \mathbf{0}$. Even if $\mathbb{E} [\mathbf{z}\epsilon | q = \gamma_0] = \mathbf{0}$ as in case (i), if $\mathbb{E} [\mathbf{z}\mathbf{x} | q = \gamma_0]$ is of full column rank, $\frac{\partial \mathbb{E} [m_i (\beta_0, \delta_0, \gamma)]}{\partial \gamma} \bigg|_{\gamma = \gamma_0}$ does not exist; e.g., in
Figure 1: $\text{plim} \bar{Q}_n(\gamma)$ When $E[\varepsilon|q] = 0$ and $z = (1, q)'$

Figure 2: $\text{plim} \bar{Q}_n(\gamma)$ When $E[\varepsilon|q] \neq 0$
Figure 3: $\text{plim} Q_n(\gamma)$ and $\text{plim} S_n(\gamma)$

the above Example 1, when $E[\varepsilon|q] = 0$ and $z = (1, q)'$,

$$
E[m_i(\beta_0, \delta_0, \gamma)] = \begin{pmatrix}
-\left(\frac{1}{2} - \gamma\right)_+ \\
-\left(\frac{1}{2}\gamma^2 - \frac{1}{3}\right)_+ \\
\left(\frac{1}{2} - \gamma\right)_+ \\
\left(\frac{1}{8} - \frac{1}{2}\gamma^2\right)_+
\end{pmatrix}
$$

(17)

is not differentiable at $\gamma_0 = 1/2$.\footnote{In the above example, when $E[\varepsilon|q] \neq 0$, $E[m_i(\beta_0, \delta_0, \gamma)]$ is differentiable at $\gamma_0$. This is due to the special design of the data generating process. Specifically, $E[\varepsilon \mid q = \gamma_0^+] = E[\varepsilon \mid q = \gamma_0^-] = 0$, and $E[\varepsilon \mid q = \gamma_0^+] = -1/2 = 1/2 - 1 = -E[\varepsilon \mid q = \gamma_0^-] - E[\varepsilon \mid q = \gamma_0^-]$.}

This makes the inference based on $\tilde{\gamma}$ difficult. We will discuss the asymptotic properties of $\tilde{\gamma}$ and the bootstrap inference based on $\tilde{\gamma}$ in a separate paper. On the contrary, in our 2SLS estimator,

$$
\frac{\partial E[g_i(\beta_0, \delta_0, \gamma)]}{\partial \gamma} \bigg|_{\gamma=\gamma_0^+} = (I_1, I_1) \frac{\partial E[m_i(\beta_0, \delta_0, \gamma)]}{\partial \gamma} \bigg|_{\gamma=\gamma_0^+} = -E[zx|q = \gamma_0] \delta_0 f_q(\gamma_0)
$$

is differentiable at $\gamma_0 = 1/2$.

Finally, we show that "sum of squares" need not have more identification power than "square of sums". 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 above Example 1 shows that the limit objective function of CH’s 2SLS need not even have a unique minimizer. On the contrary, using either the 2SLS of this paper or the GMM of HHB can generate a consistent estimator of $\gamma_0$.

The following Table 1 summarizes the identification results for all possibly consistent estimators of $\gamma_0$.\footnote{In the above example, when $E[\varepsilon|q] \neq 0$, $E[m_i(\beta_0, \delta_0, \gamma)]$ is differentiable at $\gamma_0$. This is due to the special design of the data generating process. Specifically, $E[\varepsilon \mid q = \gamma_0^+] = E[\varepsilon \mid q = \gamma_0^-] = 0$, and $E[\varepsilon \mid q = \gamma_0^+] = -1/2 = 1/2 - 1 = -E[\varepsilon \mid q = \gamma_0^-] - E[\varepsilon \mid q = \gamma_0^-]$.}
in various scenarios. The first three estimators need instruments and the last two do not. Among the last two, Perron and Yamamoto (2015) (PY in Table 1) use the least squares estimator (LSE) to estimate $\gamma$ in the structural change model even if there is endogeneity. However, as shown in Yu (2015), this strategy is only valid in the structural change model. From Table 1, it seems that the IDKE of YP has the most extensive identification power although it involves no instruments. Nevertheless, the IDKE cannot identify $\gamma_0$ in continuous TR models while HHB’s GMM estimator and our 2SLS estimator can identify $\gamma_0$ even in such models (although inferences need further investigation). Table 1 also presents some interesting differences between structural change models and TR models (case (o) vs. cases (i) and (ii)). However, as a folklore result, these two kinds of models are often considered to be similar to each other (in the asymptotic properties).

|                | (o): $\mathbb{E}[\varepsilon|z,q] = 0$ and $q \perp (z',x')$ | (i): $\mathbb{E}[\varepsilon|z,q] = 0$ but $q \not\perp (z',x')$ | (ii): $\mathbb{E}[\varepsilon|z,q] \neq 0$ but $\mathbb{E}[z_q \leq \gamma_0] = \mathbb{E}[z_q > \gamma_0] = 0$ |
|----------------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|
|                | Consistency | Literature | Consistency | Literature | Consistency | Literature |
| GMM of HHB     | No          | HHB        | Yes         | This Paper | Yes         | This Paper |
| 2SLS of This Paper | No        | This Paper | Yes         | This Paper | Yes         | This Paper |
| 2SLS of CH     | Yes         | HHE$^{22}$ | Yes         | CH         | No          | Yu (2013a) |
| LSE of PY and Yu (2015a)$^{26}$ | Yes     | PY and Yu (2015a)$^{27}$ | No         | Yu (2015a) | No         | Yu (2015a) |
| IDKE of YP     | Yes         | YP         | Yes         | YP         | Yes         | YP         |

Table 1: Identification of $\gamma_0$ by Various Possibly Valid Estimators in Different Scenarios

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

In this section, we present the limit theory for the IDKE $\hat{\gamma}$ in Method II where $k_{\pm}(0) = 0$. To facilitate the expression for the limit distribution of $\hat{\gamma}$, we define the following quantities,

$$
\Delta_i = \mathbb{E}[y_i|x_i,q_i = \gamma_0] - \mathbb{E}[y_i|x_i,q_i = \gamma_0^+] =: m_-(x_i) - m_+(x_i),
$$

$$
\Delta_f(x_i) = \Delta_i \cdot f(x_i,\gamma_0),
$$

where $\Delta_f(x_i)$ is the limit of $\hat{\Delta}_i(\gamma_0)$. Note that if $\mathbb{E}[\varepsilon|x,q]$ is not a continuous function, $\Delta_i$ need not be $(1,x',\gamma_0)\delta_0$. The asymptotic distribution of $\hat{\gamma}$ depends only on $\Delta_i$. To derive the asymptotic distribution of $\hat{\gamma}$, we impose the following assumptions on $f(u|x,q)$ which is allowed to be discontinuous at $q = \gamma_0$.

**Assumption U:**

$^{20}$We do not cover Kourtellos et al. (2016)’s estimator because as shown in Liao et al. (2017), it is not generically consistent.

$^{21}$In such models, the IDKE can be extended to take into account of also slope differences at each $\gamma \in \Gamma$ beyond level differences to identify $\gamma_0$.

$^{22}$Here, we implicitly assume $z$ and $x$ do not include $q$.

$^{23}$Both $m_1(\theta)$ and $m_2(\theta)$ are required to prove the consistency.

$^{24}$Either $m_1(\theta)$ or $m_2(\theta)$ is enough to prove the consistency.

$^{25}$Yu (2015a) strengthens this result a little bit. Specifically, let $z = (z',q)'$ and $x = (x,q)'$. If $q \perp z$ and $E[x|z,q] = g(z) + q\lambda$, i.e., $q$ need not be independent of $x$, then projecting $x$ only on $z$ in the first stage would generate a consistent estimator of $\gamma_0$.

$^{26}$$z$ is not necessary here.

$^{27}$Yu (2015a)’s result is a little bit stronger. Specifically, if $q \perp x$ and $E[\varepsilon|x,q] = g(x) + q\lambda$, i.e., $q$ need not be exogenous, then the LSE is consistent.
(a) \( f(u|x, q) \) is continuous in \( u \) for \((x', q') \in \mathcal{X} \times \Gamma \) and \((x', q') \in \mathcal{X} \times \Gamma^+ \), where \( \Gamma^+ = (\gamma - \epsilon, \gamma_0] \) and \( \Gamma^- = (\gamma_0, \gamma + \epsilon) \) for some \( \epsilon > 0 \).
(b) \( f(u|x, q) \) is Lipschitz in \((x', q') \) for \((x', q') \in \mathcal{X} \times \Gamma^\epsilon \) and \((x', q') \in \mathcal{X} \times \Gamma^+ \).
(c) \( E[u^2|x, q] \) is uniformly bounded on \((x', q') \in \mathcal{X} \times \Gamma^\epsilon \), where \( \Gamma^\epsilon = \Gamma^- \cup \Gamma^+ \).

Given Assumption U, we impose conditions on the bandwidth \( h \) below.

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

Observe that \( nh^d = \sqrt{\ln n \frac{\sqrt{nh^d}}{\ln n}} \to \infty \) when \( \sqrt{nh^d}/\ln n \to \infty \). The limit distribution of \( \hat{\gamma} \) is given in the following theorem.

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

\[
\sqrt{n/h}(\hat{\gamma} - \gamma_0) \xrightarrow{d} N(0, \Sigma)
\]

where

\[
\Sigma = \frac{\mathbb{E}[\Delta^2(x_i) f^2(x_i) \sigma^2(x_i) + \sigma^2(x_i)] q_i = \gamma_0}{f_q(\gamma_0) (\mathbb{E}[\Delta f(x_i) \Delta_i f(x_i)] q_i = \gamma_0)^2} k_+(0)^2
\]

with \( \xi_{(1)} = \int_0^1 k_+(t)^2 dt \) and \( \sigma^2(x) = \mathbb{E} [u^2|x, q = \gamma_0] \).

From Theorem 3 \( \hat{\gamma} \) converges to \( \gamma_0 \) at the rate of \( \sqrt{n/h} \); this rate is much faster than the DKE \( \tilde{\gamma} \) of DH because \( \hat{\gamma} \) uses more data information in the estimation. Specifically, the convergence rate of DKE is \( \sqrt{n/h^d-2} \) and the relative rate \( \sqrt{nh^d-2}/\sqrt{n/h} = h^{d-1} \to 0 \). Based on Theorem 2 of DH, the asymptotic variance of their estimator \( \hat{\gamma} \) is

\[
\Sigma_o = \frac{\sigma^2(x_o) + \sigma^2(x_o) \kappa^2 \xi_{(1)}}{f(x_o, \gamma_0) \Delta^2 k_+(0)^2}, \tag{18}
\]

where \( \kappa^2 = f K(u_x)^2 du_x \) with \( K(u_x) = \prod_{i=1}^{d-1} k(u_x) \), and \( \Delta_o = (1, x'_o, \gamma_0) \delta_0 \). This asymptotic variance is comparable to \( \Sigma \), but critically relies on the choice of \( x_o \). If \( \Delta = \Delta \) and \( \sigma^2(x) = \sigma^2 \), then \( \Sigma = O \left( \frac{\sigma^2(x_o) \Delta^2 k_+(0)^2}{f(x_o, \gamma_0) \Delta^2 k_+(0)^2} \right) \). As expected, \( \Sigma \) is decreasing in \( f_q(\gamma_0) \), \( |\Delta| \) and \( k_+(0) \) and increasing in \( \sigma^2 \).

The convergence rate of \( \hat{\gamma} \) is even faster than the parametric rate \( \sqrt{n} \). To understand this convergence rate, assume \( d = 1 \) for simplification, i.e., \( q \) is the only covariate. Then the convergence rate is determined by the balance between an empirical process and a deterministic process. Recall that

\[
\hat{\gamma} = \arg \max_{\gamma \in \Gamma} \hat{Q}_n(\gamma) = \arg \max_{\gamma \in \Gamma} \left\{ \hat{Q}_n(\gamma) - \hat{Q}_n(\gamma_0) \right\}.
\]

Then because \( \hat{\gamma} \) is the maximizer of \( \hat{Q}_n(\gamma) - \hat{Q}_n(\gamma_0) \) on \( \Gamma \), and \( \gamma_0 \in \Gamma \),

\[
0 \leq \hat{Q}_n(\gamma) - \hat{Q}_n(\gamma_0) = [Q_0(\gamma) - Q_0(\gamma_0)] + \left[ (\hat{Q}_n(\gamma) - \hat{Q}_n(\gamma)) - (\hat{Q}_n(\gamma_0) - \hat{Q}_n(\gamma_0)) \right],
\]

where the first term on the right-hand side is the limit process and is less than zero because \( \gamma_0 = \arg \max_{\gamma \in \Gamma} Q_0(\gamma) \), and the second term is the modulus of continuity of the empirical process and need be greater than zero. We must balance \( Q_0(\gamma) - Q_0(\gamma_0) \) and

\[
\sup_{|\gamma - \gamma_0| \leq \delta} \left[ (\hat{Q}_n(\gamma) - \hat{Q}_n(\gamma)) - (\hat{Q}_n(\gamma_0) - \hat{Q}_n(\gamma_0)) \right] =: \frac{\phi_n(\delta)}{\sqrt{n}}
\]
such that their sum is greater than zero.

In the $h$ neighborhood of $\gamma_0$, 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 $q_i = i/n$, $i = 1, \ldots, n$ and $u_i \sim N(0,1)$ and $\Delta > 0$. Now, \( \hat{\gamma} \) tries to maximize $\hat{\Delta}(\gamma) - \Delta(\gamma_0)$. For $|\gamma - \gamma_0| \leq \delta$, and $\gamma < \gamma_0$,

$$\hat{\Delta}(\gamma) - \Delta(\gamma_0) = \left[ \frac{1}{nh} \sum_{i=n(\gamma-h)}^{n\gamma} k_+ \left( \frac{i - n\gamma}{nh} \right) y_i - \frac{1}{nh} \sum_{i=n(\gamma-h)+1}^{(n+1)\gamma} k_- \left( \frac{i - n\gamma}{nh} \right) y_i \right] - \left[ \frac{1}{nh} \sum_{i=n(\gamma_0-h)}^{n\gamma_0} k_- \left( \frac{i - n\gamma_0}{nh} \right) y_i - \frac{1}{nh} \sum_{i=n\gamma_0+1}^{n(\gamma_0+h)} k_+ \left( \frac{i - n\gamma_0}{nh} \right) y_i \right]$$

$$= \Delta + \frac{1}{nh} \sum_{i=n(\gamma-h)}^{n\gamma} k_- \left( \frac{i - n\gamma}{nh} \right) u_i - \frac{1}{nh} \sum_{i=n\gamma_0+1}^{n(\gamma_0+h)} k_+ \left( \frac{i - n\gamma_0}{nh} \right) u_i$$

$$- \left[ \frac{1}{nh} \sum_{i=n(\gamma-h)}^{n\gamma_0} k_- \left( \frac{i - n\gamma_0}{nh} \right) u_i - \frac{1}{nh} \sum_{i=n\gamma_0+1}^{n(\gamma_0+h)} k_+ \left( \frac{i - n\gamma_0}{nh} \right) u_i \right]$$

$$\approx -\Delta \left( \frac{1}{1/nh} \sum_{i=n\gamma}^{n\gamma_0} k_+ \left( \frac{i - n\gamma}{nh} \right) + \frac{1}{1/nh} \sum_{i=n(\gamma-h)}^{n\gamma} \left[ k_- \left( \frac{i - n\gamma}{nh} \right) - k_- \left( \frac{i - n\gamma_0}{nh} \right) \right] \right) u_i + \frac{1}{1/nh} \sum_{i=n\gamma_0+1}^{n(\gamma_0+h)} k_+ \left( \frac{i - n\gamma_0}{nh} \right) + k_+ \left( \frac{i - n\gamma_0}{nh} \right) u_i$$

$$- \frac{1}{1/nh} \sum_{i=n\gamma_0+1}^{n(\gamma_0+h)} \left[ k_+ \left( \frac{i - n\gamma_0}{nh} \right) + k_+ \left( \frac{i - n\gamma_0}{nh} \right) \right] u_i$$

$$= -O \left( \Delta \int_0^{\gamma_0} k_+ (v) dv \right) + O_p \left( \int_0^{\gamma_0} \left( \frac{\gamma_0 - \gamma}{h} \right)^2 dv \right) + \int_0^{\gamma_0} \left( \frac{\gamma_0 - \gamma}{h} \right)^2 dv$$

If $\Delta$ is fixed, $k_+(0) = 0$ and $k_+'(0) > 0$, then $\int_0^{\gamma_0} k_+(v) dv = O \left( \left( \frac{\gamma_0 - \gamma}{h} \right)^2 \right)$ and $\int_0^{\gamma_0} \left( \frac{\gamma_0 - \gamma}{h} \right)^2 dv = \int_0^{\gamma_0} \left( \frac{\gamma_0 - \gamma}{h} \right)^2 dv = (\frac{\gamma_0 - \gamma}{h})^2$. As a result, $Q_0(\gamma) - Q_0(\gamma_0) = O (\delta^2/h^2)$ and $\phi_n (\delta) = \sqrt{\delta^2/h^2}$ since $\left( \frac{\gamma_0 - \gamma}{h} \right)^3 = \alpha \left( \frac{\gamma_0 - \gamma}{h} \right)^2$.

If $\Delta$ is fixed, $k_+(0) = 0$ and $k_+'(0) > 0$, then $\int_0^{\gamma_0} k_+(v) dv = O \left( \left( \frac{\gamma_0 - \gamma}{h} \right)^2 \right)$ and $\int_0^{\gamma_0} \left( \frac{\gamma_0 - \gamma}{h} \right)^2 dv = \int_0^{\gamma_0} \left( \frac{\gamma_0 - \gamma}{h} \right)^2 dv = (\frac{\gamma_0 - \gamma}{h})^2$. As a result, $Q_0(\gamma) - Q_0(\gamma_0) = O (\delta^2/h^2)$ and $\phi_n (\delta) = \sqrt{\delta^2/h^2}$. Solving $\frac{1}{nh} = \frac{1}{\sqrt{n}h^3}$, we get $r_n = n$. In the latter (as will be detailed in the next section), $\Delta \to 0$, so $Q_0(\gamma) - Q_0(\gamma_0) = O (\Delta^2/h)$ and $\phi_n (\delta) = \sqrt{\delta^2/h^2}$. Solving $\frac{1}{nh} = \frac{1}{\sqrt{n}h^3}$, we get $r_n = n$. Figure 4 illustrates the intuition above. For example, in Method II, because $Q_0(\gamma) - Q_0(\gamma_0)$ is quadratic in the neighborhood of $\gamma_0$ as in the regular parameter case, we expect $\hat{\gamma}$ to have an asymptotically normal distribution. The extra $h$ in the convergence rate $\sqrt{n/h}$ is 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 $\gamma$ (in the neighborhood of $\gamma_0$) such that $\gamma$ can be more easily identified than in Method II, so 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 $\gamma_0$ are $O (\Delta^2/h)$, less than $O (1/h)$ in YP), so the convergence rate is slower than in YP. In YP and Method III, the convergence rates of $\hat{\gamma}$ are actually the same as in the parametric cases. Note also that in each case, if set $h = 1$, then we get...
Figure 4: Balancing $Q_0(\gamma) - Q_0(\gamma_0)$ and $\phi_{h}(\delta) / \sqrt{n}$ in Method II, YP and Method III

the parametric counterpart in both the convergence rate and asymptotic distribution.

For inference of $\gamma$ based on the $t$ statistic, we need to estimate $\Sigma$ in Theorem 3. A straightforward way is to use its sample analog. Specifically, we estimate $\Sigma$ by

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

where

$$\widehat{\Delta}^2(\gamma) = \left( \widehat{m}_-(x_i, \gamma) \widehat{f}_-(x_i, \gamma) - \widehat{m}_+(x_i, \gamma) \widehat{f}_+(x_i, \gamma) \right),$$

$$\widehat{f}(x_i) = \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} K_{h,ij}^\tau,$$

$$\widehat{u}_i(\gamma) = y_i - \widehat{m}_-(x_i, \gamma) 1(q_i \leq \gamma) - \widehat{m}_+(x_i, \gamma) 1(q_i > \gamma),$$

with

$$\widehat{m}_\pm(x_i, \gamma) = \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} \frac{K_{h,ij}^\tau k_h^\pm (q_j - \gamma) y_j}{\sum_{j=1, j \neq i}^{n} K_{h,ij}^\tau k_h^\pm (q_j - \gamma)},$$

$$\widehat{f}_\pm(x_i, \gamma) = \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} K_{h,ij}^\tau k_h^\pm (q_j - \gamma).$$

The next theorem shows that $\widehat{\Sigma}$ is consistent.

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

Another inference method 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 framework. Our LR statistic can be used to test whether $\gamma = \gamma_0$ and is defined as

$$LR_n^{(1)}(\gamma) = nh^2 k_+^2(0) \frac{E[\Delta f(x_i)\Delta f(x_i)|q_i = \gamma_0]}{E[\Delta f(x_i)\Delta f(x_i)(\sigma^2_+(x_i) + \sigma^2_-(x_i))|q_i = \gamma_0]} \left(\hat{Q}_n(\gamma) - \hat{Q}_n(\gamma)\right).$$

Corollary 2 Under the assumptions of Theorem 3

$$LR_n^{(1)}(\gamma_0) \xrightarrow{d} \chi^2_1.$$

To construct the CI for $\gamma$ based on this ratio, we need to estimate $\frac{E[\Delta f(x_i)\Delta f(x_i)|q_i = \gamma_0]}{E[\Delta f(x_i)\Delta f(x_i)(\sigma^2_+(x_i) + \sigma^2_-(x_i))|q_i = \gamma_0]}$. A natural estimator is $\frac{1}{nh} \sum_{i=1}^{nh} k_n(q_i - \hat{\gamma}) \hat{\Delta}^2 f(x_i) \hat{f}^{-1}(x_i, \hat{\gamma}) f(x_i)$, which is consistent from Theorem 4. Now, the $(1 - \alpha)$ LR-CI for $\gamma$ is

$$\left\{ \gamma : \hat{LR}_n^{(1)}(\gamma) \leq \text{crit} \right\},$$

where $\hat{LR}_n^{(1)}(\gamma)$ replaces $\frac{E[\Delta f(x_i)\Delta f(x_i)|q_i = \gamma_0]}{E[\Delta f(x_i)\Delta f(x_i)(\sigma^2_+(x_i) + \sigma^2_-(x_i))|q_i = \gamma_0]}$ in $LR_n^{(1)}(\gamma)$ by its estimate, and crit is the $(1 - \alpha)$ quantile of $\chi^2_1$.

5 Inference Based on the IDKE with Shrinking Threshold Effects

In the last section, we adjust the IDKE by letting $k_+(0) = 0$ to construct CIs for $\gamma$. Nevertheless, the threshold effect is still assumed to be fixed in Method II. In this section, we adjust the IDKE from another perspective - assuming the threshold effect shrinking to zero with the sample size but $k_+(0) > 0$.

First of all, what is the meaning of shrinking threshold effects? As argued in Section 2.4 of YP, the local shifter $1(q > \gamma)$ plays the role of an instrument. Since when $q$ shifts from the left side of $\gamma$ to the right side, the shift in the mean of $y$ shrinks to zero, this is a weak IV problem in the threshold regression context. A natural question in such a context is the identifiability of $\gamma$ as $\delta$ shrinks to zero. In other words, we try to find the minimum magnitude of $\delta$ that can ensure the identification of $\gamma$. For this purpose, we cast the model into the following general framework.

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

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

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

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

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

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

Only if $\delta$ is large enough, the right side will converge to zero; the best rate of convergence of $R(\hat{\gamma}, \mathcal{P})$ to zero is called the optimal rate of convergence or the minimax rate of convergence. Now $P \in \mathcal{P}$ in our model is
characterized by $\delta$ and $g(x, q)$ as follows

$$
\mathcal{P}(s, B) = \left\{ P_{g, \delta, \gamma} : \frac{dP_{g, \delta, \gamma}}{d\mu} = f(x, q)\varphi_{x,q} (y - g(x, q) - x'\delta(\gamma \leq \gamma)), g(x, q) \in C_s(B, \mathcal{X} \times \Gamma_\varepsilon), \|\delta\| \leq B, \gamma \in \Gamma \right\},
$$

where $\mu$ is Lebesgue measure on $\mathbb{R}^{d+1}$, $\varphi_{x,q}$ is the conditional density of $u$ given $(x', q)'$, and $C_s(\cdot, \cdot)$ is defined in Section 2.2.

To ease the statement of our theorem, define

$$
\delta_n = \sqrt{\int (m_-(x) - m_+(x))^2 f(x|\gamma)dx},
$$

where $\gamma = \gamma(P)$. This definition of $\delta_n$ allows $g(x, q)$ to be kinked or discontinuous at $\gamma$. If $g(x, q)$ is smooth in the neighborhood of $\gamma$, then $\delta_n = |\mathbb{E}[x|q = \gamma]'|\delta = O(|\delta|)$.

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

$$
\liminf_{n \to \infty} \sup_{P \in \mathcal{P}(s, B)} P \left( n\delta_n^2 |\hat{\gamma} - \gamma(P)| > \frac{\varepsilon}{2} \right) \geq C,
$$

and if $n^{-\frac{1}{s}}\delta_n = O(1)$, then

$$
\liminf_{n \to \infty} \sup_{P \in \mathcal{P}(s, B)} P \left( |\hat{\gamma} - \gamma(P)| > \frac{\varepsilon}{2} \right) \geq C,
$$

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

We first clarify a key difference between the parametric and nonparametric threshold model 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 $n^{-1/2}$), $\gamma$ can be identified; in the latter, however, we require the jump size larger than $n^{-\frac{1}{s}}\delta_n$ to identify $\gamma$. In other words, the minimum rate of convergence for $\gamma$ in the nonparametric model must be larger than $n^{-\frac{1}{s}}\delta_n$ rather than any rate diverging to infinity in the parametric model. In the parametric model, $s = \infty$, so $n^{-\frac{1}{s}}\delta_n = n^{-1/2}$ and $n\delta_n^2 = O(1)$, i.e., the parametric result is a special case of Theorem 5. Such a difference between the parametric model and nonparametric model do not seem to be recognized in the literature, e.g., Müller and Song (1997) show that the convergence rate of the DKE is $n\delta_n^2$ when $x$ is empty by implicitly assuming $\delta_n$ to be larger than $n^{-\frac{1}{s}}\delta_n$. When $\gamma$ can be identified, the optimal rate of convergence for $\gamma$ is the same as in the parametric case. Actually, this rate is achievable by the IDKE as shown below.

To facilitate expression of the limit distribution of $\hat{\gamma}$, we define the following quantities

$$
D_n = \mathbb{E}[\Delta f(x_i)\Delta_i f(x_i)|q_i = \gamma_0]/\delta_n^2, \\
V_{1n} = \mathbb{E}[\Delta^2 f(x_i)f^2(x_i)|\sigma^2(x_i)|q_i = \gamma_0]/\delta_n^2, \\
V_{2n} = \mathbb{E}[\Delta^2 f(x_i)f^2(x_i)|\sigma^2(x_i)|q_i = \gamma_0]/\delta_n^2.
$$

We also impose the following conditions on the bandwidth $h$.

**Assumption H**: $h \to 0$, $\sqrt{n}h^d / \ln n \to \infty$, $\delta_n \to 0$, $\delta_n/h \to \infty$, $n^{3/4}h^d \delta_n^2 \to \infty$.

This assumption implies $n\delta_n^2 = \left(\sqrt{\ln n}\delta_n^2/h^d\right)^2 \left(\sqrt{\ln n}/\ln n\right) \to \infty$ with $\sqrt{n}h^d / \ln n \to \infty$, $\delta_n/h \to \infty$ and $d \geq 2$. The limit distribution of $\hat{\gamma}$ is given in the next result.
Corollary 3
Under the assumptions of Theorem 6 with Assumption H and of Bai (1997):

\[ \text{difference between the magnitude of } \gamma \text{, in other words, we use a different normalization (the convergence rate is hard to compare, depending on the magnitude of } \delta_n \text{) on } (\hat{\gamma} - \gamma_0) \text{ to achieve a nondegenerate asymptotic distribution.} \]

Figure 5 shows the difference between \( N(0,1) \) and \( \arg\max_v Z(v) \), where the density of \( \arg\max_v Z(v) \) is reported in Appendix B of Bai (1997):

\[
p(x) = \begin{cases} 
-\frac{1}{2} \Phi \left( -\frac{\sqrt{x}}{\lambda} \right) + \frac{1}{2} \left( 1 + \frac{3}{2} \right) \exp \left( \frac{1}{2} \frac{1}{\lambda} (1 + \frac{1}{2}) |x| \right) \Phi \left( -\frac{(1+\frac{1}{2})\sqrt{|x|}}{\lambda} \right), & \text{if } x < 0, \\
-\frac{1}{2} \Phi \left( -\frac{1}{2} \sqrt{\lambda} \right) + (1 + \frac{1}{2\lambda}) \exp \left( \frac{1+\lambda}{2\lambda} x \right) \Phi \left( -\left( \sqrt{\lambda} + \frac{1}{2\sqrt{\lambda}} \right) \sqrt{x} \right), & \text{if } x > 0,
\end{cases}
\]

with \( \Phi(\cdot) \) being the cdf of \( N(0,1) \).

For comparison, we state also the asymptotic distribution of the DKE in the following corollary. For this purpose, we define \( \delta_o = (1,x_o,\gamma_0) \delta_0 \), and adjust Assumption H' to:

Assumption H': \( h \to 0, \sqrt{n} h / \ln n \to \infty, \delta_o \to 0, \delta_o / h \to \infty, n^{3/4} h^d \delta_o \to \infty. \)

Corollary 3
Under the assumptions of Theorem 6 with Assumption H' replaced by H',

\[
\frac{nh^{d-1} \delta_o^2 f(x_o, \gamma_0)}{\sigma_o^2(x_o)} (\hat{\gamma} - \gamma_0) \xrightarrow{d} \arg\max_v Z_o(v),
\]

where

\[
Z_o(v) = \begin{cases} 
W_1(-v) - \frac{|v|}{2\kappa}, & \text{if } v \leq 0, \\
\sqrt{\lambda_o} W_2(v) - \frac{|v|}{2\kappa}, & \text{if } v > 0,
\end{cases}
\]

with \( \lambda_o = \frac{\sigma_o^2(x_o)}{\sigma_o^2(x_o)} \), \( \kappa^2 \) being defined in (18), and \( W_\ell(v), \ell = 1,2, \) being defined in Theorem 6.

Compared with the convergence rate in Method III \( (nh_o^2) \), the convergence rate of the DKE \( (nh^{d-1} \delta_o^2) \) is much slower especially when \( d \) is large. It is also interesting to notice that the asymptotic distribution in

\footnote{Actually, when \( \delta_o \to 0 \), the result in (20) is still correct. Now, the convergence rate is \( \sqrt{nh_o^2}/h \). Also, the CI based on \( LR_n^{(1)}(\gamma) \) is still valid.}
Method III does not depend on the kernel choice while the asymptotic distribution of the DKE does depend on the kernel choice on $x$ (although not on $q$). These results resound Theorem 1 and Corollary 1 of YP where $\delta_n$ is fixed and $k_{\pm}(0) > 0$. On the contrary, when $k_{\pm}(0) = 0$, from Theorem 3, the asymptotic distribution of the IDKE depends on the kernel choice on $q$, and from (18), the asymptotic distribution of the DKE depends on the kernel choice on both $x$ and $q$. In other words, whether $k_{\pm}(0) = 0$ indeed affects the role of the kernel on $q$ in the data usage (or efficiency) of the estimators.

We next discuss the inference on $\gamma$ based on our IDKE. Although we can construct the Wald-type CI by inverting the asymptotic distribution of $\hat{\gamma}$ in Theorem 6, Hansen (2000) shows that such kind of CIs perform poorly due to the identification failure when $\delta_n = 0$. He suggests to construct CIs for $\gamma$ by inverting the LR statistic which in our case is defined as

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

**Corollary 4** Under the assumptions of Theorem 6

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

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

To construct CIs for $\gamma$, we need to estimate $D_n/V_{1n}$ and $\lambda$. By similar procedures as in the last section, we can show that

$$\frac{\hat{D}_n}{V_{1n}} = \frac{n}{n} \sum_{i=1}^{n} k_h(q_i - \hat{\gamma}) \Delta_{\hat{\gamma}} \tilde{f}(x_i),$$

$$\hat{\lambda} = \frac{n}{n} \sum_{i=1}^{n} k_h(q_i - \hat{\gamma}) \Delta_{\hat{\gamma}} \tilde{f}^2(x_i).$$

Figure 5: Comparison Between the PDFs of $N(0,1)$ and $\arg\max_v Z(v)$: $\lambda = 0.5, 1, 2$
are the required consistent estimators, where \( \hat{\Delta}_i(\gamma) \), \( \hat{f}(x_i, \hat{\gamma}) \), \( \hat{f}(x_i) \) and \( \hat{u}_i \) are defined in the last section. If \( \sigma_+^2(x_i) = \sigma_{10}^2 \) and \( \sigma_-^2(x_i) = \sigma_{20}^2 \), then \( \lambda \) can be simply estimated by

\[
\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} \frac{k_i^+(q_i - \hat{\gamma})\hat{u}_i^2}{\sum_{i=1}^{n} k_i^-(q_i - \hat{\gamma})\hat{u}_i^2}.
\]

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

\[
\left\{ \gamma : \hat{L}R_n^{(2)}(\gamma) \leq \text{crit} \right\},
\]

where \( \hat{L}R_n^{(2)}(\gamma) \) replaces \( D_n/V_{1n} \) in \( LR_n^{(2)}(\gamma) \) by its estimates, and \( \text{crit} \) is the \((1 - \alpha)\) quantile of \( M \) with \( \lambda \) being substituted by its estimate.

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

\[
LR_n^{(1)}(\gamma) = nh \frac{k_+^{(0)}(\xi)}{\xi(1)} \frac{D_n}{V_{1n} + V_{2n}} \left( \hat{Q}_n(\hat{\gamma}) - \hat{Q}_n(\gamma) \right).
\]

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

\[
\frac{LR_n^{(1)}(\gamma)}{LR_n^{(2)}(\gamma)} = 4 \frac{k_+^{(0)}k_+^{(0)}(\xi)}{\xi(1)} \frac{V_{1n}}{V_{1n} + V_{2n}} =: R_n.
\]

If the model is locally homoskedastic in each regime, then \( R_n = 4 \frac{\sigma_{10}^2 \xi(1)}{\sigma_{10}^2 + \sigma_{20}^2} \), which is further simplified to \( 2 \frac{k_+^{(0)}k_+^{(0)}(\xi)}{\xi(1)} \) when the model is locally homoskedastic in both regimes. However, \( \hat{Q}_n(\hat{\gamma}) - \hat{Q}_n(\gamma) \) are not the same in \( LR_n^{(1)}(\gamma) \) and \( LR_n^{(2)}(\gamma) \) because the employed kernels are different and \( \hat{\gamma} \)'s are different.

To compare the asymptotic distributions of these two LR statistics, we plot their asymptotic pdfs in Figure 6 and report their 95% critical values in Table 2.

<table>
<thead>
<tr>
<th>Test Stat.</th>
<th>( LR_n^{(1)} )</th>
<th>( LR_n^{(2)}(\lambda = 0.5) )</th>
<th>( LR_n^{(2)}(\lambda = 1) )</th>
<th>( LR_n^{(2)}(\lambda = 2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>95% crit</td>
<td>3.841</td>
<td>3.040</td>
<td>3.676</td>
<td>6.081</td>
</tr>
</tbody>
</table>

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

We close this section by comparing the IDKE with the LSE in the parametric case (see, e.g., Hansen (2000)). From YP,

\[
\hat{\gamma}_{LSE} = \arg \max_{\gamma} \left( \delta' X' \left[ X (X'X)^{-1} X'_{\geq \gamma}X_{\geq \gamma} (X'X)^{-1} X_{\leq \gamma}X_{\leq \gamma} (X'X)^{-1} X' \right] \left( X\delta \right) \right),
\]

where \( \hat{\delta} \) is the LSE of \( \delta \) based on the splitting of \( \gamma \) so that \( X\hat{\delta} \) mimics \( \{\hat{\Delta}_i\}_{i=1}^n \) and \( X \) and \( X_{\leq \gamma} \) (similarly for \( X_{\geq \gamma} \)) are defined in (2). 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 \( \sigma^2 D_n/V_{1n} \) reduces to \( \frac{\|\Delta_i^2|_{q_i=\gamma_0}\|^2}{\|\Delta_i^2|_{q_i=\gamma_0}\|^2} \) and \( \lambda \) reduces to \( \frac{\|\Delta_i^2|_{q_i=\gamma_0}\|^2}{\|\Delta_i^2|_{q_i=\gamma_0}\|^2} \) if \( \sigma_+^2(x_i) = \sigma_{10}^2 \) and \( \sigma_-^2(x_i) = \sigma_{20}^2 \), then \( \delta^2 D_n^2/V_{1n} \) further reduces to \( \frac{\|\Delta_i^2|_{q_i=\gamma_0}\|^2}{\|\Delta_i^2|_{q_i=\gamma_0}\|^2} \) and \( \lambda \) further reduces to \( \frac{\sigma_+^2}{\sigma_{10}^2} \). A natural question is how to generate the same asymptotic distribution as in the parametric case. By carefully checking our proofs, we
can show that if the objective function of the IDKE changes to
\[
\frac{1}{n} \sum_{i=1}^{n} \frac{\hat{f}(x_i, \gamma)}{f(x_i)} \left[ \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} y_j K_{h,i,j}^{-} \gamma + \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} y_j K_{h,i,j}^{+} \gamma \right] - \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} K_{h,i,j}^{-} \gamma^2 \bigg],
\]
then the asymptotic distribution of the IDKE is the same as that of the parametric LSE, where \( \hat{f}(x_i, \gamma) \) is a consistent estimator of \( f(x, \gamma) \), e.g., \( \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} K_{h,i,j}^{-} \gamma \), and \( \hat{f}(x_i) \) is a consistent estimator of \( f(x_i) \), e.g., \( \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} K_{h,i,j}^{+} \gamma \). Asymptotically, we impose a weight \( \frac{1}{f(x_i, \gamma_0) f(x_i)} \) on \( \Delta_f(x_i)^2 \) or impose a weight \( f_{q|x}(\gamma_0 | x_i) \) on \( \Delta^2 \). This weight is very intuitive when there are more data points in the neighborhood of \( q = \gamma_0 \) at \( x_i \), we impose a larger weight on \( \Delta^2 \). Actually, 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. In other words, the complicated weights in the square bracket of (21) are asymptotically equivalent to \( \{ f_{q|x}(\gamma_0 | x_i) \}_{i=1}^{n} \); such an equivalence result is not obvious from the least squares objective function.

6 Two Specification Tests

In this section, we study the limit theory for the two specification tests in Section 2.2. We first specify some necessary regularity conditions.

\[\text{using such an objective function, the asymptotic distribution of the IDKE with } k_{\pm}(0) = 0 \text{ in Section 2 would also change. For example, } \Sigma \text{ would change to } \frac{E[|\Delta^2 u_i^2|_{|y_i = \gamma_0|} + E[|\Delta^2 u_i^2|_{|y_i = \gamma_0|}]] \xi(1)}{f_{\gamma_0}(E[|\Delta^2 u_i^2|_{|y_i = \gamma_0|}]^2 k_{\pm}(0)^2^2)} \text{, and } LR(1)^{(1)}(\gamma) \text{ changes to } \frac{\hat{Q}_n (\gamma) - \hat{Q}_n (\gamma)}{\hat{Q}_n (\gamma) - \hat{Q}_n (\gamma)}.\]
**Assumption F'**: $f(x,q) \in C_1 (B, \mathcal{X} \times \mathcal{Q})$.

**Assumption F''**: $f(x,q) \in C_1 (B, \mathcal{X} \times \Gamma_r)$ with $\lambda \geq 1$, and $0 < \bar{f} \leq f(x,q) \leq \bar{f} < \infty$ for $(x',q') \in \mathcal{X} \times \Gamma_r$.

**Assumption G''**: $g(x,q) \in C_s (B, \mathcal{X} \times \mathcal{Q})$ with $s \geq 2$.

**Assumption U'**:

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

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

(c) $\mathbb{E}[u^2|x,q]$ is uniformly bounded on $(x',q') \in \mathcal{X} \times \mathcal{Q}$.

For $I_n^{(2)}$, we can replace Assumption U' by Assumption U and replace $\mathcal{Q}$ by $\Gamma_r$ in Assumptions G''.

**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 $H_0^{(2)}$. 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 $\hat{e}_i \hat{e}_j$ behave like a squared term which 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. But the null hypothesis is better modified to the equivalence $m_-(x) = m_+(x)$ for all $x \in \mathcal{X}$ and $g$ in Assumption G'' need not 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 $q = \gamma_0$.

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

$l_h(\cdot,t)$ may be a higher order kernel to reduce the bias in $\hat{g}_t$.

### 6.1 Limit Theory for the Two Tests

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

**Theorem 7** Under Assumptions B1, F', G', K, S, and U', the following hold:

(i) $I_n^{(1)} \overset{d}{\to} N\left(0, \Sigma^{(1)}\right)$

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

$$\Sigma^{(1)} = 2 \int k^{2d}(u)du \mathbb{E} \left[ f(x,q) \sigma^4(x,q) \right]$$
Theorem 8

Under Assumptions B2, \( F \) cases that do generate power include (i) specification testing without threshold effects – see, e.g., Bierens and Ploberger (1997, p. 1135). Possible under \( H \) point from following special example to illustrate. Suppose according to this result, \( I \) is \( \frac{2h^d}{n(n-1)} \sum_i \sum_{j \neq i} K^2_{h,i,j} c_i^2 c_j^2 \).

As a result, the test based on the studentized test statistic \( T_{n}^{(1)} = \frac{I_{n}^{(1)}}{v_{n}^{(1)}} \)

\[ i_{n}^{(1)} = 1 \left( T_{n}^{(1)} > z_{\alpha} \right) , \]

has significance level \( \alpha \), where \( z_{\alpha} \) is the \( 1 - \alpha \) quantile of the standard normal distribution\(^{31}\)

(ii) If under \( H_{1}^{(1)} \), \( m(x, q) - \bar{m}(x, q) = n^{-1/2} h^{-d/4} \delta_n(x, q) \) such that \( \int \delta_n(x, q)^2 f(x, q)^2 dx dq \to \delta \), then

\[ I_{n}^{(1)} \overset{d}{\to} N \left( \delta, \Sigma^{(1)} \right) \text{ and } T_{n}^{(1)} \overset{d}{\to} N \left( \delta / \sqrt{\Sigma^{(1)}}, 1 \right) , \]

so that the test \( i_{n}^{(1)} \) is consistent and \( P_{m} \left( T_{n}^{(1)} > z_{\alpha} \right) \to 1 \) for any \( m(\cdot) \) such that \( \int (m(x, q) - \bar{m}(x, q))^2 f(x, q)^2 dx dq \neq 0 \). Furthermore, the result continues to hold when \( z_{\alpha} \) is replaced by any nonstochastic constant \( C_{n} = o \left( nh^{d/2} \right) \).

According to this result, \( I_{n}^{(1)} \) does not have power if \( \mathbb{E} \left[ (m(x, q) - \bar{m}(x, q))^2 f(x, q) \right] = 0 \). Consider the following special example to illustrate. Suppose \( m(x, q) \) under \( H_{0}^{(1)} \) is \( x' \beta + x' \delta_1 (q \leq \gamma) \), and the alternative is \( m(x, q) = x' \beta + x' \delta_0 (q \leq \gamma) + x' \xi + x' \zeta 1 (q \leq \gamma) \), then obviously, \( \mathbb{E} \left[ (m(x, q) - \bar{m}(x, q))^2 f(x, q) \right] = 0 \) under \( H_{1}^{(1)} \) and \( I_{n}^{(1)} \) does not have any 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 power include (i) \( m(x, q) \) takes the parametric form but has a different threshold point from \( \bar{m}(x, q) \); (ii) \( m(x, q) \) takes a nonparametric form.

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

(i)

\[ I_{n}^{(2)} \overset{d}{\to} N \left( 0, \Sigma^{(2)} \right) \]

uniformly over \( \mathcal{H}_{0}^{(2)} \), where

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

and can be consistently estimated by

\[ v_{n}^{(2)} = \frac{2h^d}{n(n-1)} \sum_i \sum_{j \neq i} 1_{i}^T 1_{j}^T K^2_{h,i,j} c_i^2 c_j^2 . \]

As a result, the test based on the studentized test statistic \( T_{n}^{(2)} = \frac{I_{n}^{(2)}}{v_{n}^{(2)}} \)

\[ i_{n}^{(2)} = 1 \left( T_{n}^{(2)} > z_{\alpha} \right) , \]

has significance level \( \alpha \), where \( z_{\alpha} \) is the \( 1 - \alpha \) quantile of the standard normal distribution.

\(^{31}\) The test is a one-sided because \( I_{n}^{(1)} \) is based on the \( L^2 \)-distance between \( m(\cdot) \) and \( \bar{m}(\cdot) \).
(ii) If under $H_0^{(2)}$, $m_-(x) - m_+(x) = n^{-1/2}h^{-d/4}b^{-1/2}\delta_n(x)$ such that $\int \delta_n(x)^2 f(x, \gamma_0) dx \to \delta$, then

$$I_n^{(2)} \to_d N \left( \kappa \delta, \Sigma^{(2)} \right) \quad \text{and} \quad T_n^{(2)} \to_d N \left( \kappa \delta / \sqrt{\Sigma^{(2)}}, 1 \right),$$

where $\kappa = 2 \int_0^1 \left( \int_u^1 F(u) du \right)^2 dv$, and the test $t_n^{(2)}(x)$ is consistent with $P_m \left( T_n^{(2)} > z_\alpha \right) \to 1$ for any $m$ such that $\int (m_-(x) - m_+(x))^2 f(x, \gamma_0) dx \neq 0$. The result continues to hold when $z_\alpha$ 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 nontrivial power – specifically, $n^{-1/2}h^{-d/4}b^{-1/2}/\sqrt{n^{-1/2}h^{-d/4}} = b^{-1/2} \to \infty$.

### 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 $hd/2$ and under $H_0^{(2)}$ is $nh^{d/2}b^2$. 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 $T_n^{(l)}$ 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 $\hat{e}_i$ and $\hat{y}_i$ having different definitions in the two tests.

**Algorithm WB:**

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

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

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

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

**Step 3:** Use the bootstrap samples to compute the statistics

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

$$I_n^{(2)*} = \frac{nh^{d/2}}{n(n-1)} \sum_i \sum_{j \neq i} 1^n_i 1^n_j K_{h,ij} \hat{e}_{i}^* \hat{e}_{j}^*.$$

---

32 To construct $I_n^{(2)*}$, we need only the data with $q_i \in \left[ \gamma - b, \gamma + b \right]$. 
and the estimated asymptotic variances

\[ v_n^{(1)*2} = \frac{2h^d}{n(n-1)} \sum_{i,j \neq i} K_{h,ij}^2 \tilde{\varepsilon}_i^2 \tilde{\varepsilon}_j^2, \]

\[ v_n^{(2)*2} = \frac{2h^d}{n(n-1)} \sum_{i,j \neq i} 1^\Gamma_i 1^\Gamma_j K_{h,ij}^2 \tilde{\varepsilon}_i^2 \tilde{\varepsilon}_j^2. \]

The studentized bootstrap statistics are \( T_n^{(i)*} = T_n^{(i)} / v_n^{(i)*} \). Here, the same \( b \) and \( h \) are used as in \( T_n^{(i)} \) and \( v_n^{(i)*2} \) in Theorem 7 and 8.

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

In Step 1, a popular way to simulate \( U_i^* \) in the second test is based on \( \tilde{e}_i \)'s centralized counterpart \( \bar{e}_i = \tilde{e}_i - \bar{e} \) rather than \( \tilde{e}_i \) itself, where \( \bar{e} = \sum_{i=1}^n \tilde{e}_i 1^\Gamma_i / \sum_{i=1}^n 1^\Gamma_i \), \( \Gamma_b = (\gamma - b, \gamma + b) \); see, e.g., Gijbels and Goderniaux (2004) and Su and Xiao (2008). Such a formulation can lead to \( \sum_{i=1}^n \tilde{e}_i 1^\Gamma_i / \sum_{i=1}^n 1^\Gamma_i = 0 \) 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 \( T_n^{(i)*} \) will mimic the null distribution of \( T_n^{(i)} \) even when the null hypothesis is false. When the null is false, \( \tilde{e}_i \) is not a consistent estimate of \( e_i \) or \( U_i \). Nevertheless, the following theorem shows that the above bootstrap procedure is valid. This is because our studentized test statistic \( T_n^{(i)} \) is invariant to the variance of \( e \). But the wild bootstrap procedure is not valid if the test statistic \( J_n^{(i)} \) is used instead of \( T_n^{(i)} \).

**Theorem 9** Under the assumptions of Theorem 3 and 8

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

where \( \Phi(\cdot) \) is the cumulative distribution function of a standard normal random variable.

## 7 Simulations

We conduct some simulations in this section to check the performances of our CIs and testing procedures. We will concentrate on the procedures whose performances are unclear from the literature. For inference, we will check only the CIs which invert the two LR statistics in Section 4 and 5. We will not compare the performances of the IDKE and the DKE because YP have already shown that the former performs much better than the latter in finite samples. We will not check the performances of the CIs based on bootstrapping our 2SLS or efficient GMM because there are enough evidences in the literature on the

---

33 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.

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

35 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 \( \gamma_0 \).
performances of bootstrap when the moment conditions are nonsmooth. We do not check the performances of the CIs based on inverting the \( t \) statistics either because as argued in Hansen (2000), their performances are not satisfactory when the identification is weak. For the two specification tests, we will 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. Another reason for why we concentrate on these two CI construction methods and these two specification tests is because neither of them involves instruments. As mentioned in the introduction, good instruments are hard to find and justify in practice.

We will use a similar data generating process (DGP) as in YP. 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 \( \varepsilon \mid (x, q) \sim N(-\delta_2 q^2, 0.1^2) \). In CI construction, we let \( \delta_1 = 0.1 \) and 0.2, indicating the small and large threshold effects (see Table 5 below), respectively, and \( \delta_2 = 1 \), indicating severe endogeneity (see Table 4 below). In testing endogeneity, we let \( \delta_1 = 0.2 \) and \( \delta_2 = 0.2, 0.5 \) 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 \( k_x(0) = 0 \),

\[
    k_-(x, r) = -x(1 + x)1(-1 \leq x \leq r) \left/ \left( \frac{1}{6} - \frac{1}{2} r^2 - \frac{1}{3} r^3 \right) \right.,
\]

and for the IDKE with shrinking threshold effects,

\[
    k_-(x, r) = \frac{3}{4} (1 - x^2)1(-1 \leq x \leq r) \left/ \left( \frac{1}{2} + \frac{3}{4} r - \frac{1}{4} r^3 \right) \right., \quad 0 \leq r \leq 1,
\]

which degenerates to the Epanechnikov kernel when \( r = 1 \); \( k_+(x, r) = k_-(x, r) \).\footnote{Although the literature (e.g., Zheng, 1996; Li and Wang, 1998) provides simulation results when the approximation function \( \hat{m}(x, q) \) in (3) is smooth, there is no corresponding results when \( \hat{m}(x, q) \) is discontinuous. Also, although Porter and Yu (2015) investigate the finite-sample performance of a similar structural change test as \( I_n^{(2)} \) in this paper, no covariates \( x \) are included there.} In both tests, the kernel in \( K_{h,i} \) is specified in (22), and in the second test, \( \hat{y}_i \) is estimated by the local linear smoother which implies a second-order boundary kernel in \( L_{h,ij} \) as required in Assumption L. Following DH, three bandwidths \( b \) 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 \( h/b \to 0 \), \( nh^{d/2} \delta^2 \to 0 \) with \( \eta = 2 \).\footnote{These kernel functions imply \( \xi(1) = 12 \) and \( k'_x(0) = 6 \) in Corollary 2 and \( k_x(0) = 1.5 \) in Corollary 3. So \( \frac{L_k^{1.5}(\gamma)}{L_k^{2.5}(\gamma)} = \frac{2k_x(0)k'_x(0)}{\xi(1)} = 1.5 \) in our DGP.} The simulation study in Müller (1991) 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 considered. In Algorithm WB, when \( n = 500 \), \( B = 399 \), and when \( n = 1000 \), \( B = 199 \). For CI construction, the level of significance is set as 95%, and for testing, the level of significance is set as 5%.

### 7.1 Two LR-Based CIs

The coverage and length of the two LR-based CIs are reported in Table 3. The results show that both methods perform well in coverage. The choice of bandwidth and expansion of sample sizes improves marginally in coverage. On the other hand, different bandwidths and samples sizes have big impacts on CI lengths.

\[\text{Notice that the range of bandwidths chosen is quite large, being the ratio between the first and last one equal to 2. In the estimation of } \gamma, \text{ the bandwidth is smaller than the usual optimal bandwidth (which is of rate } n^{-1/4} \text{) as suggested in Porter and Yu (2015). In the second test, } N = n \times (2 \times \frac{1}{2} C^{1/2} n^{-1/4})^2 = Cn^{1/2} \text{ data points are used to obtain } \hat{y}_i. \text{ When } C = 2 \text{ and } n = 500, N \approx 45. \text{ When } C = 4 \text{ and } n = 1000, N \approx 126.\]
Specifically, under our DGP, the medium bandwidth seems to perform satisfactorily in length among various scenarios and a larger sample size shrinks the length significantly. Another phenomenon deserved to mention is that the length decreases sharply when the jump size doubles in both methods. This is expectable because larger threshold effects make the threshold point easier to identify. Comparing Method II to Method III, the later behaves a little bit better in coverage. This may stem from the fact that the latter makes full use of the data information around $\gamma_0 (k_+(0) > 0)$ while the former makes only marginal use of such information $(k_+(0) = 0)$. As a cost, the CIs in Method III are generally longer than those in Method II.

<table>
<thead>
<tr>
<th>$k_+(0) = 0$</th>
<th>Coverage</th>
<th>Length ($\times 10^{-2}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>500</td>
<td>1000</td>
</tr>
<tr>
<td>$\delta_1$</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>$C = 2$</td>
<td>0.998</td>
<td>0.962</td>
</tr>
<tr>
<td>$C = 3$</td>
<td>0.994</td>
<td>0.958</td>
</tr>
<tr>
<td>$C = 4$</td>
<td>0.984</td>
<td>0.954</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\delta_n \rightarrow 0$</th>
<th>Coverage</th>
<th>Length ($\times 10^{-2}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>500</td>
<td>1000</td>
</tr>
<tr>
<td>$\delta_1$</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>$C = 2$</td>
<td>1</td>
<td>0.980</td>
</tr>
<tr>
<td>$C = 3$</td>
<td>1</td>
<td>0.986</td>
</tr>
<tr>
<td>$C = 4$</td>
<td>0.998</td>
<td>0.984</td>
</tr>
</tbody>
</table>

Table 3: Comparison of Inference Methods: Coverage and Average Length of Nominal 95% Confidence Intervals for $\gamma$ (Based on 500 Repetitions): $\delta_2 = 1$

### 7.2 Two Nonparametric Specification Tests

The size and power of the two nonparametric specification tests are reported in Tables 4 and 5, respectively. The results show that all tests have the size close to 5% except in the second test with a large $h$ where the test is under-sized. A large $h$ implies a large bias in $\hat{I}(2)$ so that 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 $N (0, 0.1^2)$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>500</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta_2$</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>$C = 2$</td>
<td>5.2</td>
<td>8</td>
</tr>
<tr>
<td>$C = 3$</td>
<td>5.8</td>
<td>10.2</td>
</tr>
<tr>
<td>$C = 4$</td>
<td>5.2</td>
<td>10.8</td>
</tr>
</tbody>
</table>

Table 4: Size and Power of $T_n^{(1)}$ (%): Significance Level = 5%, $\delta_1 = 0.2$ (Based on 500 Repetitions)
8 Conclusion

In this paper, we propose three inference methods on the threshold point when endogeneity is present. All three methods are valid regardless of the endogeneity of the threshold variable; to our knowledge, they are the only three inference methods in the literature with 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 instrument. In discussing the 2SLS method, we also clarify some mysteries in the literature. For example, why the usual GMM cannot identify the threshold point in structural change models; 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 while only one group of moments is enough when the threshold variable is endogenous. We also clarify why the bootstrap is valid for our 2SLS while is generally invalid for the usual GMM. In discussing the two IDKE-smoothing methods, we show that the two IDKEs use different normalizations to obtain operable asymptotic distributions under different assumptions and explain why their convergence rates are different; we also suggest to construct confidence intervals based on inverting the likelihood ratio statistics which are different only by a constant in these two methods. Our three inference methods provide much flexibility to practitioners. When instruments are available, the 2SLS combined with the bootstrap can be used; when instruments are absent, the other two methods can be used.

We further propose two specification tests; the first one is to check the existence of endogeneity and the second one is to check the presence of threshold effects. Our results show that it is possible to test for threshold effects in the absence of instrumentation even when endogeneity is present. And one important implication of the test for endogeneity in empirical work is that it helps to assess whether instruments are needed to achieve consistent estimation of the structural coefficients. Both tests are similar to score tests and are conveniently asymptotically normal, although for improving finite sample performance, a wild bootstrap procedure is suggested to obtain critical values.
References


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


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


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

