THRESHOLD REGRESSION WITH ENDOGENEITY

By

Ping Yu and Peter C. B. Phillips

December 2014

COWLES FOUNDATION DISCUSSION PAPER NO. 1966

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Threshold Regression with Endogeneity*

Ping Yu†
The University of Hong Kong

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

Started: August 2009
First Draft: October 2012
This Version: August 2014

Abstract

This paper studies estimation and specification testing in threshold regression with endogeneity. Three key results differ from those in regular models. First, both the threshold point and the threshold effect parameters are shown to be identified without the need for instrumentation. Second, in partially linear threshold models, both parametric and nonparametric components rely on the same data, which *prima facie* suggests identification failure. But, as shown here, the discontinuity structure of the threshold itself supplies identifying information for the parametric coefficients without the need for extra randomness in the regressors. Third, instrumentation plays different roles in the estimation of the system parameters, delivering identification for the structural coefficients in the usual way, but raising convergence rates for the threshold effect parameters and improving efficiency for the threshold point. Specification tests are developed to test for the presence of endogeneity and threshold effects without relying on instrumentation of the covariates. 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. Simulation studies corroborate the theory and the asymptotics. An empirical application is conducted to explore the effects of 401(k) retirement programs on savings, illustrating the relevance of threshold models in treatment effects evaluation in the presence of endogeneity.

KEYWORDS: Threshold regression, Endogeneity, Local shifter, Identification, Efficiency, Integrated difference kernel estimator, Regression discontinuity design, Optimal rate of convergence, Partial linear model, Specification test, U-statistic, Wild bootstrap, Threshold treatment model, 401(k) plan.

JEL-CLASSIFICATION: C12, C13, C14, C21, C26

*Thanks go to Rosanne Altshuler, Denise Doiron, Yunjong Eo, Denzil Fiebig, Bruce Hansen, Roger Klein, John Landon-Lane, Paul S.H. Lau, Jay Lee, Taesuk Lee, Chu-An Liu, Bruce Mizrach, James Morley, Tatsushi Oka, Woong Yong Park, Jack Porter, Wing Suen, Norman Swanson, Denis Tkachenko, Andrew Weiss, Ka-fu Wong, Steven Pai Xu and seminar participants at ANU, HKU, NUS, Rutgers, UNSW, UoA and 2013NASM for helpful comments.
†School of Economics and Finance, The University of Hong Kong, Pokfulam Road, Hong Kong; email: pingyu@hku.hk.
‡Cowles Foundation for Research in Economics, Yale University; email: peter.phillips@yale.edu

1 Introduction

In recognition of potential shifts in economic relationships, threshold models have become increasingly popular in econometric practice both in time series and cross section applications. A typical use of thresholds in time series modeling is to capture asymmetric effects of shocks over the business cycle (e.g., Potter, 1995). Other time series applications involving threshold autoregressive modeling of interest arbitrage, purchasing power parity, exchange rates, stock returns, and transaction cost effects are discussed in a recent overview by Hansen (2011). Threshold models are particularly common in cross sectional applications. For example, following a seminal contribution by Durlauf and Johnson (1995) on cross country growth behavior, Hansen (2000) showed how growth patterns of rich and poor countries can be distinguished by thresholding in terms of initial conditions relating to per capita output and adult literacy. Much of the relevance of threshold modeling in empirical work is explained by the preference policy makers and administrators have for threshold-related policies. For example, tax rates and welfare programs are commonly designed to depend on threshold income levels, merit-based university scholarships often depend on threshold GPA levels, and need-based aid programs generally depend on threshold levels of family income.

The usual threshold regression model splits the sample according to the realized value of some observed threshold variable q. The dependent variable y is determined by covariates $\mathbf{x} = (1, x', q) \in \mathbb{R}^{d+1}$ in the split-sample regression

$$y = \mathbf{x}'\beta_1 1(q \leq \gamma) + \mathbf{x}'\beta_2 1(q > \gamma) + \varepsilon,$$

where d is the dimension of the nonconstant covariates (x, q), the indicators $1(q \leq \gamma)$ and $1(q > \gamma)$ define two regimes in terms of the value of q relative to a threshold point given by the parameter γ, the coefficients β_1 and β_2 are the respective threshold parameters, and ε is a random disturbance. The model is therefore a simple nonlinear variant of linear regression and can conveniently be rewritten as

$$y = \mathbf{x}'\beta + \mathbf{x}'\delta 1(q \leq \gamma) + \varepsilon, \tag{1}$$

with regression coefficient $\beta = \beta_2$ and discrepancy coefficient $\delta = \beta_1 - \beta_2$. The central parameters of interest are $\theta \equiv (\beta', \delta', \gamma)'$.

An asymptotic theory of estimation and inference is now fairly well developed for linear threshold models such as (1) with exogenous regressors – see Chan (1993), Hansen (2000), Yu (2012a) and the references therein. In this framework, $\mathbf{x}$ is typically taken as exogenous in the sense that the orthogonality condition $\mathbb{E}[\varepsilon|x, q] = 0$ holds, thereby enabling least squares estimation which can be used to consistently estimate θ and facilitate inference. While the assumption is convenient, exogeneity is often restrictive in practical work and limits the range of suitable empirical applications of modeling with threshold effects. For instance, the empirical growth models used in Papageorgiou (2002) and Tan (2010) both suffer from endogeneous regressor problems, as argued in Frankel and Romer (1999) and Acemoglu et al. (2001). Endogenous regressor issues also arise in treatment effect models where there are often important policy implications, as evidenced in the empirical application to tax-deferred savings programs considered later in the paper. In fact, whenever endogeneity in the regressors is relevant in a linear regression framework, it will inevitably be present in the corresponding threshold model under the null of zero discrepancy.

Endogeneity is considered in some existing work. For instance, Caner and Hansen (2004) use the asymptotic framework of Hansen (2000), where δ shrinks to zero, to explore the case where q is exogenous but x may be endogenous. In the same framework, except that q may also be endogenous, Kourtellos et al. (2009) consider a structural model with parametric assumptions on the data distribution and apply a sample selection technique (Heckman, 1979) to estimate γ. Kapetanios (2010) tests exogeneity of the instruments used in threshold regression by bootstrapping a Hausman-type test statistic within the Hansen (2000) framework. The common solution to the endogeneity problem in all this work is to employ instruments and to apply two-stage-least squares (2SLS) estimation, just as in linear regression (For related work on 2SLS estimation of structural change regression without thresholding, see Boldea et al. (2012), Hall et al. (2012) and Perron and Yamamoto (2012a)). However, Yu (2012b) shows that three typical 2SLS estimators of γ are generally inconsistent. This finding motivates us to search for general consistent estimators of γ. One of the main contributions of the present paper is to show that when only γ and δ are of interest, as in the typical case,[1] these parameters are both identified even without instruments. This result is meaningful to practitioners since good instruments are often hard to find and justify in practical work. A second contribution of the paper is to show how the parameters may be consistently estimated and inference conducted, thereby opening up many potential empirical applications.

Throughout the paper we assume that δ is fixed as in Chan (1993) and the data are i.i.d. sampled. If $\mathbb{E}[\varepsilon|x,q] \neq 0$, we can write model (1) in the form

$$y = m(x,q) + e = g(x,q) + \mathbf{x}'\delta 1\,(q \leq \gamma) + e, \tag{2}$$

where $m(x,q) = g(x,q) + \mathbf{x}'\delta 1\,(q \leq \gamma)$, $g(x,q) = \mathbf{x}'\beta + \mathbb{E}[\varepsilon|x,q]$ is any smooth function, and $e = \varepsilon - \mathbb{E}[\varepsilon|x,q]$ satisfies $\mathbb{E}\,[e|x,q] = 0$. This formulation falls within the framework of the general nonparametric threshold model

$$y = g(x,q) + \delta(x,q) 1\,(q \leq \gamma) + e, \tag{3}$$

where $g(\cdot)$ and $\delta(\cdot)$ are smooth functions. The special feature of (2) is that the jump size function $\delta(\cdot)$ at the threshold point has the linear parametric form $\mathbf{x}'\delta$.

Estimation of the threshold parameter γ in nonparametric regression is presently an unresolved problem in the literature. Our approach introduces a new estimator called the *integrated difference kernel estimator* (IDKE) that can be used to produce a consistent estimator of γ irrespective of whether q is endogenous. Moreover, the construction of this estimator does not depend on the linearity feature that $\delta(x,q) = \mathbf{x}'\delta$ in (2) so that the method can be applied in the general nonparametric threshold regression model (3). More strikingly, we show that this estimator is n-consistent and has a limiting distribution similar to the least squares estimator (LSE) when the exogeneity condition $\mathbb{E}[\varepsilon|x,q] = 0$ holds. The approach makes use of the jump information in the vicinity of the threshold point to identify γ, so that only the local information around γ is used for identification. Jumps such as those in (2) and (3) produce a form of nonstationarity in the process which can be used to aid identification and estimation. In this sense, the feasibility of consistent estimation without explicit instrumentation relates to recent findings by Wang and Phillips (2009) and Phillips and Su (2011) who show that nonparametric relationships involving nonstationary shifts are identified without instruments and can be consistently estimated by using only local information.

Given a consistent estimator of the threshold parameter γ, we propose two estimators of δ that are suggested from the partial linear model structure of (2) that applies for known γ.[2] An important difference between (2) and the usual partial linear structure is that both parametric and nonparametric components of $m(x,q) = g(x,q) + \mathbf{x}'\delta 1\,(q \leq \gamma)$ rely on the same data (x,q). It is well-known that extra randomness beyond (x,q) is usually required in the linear regressors of a partial linear model to ensure a sufficient signal to identify the linear coefficients. In the present model the linear component $\mathbf{x}'1\,(q \leq \gamma)$ is fully determined by (x,q) given γ, a fact that may *prima facie* suggest identification failure. However, the key argument for

[1] See Yu and Zhao (2013) for an example in treatment effects evaluation.

[2] In the notation of Robinson (1988), $Z = (x',q)$, $X = \mathbf{x}'1\,(q \leq \gamma)$, $\beta = \delta$ and $\theta\,(Z) = g(x,q)$.

identification failure is that the systematic part of the model (2) can be written as

$$m(x,q) = \mathbf{x}'\delta 1(q \leq \gamma) + g(x,q) = [\mathbf{x}'\delta 1(q \leq \gamma) + \eta(x,q)] + [g(x,q) - \eta(x,q)]$$

for all $\eta(x,q)$, suggesting that the (partial linear) component $\mathbf{x}'\delta 1(q \leq \gamma)$ cannot be separated from $g(x,q)$ in the composite function $m(x,q)$. But this argument assumes that $\eta(x,q)$ is smooth (as is assumed for the nonparametric component $g(x,q)$) and it ignores the identifying information for δ in the discontinuity structure of the component $\mathbf{x}'\delta 1(q \leq \gamma)$ that arises from the jump in $m(x,q)$ at $q = \gamma$. It is this jump discontinuity that assures identification of the linear coefficients δ.

Although the coefficient vector δ is identified, our two estimators do not achieve the usual semiparametric $\sqrt{n}$ rate since these estimators use only local information in the neighborhood of $q = \gamma$. Further, the usual semiparametric consistency proof (Robinson, 1988) relies on the assumption that $\mathbb{E}[\mathbf{x}'\delta 1(q \leq \gamma)|x,q]$ is smooth in $(x',q)'$, but smoothness fails in the present case and the usual proof is no longer applicable. Instead, the new proof provided here is based on projections of U-statistics. A final contribution of the paper is to show that the optimal rate of convergence of δ is nonparametric, i.e., slower than $\sqrt{n}$, and that this rate is achieved by our suggested estimators. Section 3.3 of Porter (2003) and Section 2 of Yu (2010) contain some related discussion on this point in the simple case where q is the only covariate.

When instruments are available, the coefficients δ can be estimated at a $\sqrt{n}$ rate. In this case, for the linear endogeneous threshold model (1), β can also be estimated at a $\sqrt{n}$ rate. So the role of instruments in the model (1) is to provide identification for β and to improve the convergence rate of estimates of δ. As for the threshold parameter γ in (1), our results show that γ can be estimated at the rate n even if no instruments are available - so instruments have no import on this convergence rate. Instead, as with the earlier finding in Yu (2008), the role of instrumentation for γ is not to improve the convergence rate or to provide identification, but to improve efficiency. In summary, instrumentation plays different roles in the estimation of the system parameters β, δ and γ: only for β do instruments have the conventional role of delivering identification, whereas for δ and γ the presence of instruments serves to improve convergence rates or efficiency.

A further contribution of the paper is to specification testing. Two tests are considered. One tests for endogeneity and the other tests for the presence of threshold effects in the absence of instruments. Both tests are valuable in practical work where endogeneity and threshold effects may be suspected. The second test extends parametric structural change tests that are presently in the literature to the nonparametric case. Both tests are constructed under the null and are similar in spirit to score tests. The limit distributions of the tests are derived under the null and their local power functions are provided. For finite sample implementation, we recommend using the wild bootstrap to obtain critical values and the validity of this use of the bootstrap is established.

A brief simulation study is included to test the adequacy of the asymptotic theory of the estimation and test procedures in finite samples in the presence of threshold effects and endogeneity. The results confirm that the IDKE estimation procedure has good bias and root mean squared error properties in finite samples and that the test statistics have good size and detective power for threshold effects and endogeneity. An empirical application is conducted to explore the effects of 401(k) retirement programs on savings, giving particular attention to the important policy question of whether contributions to tax-deferred retirement plans represent additional savings or simply crowd out other types of savings, and illustrating the relevance of threshold models in treatment effects evaluation in the presence of endogeneity.

The remainder of the paper is organized as follows. In Section 2, we construct estimators of γ and δ and derive their limit distributions. Section 3 investigates the role of instruments. Section 4 develops specification tests. Section 5 covers some extensions and simplifications of our analysis. Section 6 reports the results of some finite sample simulations. Section 7 presents an empirical application to explore the effects of 401(k) retirement programs on savings. Section 8 concludes. Proofs with supporting propositions and lemmas are given in Appendices A, B and C, respectively.

A word on notation. C denotes a generic positive constant whose value may change in each occurrence. The parameters β and δ are partitioned conformably with the intercept and variables as $\left(\beta_\alpha, \beta_x', \beta_q\right)'$ and $\left(\delta_\alpha, \delta_x', \delta_q\right)'$. The symbol ℓ is used to indicate the two regimes in (1) or the two specification tests, and is not written out explicitly as '$\ell = 1, 2'$ except in Section 7 where there are three regimes. We use $f, f_{x|q}$, and f_q for the joint, conditional, and marginal probability densities of (x, q) , $x|q$, and q, respectively; $\|\cdot\|$ denotes the Euclidean norm unless otherwise specificied; and $\approx$ signifies that higher-order terms are omitted or a constant term is omitted, depending on the context.

2 The Integrated Difference Kernel Estimator

This section introduces a new methodology for consistently estimating γ and δ when instruments are absent. The method involves a nonparametric kernel estimator that we call the integrated difference kernel estimator (IDKE) A related estimator of γ that is already in the literature is the difference kernel estimator (DKE) of Qiu et al. (1991). As we explain below, some difficulties that arise in applying the DKE in the present case help to motivate the construction of the IDKE. The discussion that follows provides an intuitive rationale for the identification and consistent estimation of γ and δ without instruments, gives the limit theory for the IDKE and associated coefficient estimates, and relates these results to those of other approaches, including the LSE and the partial linear estimator (PLE).

2.1 Difficulties in Applying the DKE

When there are no other covariates besides q, the DKE is a popular procedure for estimating γ. Porter and Yu (2011) provide some discussion and references to the related literature. In this simple case, we have the model $y = g(q) + (1, q)\,\delta 1\,(q \leq \gamma) + e$ with $\mathbb{E}[e|q] = 0$. The DKE is defined as the extremum estimator

$$\widehat{\gamma}_{DKE} = \arg\max_{\gamma} \widehat{\Delta}^2(\gamma), \tag{4}$$

where $\widehat{\Delta}(\gamma) = \widehat{\mathbb{E}}[y|q = \gamma-] - \widehat{\mathbb{E}}[y|q = \gamma+]$ with $\widehat{\mathbb{E}}[y|q = \gamma-] = n^{-1}\sum_{j=1}^n w_j(\gamma) y_j d_j(\gamma)$ and $\widehat{\mathbb{E}}[y|q = \gamma+] = n^{-1}\sum_{j=1}^n w_j(\gamma) y_j \left(1 - d_j(\gamma)\right)$ being estimators of $\mathbb{E}[y|q = \gamma-]$ and $\mathbb{E}[y|q = \gamma+]$, and $d_j(\gamma) = 1\,(q_j \leq \gamma)$. In the definition of $\widehat{\mathbb{E}}[y|q = \gamma\pm]$, the weight function $w_j(\gamma) = k_h\left(q_j - \gamma\right) \Big/ \sum_{l=1}^n k_h\left(q_l - \gamma\right)$, where $k_h(\cdot) = k\left(\cdot/h\right)/h$ is a rescaled kernel density, and h is the bandwidth. Due to the weighted average nature of kernel smoothers, $\widehat{\Delta}(\gamma)$ would be near zero if there were no jump at γ. Otherwise, the difference would be near the magnitude of the jump $\Delta_0 = (1, \gamma_0)\,\delta_0$ which is assumed to be nonzero. This difference ensures that the estimator $\widehat{\gamma}_{DKE}$ is consistent. Porter and Yu (2011) have recently shown that $\widehat{\gamma}_{DKE}$ converges at rate n and the asymptotic distribution is related to a compound Poisson process. This limit theory is explained by interpreting γ as a 'middle' boundary point of q (see Yu, 2012a). For boundary point estimation, it is well-known that only data in an $O\left(n^{-1}\right)$ neighborhood is informative, so the h neighborhood in the construction of the DKE is typically large enough to ensure the n-consistency of $\widehat{\gamma}_{DKE}$. Given $\widehat{\gamma}_{DKE}$, the literature has also considered the estimation of the jump magnitude Δ_0. But no estimator of δ_0 is presently available.

When there are additional covariates, Delgado and Hidalgo (2000) suggested that the DKE continue to

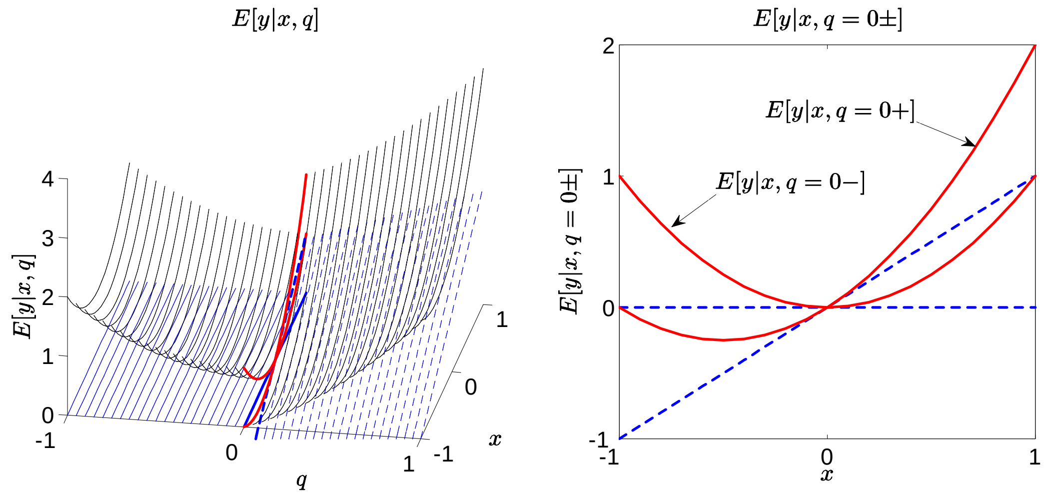

Figure 1: $E[y|x,q]$ and $E[y|x,q=\gamma_0\pm]$ With Endogeneity: The Blue Lines Represent the Case Without Endogeneity

be used to estimate γ.[3] In this case, the procedure can be employed by fixing some point (say x_o) in the support of x and redefining $\widehat{\Delta}(\gamma)$ as $\widehat{\mathbb{E}}[y|x_o, q=\gamma-]-\widehat{\mathbb{E}}[y|x_o, q=\gamma+]$, where $\widehat{\mathbb{E}}[y|x_o, q=\gamma\pm]$ is an estimator of the conditional mean of y given $x=x_o$ and $q=\gamma\pm$. The objective function converges to zero when $\gamma\neq\gamma_0$, and to $\Delta_o^2=(\mathbb{E}[y|x_o,q=\gamma_0-]-\mathbb{E}[y|x_o,q=\gamma_0+])^2$ when $\gamma=\gamma_0$, so $\widehat{\gamma}_{DKE}$ is consistent if $\Delta_o\neq 0$.

There are several difficulties in applying the DKE in this way. First, the selection of x_o raises difficulties, as shown in the following example. Suppose $y=(x+q)\,1\,(q>\gamma)+\varepsilon$, where $\gamma_0=0$, the supports of x and q are both $[-1,1]$, and endogeneity takes the form $\mathbb{E}\,[\varepsilon|x,q]=x^2+q^2$. Figure 1 shows $\mathbb{E}[y|x,q]$ and $\mathbb{E}[y|x,q=\gamma_0\pm]$. To identify γ successfully we need to select x_o so that Δ_o^2 is large, which means that x_o should be on the boundary of x's support. On the other hand, we also need $f_{x|q}(x_o|\gamma_0)$ to be large so that there is sufficient data to identify γ. When the density of $f_{x|q}(x|\gamma_0)$ takes on a bell shape, as in a typical case, x_o should ideally be in the middle of x's support on this criterion. Hence, the selection of x_o poses a dilemma and a potential tradeoff that is presently unresolved from both theory and practical perspectives. Second, consistency of $\widehat{\gamma}_{DKE}$ requires that $\Delta_o\neq 0$, but Δ_o can be 0 as shown in the example of Figure 1. Delgado and Hidalgo (2000) apply the DKE to estimate γ, assuming that $\left(\delta_{x0}',\delta_{q0}\right)'=\mathbf{0}$ and $\delta_{\alpha0}\neq 0$ so that $\Delta_o=\delta_{\alpha0}\neq 0$ does not depend on the choice of x_o. Moreover, their kernel function uses data in the neighborhood of $q=\gamma_0$ inefficiently, so that the convergence rate of $\widehat{\gamma}_{DK\mathbb{E}}$ is quite slow, as discussed further in Section 2.3 below. Furthermore, given $\widehat{\gamma}_{DKE}$, the induced estimator of $\delta_{\alpha0}$ uses only data in the neighborhood of $(x_o',\widehat{\gamma}_{DKE})'$, so the convergence rate of $\widehat{\delta}_{\alpha,DKE}$ is also very slow, especially when the dimension of x is large.

2.2 Construction of the IDKE of γ

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

Definition: $k_h(\cdot,\cdot)$ *is called a* univariate generalized kernel function of order p *if* $k_h(u,t)=0$ *if* $u>t$ *or*

[3] One might consider neglecting the data of x, and using only the data of q and y to estimate γ. This will generate the DKE of Porter and Yu (2010). Now, the jump size $E[\mathbf{x}'\delta|q=\gamma_0]$ is an average of the jumps at all x values, so may be zero or small, which results in identification failure or weak identification. Even if $E[\mathbf{x}'\delta|q=\gamma_0]$ is large, this DKE might be less efficient than the IDKE in the next subsection because the jump information at γ_0 is not fully explored; see footnote 7 for further analysis.

$u < t-1$ *and for all* $t \in [0,1]$,

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

A popular example of a generalized kernel function is as follows. Define

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

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

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

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

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

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

(iv) implies that $k_-(\cdot,r)$ is Lipschitz on $(-\infty,r]$ and $k_+(\cdot,r)$ is Lipschitz on $[r,\infty)$. This assumption is important in deriving the asymptotic distribution of the IDKE of γ; see Section 4.2.2 of Porter and Yu (2011) 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: $(y,x',q)' \in \mathbb{R} \times \mathcal{X} \times \mathcal{Q} \subset \mathbb{R}^{d+1}$, $\mathcal{X} = [0,1]^{d-1}$, $\mathcal{Q} = [\underline{q}, \overline{q}]$, and $\gamma \in \Gamma = [\underline{\gamma}, \overline{\gamma}] \subset \mathcal{Q}$, $\delta \in \Lambda \subset \mathbb{R}^{d+1}$, where $\underline{q}$ can be $-\infty$ and $\overline{q}$ can be ∞, and Γ and Λ are compact.

Since δ_0 is assumed to be fixed, we work with the discontinuous threshold regression of Chan (1993) instead of the small-threshold-effect framework of Hansen (2000). We do not restrict $\delta_0 \neq 0$ in Assumption S, where $\neq$ here means that at least one element is unequal; an enhanced restriction on δ_0 is imposed in Assumption I of Section 2.3 below. 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 much additional notation. 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 moments of x.

Define

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

and

$$k_h(u,t) = \begin{cases} \frac{1}{h} k\left(\frac{u}{h}\right), & \text{if } h \le t \le 1-h, \\ \frac{1}{h} k_+\left(\frac{u}{h}, \frac{t}{h}\right), & \text{if } 0 \le t \le h, \\ \frac{1}{h} k_-\left(\frac{u}{h}, \frac{1-t}{h}\right), & \text{if } 1-h \le t \le 1. \end{cases} \quad (5)$$

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 to estimate γ.[4] Formally, we require

Assumption K: $k_h(u,t)$ takes the form of (5) with $p=1$ and $k_+(0)=k_-(0)>0$.

The condition $k_+(0)=k_-(0)>0$ differs from that in Delgado and Hidalgo (2000). The following subsection discusses the impact of this condition on the asymptotic distributions of estimators of γ.

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

$$\begin{aligned}\widehat{\gamma} &= \arg\max_{\gamma} \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{n-1}\sum_{j=1,j\neq i}^{n} y_j K_{h,ij}^{\gamma-} - \frac{1}{n-1}\sum_{j=1,j\neq i}^{n} y_j K_{h,ij}^{\gamma+}\right]^2 \qquad (6)\\ &\equiv \arg\max_{\gamma} \frac{1}{n}\sum_{i=1}^{n}\widehat{\Delta}_i^2(\gamma) \equiv \arg\max_{\gamma} \widehat{Q}_n(\gamma),\end{aligned}$$

where

$$\begin{aligned}K_{h,ij}^{\gamma+} &= \prod\nolimits_{l=1}^{d-1} k_h(x_{lj}-x_{li},x_{li})\cdot k_h^{+}(q_j-\gamma) \equiv K_{h,ij}^{x} k_h^{+}(q_j-\gamma),\\ K_{h,ij}^{\gamma-} &= \prod\nolimits_{l=1}^{d-1} k_h((x_{lj}-x_{li},x_{li})\cdot k_h^{-}(q_j-\gamma) \equiv K_{h,ij}^{x} k_h^{-}(q_j-\gamma),\end{aligned}$$

with

$$K_{h,ij}^{x} = \prod\nolimits_{l=1}^{d-1} k_h(x_{lj}-x_{li},x_{li}) \equiv K_h^x(x_j-x_i,x_i) \equiv \frac{1}{h^{d-1}}K^x\left(\frac{x_j-x_i}{h},x_i\right).$$

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. From Yu (2008), it is known that to find $\widehat{\gamma}$ we need only check the middle points of the contiguous q_i's in the optimization process. In other words, the argmax operator (or argmin operator in Theorem 1 which gives the asymptotic distribution of $\widehat{\gamma}$) is a middle-point operator. The summation in the parenthesis of (6) excludes $j=i$, which is a standard strategy in converting a V-statistic to a U-statistic. Also, the normalization factor $\sum_{j=1,j\neq i}^{n} K_{h,ij}^{\gamma\pm}$ does not appear in the construction of $\widehat{\gamma}$, thereby avoiding random denominator issues and simplifying the derivation of the limit distribution of $\widehat{\gamma}$, a technique that dates back at least to Powell et al. (1989). This form of $\widehat{\gamma}$ has some practical advantages especially when d is large. Since the conditional mean is estimated at the boundary point $q=\gamma$, the local linear smoother (LLS) might be considered to ameliorate bias. However, when d is large, there are not many data points in a h neighborhood of $(x_i',\gamma)'$. As a result, not only does the LLS lose degrees of freedom (by estimating more parameters) but its denominator matrix tends to be close to singular. Of course, use of the LLS does not affect the asymptotic distribution of $\widehat{\gamma}$ in an essential way, and only has higher-order effects on the distribution of $\widehat{\gamma}$.

The objective function in (6) may be viewed as a nonparametric extension of the objective function of the parametric LSE of γ. With some preliminary algebra, it can be shown that the parametric LSE of γ satisfies

$$\widehat{\gamma}_{LS\mathbb{E}}^{P} = \arg\max_{\gamma}\left(\widehat{\delta}'\mathbf{X}'\right)\left[\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}_{>\gamma}'\mathbf{X}_{>\gamma}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}_{\leq\gamma}'\mathbf{X}_{\leq\gamma}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\left(\mathbf{X}\widehat{\delta}\right),$$

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

where $\widehat{\delta}$ is the LSE of δ based on the splitting of γ, and $\mathbf{X}$, $\mathbf{X}_{\leq\gamma}$ and $\mathbf{X}_{>\gamma}$ are $n \times (d+1)$ matrices that stack the vectors $\mathbf{x}_i'$, $\mathbf{x}_i' 1(q_i \leq \gamma)$ and $\mathbf{x}_i' 1(q_i \leq \gamma)$, respectively. The objective function of $\widehat{\gamma}_{LSE}^P$ uses the weighted average form of $\mathbf{X}\widehat{\delta}$ which is the conditional mean differences at all $\mathbf{x}_i$'s.[5] The weights in (6) are essentially given by $f(x_i, \gamma)$ (the probability limit of $n^{-1} \sum_{j=1, j\neq i}^{n} K_{h,ij}^{\gamma\pm}$), so that greater weight is placed on the conditional mean difference when there is more data around $(x_i', \gamma)'$. This weighting scheme is intuitively appealing for estimating the threshold parameter γ.

2.3 Limit Theory for IDKE and DKE

We start with some intuitive discussion on the validity of $\widehat{\gamma}$. 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 Lipschitz and satisfies $0 < \underline{f} \leq f(x,q) \leq \overline{f} < \infty$ for $(x,q) \in \mathcal{X} \times \Gamma_\epsilon$, where $\Gamma_\epsilon \equiv \left(\underline{\gamma} - \epsilon, \overline{\gamma} + \epsilon\right)$ for some $\epsilon > 0$ and some fixed quantities $(\underline{f}, \overline{f})$.
Assumption G: $g(x,q)$ is Lipschitz on $\mathcal{X} \times \Gamma_\epsilon$.

Assumption F implies that $0 < \underline{f}_q \leq f_q(\gamma) \leq \overline{f}_q < \infty$ for $\gamma \in \Gamma_\epsilon$ and fixed $\left(\underline{f}_q, \overline{f}_q\right)$, and the conditional density $f_{x|q}(x|q)$ is bounded below and above for $(x,q) \in \mathcal{X} \times \Gamma_\epsilon$. 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 γ_0 is not on the boundary of $\mathcal{Q}$. Under these two assumptions, we expect the objective function $\widehat{Q}_n(\gamma)$ to converge to

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

Since $f(x)$ and $f(x,\gamma)$ are continuous in x and γ, there will be a jump in the limit only if $\gamma = \gamma_0$ which provides identifying information. As a result, the threshold point can be identified and consistently estimated by maximizing $\widehat{Q}_n(\gamma)$. 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, use of all the data ensures that the IDKE has greater identifying capability than the DKE. 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 γ_0.

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

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

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

$$\begin{aligned} \overline{z}_{1i} &= \left[2\,(1, x_i', \gamma_0)\,\delta_0 e_i + \delta_0'\,(1, x_i', \gamma_0)'\,(1, x_i', \gamma_0)\,\delta_0\right] f(x_i, \gamma_0) f(x_i), \\ \overline{z}_{2i} &= \left[-2\,(1, x_i', \gamma_0)\,\delta_0 e_i + \delta_0'\,(1, x_i', \gamma_0)'\,(1, x_i', \gamma_0)\,\delta_0\right] f(x_i, \gamma_0) f(x_i). \end{aligned}$$

[5] To show the weights more clearly, let $\mathbf{x} = 1$. Then the objective function is equivalent to $\widehat{\delta}\left(\frac{n_1}{n} \cdot \frac{n_2}{n}\right)\widehat{\delta}$, where $n_1 = \sum_{i=1}^{n} 1(q_i \leq \gamma)$, and $n_2 = n - n_1$. If $\mathbf{x} = x$, then the weights are $\frac{\sum_{i=1}^{n} x_i^2 1(q_i \leq \gamma)}{\sum_{i=1}^{n} x_i^2} \frac{\sum_{i=1}^{n} x_i^2 1(q_i > \gamma)}{\sum_{i=1}^{n} x_i^2} \frac{\mathbf{X}\mathbf{X}'}{\sum_{i=1}^{n} x_i^2}$, where $\mathbf{X} = (x_1, \cdots, x_n)'$.

Here, $\overline{z}_{1i}$ represents the effect on $\widehat{Q}_n(\gamma)$ when the threshold point is displaced on the left of γ_0, and $\overline{z}_{2i}$ represents the converse. If we assume $f(e|x,q)$ is continuous in x and q, then $\overline{z}_{\ell i}$ and q_i have a continuous joint density $f_{\overline{z}_\ell,q}(\overline{z}_\ell,q)$. We now define $z_{1i}=\lim_{\Delta\uparrow 0}\overline{z}_{1i}1\{\gamma_0+\Delta<q_i\leq\gamma_0\}$, the limiting conditional value of $\overline{z}_{1i}$ given $\gamma_0+\Delta<q_i\leq\gamma_0$, $\Delta<0$ with $\Delta\uparrow 0$, and $z_{2i}=\lim_{\Delta\downarrow 0}\overline{z}_{2i}1\{\gamma_0<q_i\leq\gamma_0+\Delta\}$, the limiting conditional value of $\overline{z}_{2i}$ given $\gamma_0<q_i\leq\gamma_0+\Delta$, $\Delta>0$ with $\Delta\downarrow 0$. It follows that the density of the quantity $z_{\ell i}$ is $f_{\overline{z}_\ell,q}(z_\ell,\gamma_0)/f_q(\gamma_0)$, the conditional density of $\overline{z}_\ell$ given $q=\gamma_0$. The following assumption allows $f(e|x,q)$ to be discontinuous at $q=\gamma_0$.

Assumption E:

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

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

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

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

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

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

Theorem 1 *Under Assumptions E, F, G, H, I, K and S,*

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

where

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

is a cadlag process with $D(0)=0$, $\{z_{1i},z_{2i}\}_{i\geq 1}$, $N_1(\cdot)$ and $N_2(\cdot)$ are independent of each other, and $N_\ell(\cdot)$ is a Poisson process with intensity $f_q(\gamma_0)$.

The intuition for the rate n consistency of $\widehat{\gamma}$ is similar to that given in Porter and Yu (2011) where the DKE is considered and q is the only covariate. If we neglect the factor $f(x_i,\gamma_0)f(x_i)$ in $z_{\ell i}$, the asymptotic distribution is the same as that of the LSE in the parametric model, see Section 4.1 of Yu (2008). The factor $f(x_i,\gamma_0)$ appears in the limit theory because the random denominator in the kernel has been eliminated in estimating the jumps of $\mathbb{E}[y|x,q]$; see (6). If the LLS is used in the construction of $\widehat{\gamma}$, the factor $f(x_i,\gamma_0)$ will not appear. The factor $f(x_i)$ appears because the summation in (6) is over all the x_i's, and the U-statistic projection generates the marginal density of x.

We remark that this theorem is relevant in very general frameworks. For example, it applies irrespective of whether q is endogenous. It also applies to nonparametric threshold regression with endogeneity and nonadditive errors, that is modifying (1) to

$$y=g_1(x,q,\varepsilon_1)1(q\leq\gamma)+g_2(x,q,\varepsilon_2)1(q>\gamma),$$

where g_1 and g_2 are different smooth functions and ε_1 and ε_2 are error terms with $\mathbb{E}[\varepsilon_\ell|x,q]\neq 0$. The only difference in the asymptotic distribution in this case is that the jump size at $(x_i',\gamma_0)'$ in $\overline{z}_{\ell i}$ changes from

$(1, x_i', \gamma_0)\,\delta_0$ to the corresponding nonparametric form $\mathbb{E}[g_1(x_i, q_i, \varepsilon_{1i})|x_i, q_i = \gamma_0] - \mathbb{E}[g_2(x_i, q_i, \varepsilon_{2i})|x_i, q_i = \gamma_0]$.

For comparison, we state the following corollary for the asymptotic distribution of the DKE

$$\widetilde{\gamma} = \arg\max_{\gamma} \widehat{\Delta}_o^2(\gamma),$$

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

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

and where x_o is some fixed point in the interior of $\mathcal{X}$. For ease of expression in the following corollary, define $K(u_x) = \prod_{l=1}^{d-1} k(u_{x_l})$.

Corollary 1 *Suppose $(1, x_o', \gamma_0)\,\delta_0 \neq 0$ and $d > 1$. Then, under the same assumptions as in Theorem 1,*

$$nh^{d-1}(\widetilde{\gamma} - \gamma_0) \xrightarrow{d} \arg\min_{v} D(v),$$

where

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

is a cadlag process with $D(0) = 0$, $z_{1i} = \left[2(1, x_o', \gamma_0)\,\delta_0 e_i^- + \delta_0'(1, x_o', \gamma_0)'(1, x_o', \gamma_0)\,\delta_0\right] K(U_i^-)$ with e_i^- following the conditional distribution of e_i given $x_i = x_o$ and $q_i = \gamma_0-$ and U_i^- following the uniform distribution on the support of $K(\cdot)$, $z_{2i} = \left[-2(1, x_o', \gamma_0)\,\delta_0 e_i^+ + \delta_0'(1, x_o', \gamma_0)'(1, x_o', \gamma_0)\,\delta_0\right] K(U_i^+)$ with e_i^+ following the conditional distribution of e_i given $x_i = x_o$ and $q_i = \gamma_0+$ and U_i^+ following the same distribution as U_i^-, $\left\{e_i^-, e_i^+, U_i^-, U_i^+\right\}_{i \ge 1}$, $N_1(\cdot)$ and $N_2(\cdot)$ are independent of each other, and $N_\ell(\cdot)$ is a Poisson process with intensity $2^{d-1} f(x_o, \gamma_0)$.

When $d > 1$, the convergence rate of $\widetilde{\gamma}$ is slower than n although its asymptotic distribution is still related to the compound Poisson process. This is because less data is used in the estimation of γ. Nevertheless, the convergence rate is still faster than that of Delgado and Hidalgo (2000). In their setup in terms of the DKE, $k_+(0) = k_-(0) = 0$,[6] so that data in the neighborhood of γ_0 are not used in estimating γ_0. Their convergence rate is $\sqrt{nh^{d-2}}$ and the relative rate $\sqrt{nh^{d-2}}/nh^{d-1} = 1/\sqrt{nh^d} \to 0$. Compared to the asymptotic distribution of $\widehat{\gamma}$, x_i in $\overline{z}_{\ell i}$ is changed to x_o, the distribution of e_i is conditional on $x_i = x_o$ and $q_i = \gamma_0$ rather than only on $q_i = \gamma_0$, and the intensity of $N_\ell(\cdot)$ is related to $f(x_o, \gamma_0)$ rather than $f_q(\gamma_0)$. Those changes occur because only data in the neighborhood of x_o is used to estimate the threshold point. The appearance of $U_i^\pm$ in $z_{\ell i}$ may at first appear mysterious. But note that the conditional distribution of $(x_i - x_o)/h$ given that it falls in the support of $K(\cdot)$ converges to a uniform distribution, which leads directly to the presence of $U_i^\pm$ in $z_{\ell i}$. The factor 2^{d-1} in the intensity of $N_\ell(\cdot)$ measures the volume of the support of $K(\cdot)$. When the support of $K(\cdot)$ is large, more data is used in estimation and the intensity is larger. However, use of $K(\cdot)$ with a larger support may not add efficiency to $\widetilde{\gamma}$ since $K(U_i^\pm)$ in $z_{\ell i}$ tends to be smaller. To consider a simpler form of the limit process $D(v)$, let $K(\cdot)$ be a uniform kernel on $[-1/2, 1/2]^{d-1}$, in which case both $K(U_i^\pm)$ in $z_{\ell i}$ and 2^{d-1} in the intensity of $N_\ell(\cdot)$ disappear.

[6] This assumption guarantees that the DKE is asymptotically normally distributed. Moreover, the convergence rate requires further conditions on the derivatives that $k_+'(0) > 0$ and $k_-'(0) < 0$. Otherwise, the convergence rate is even slower.

When $d=1$ (that is when there are no other covariates except q), Porter and Yu (2011) derive the asymptotic distribution of the DKE. In that case, the convergence rate is $nh^{d-1}=n$, $z_{1i}=2\left(1,\gamma_0\right)\delta_0 e_i^- + \delta_0'\left(1,\gamma_0\right)'\left(1,\gamma_0\right)\delta_0$ with e_i^- following the conditional distribution of e_i given $q_i=\gamma_0-$, $z_{2i}=-2\left(1,\gamma_0\right)\delta_0 e_i^+ + \delta_0'\left(1,\gamma_0\right)'\left(1,\gamma_0\right)\delta_0$ with e_i^+ following the conditional distribution of e_i given $q_i=\gamma_0+$, and the intensity of $N_\ell(\cdot)$ is changed to $f_q(\gamma_0)$. This asymptotic distribution then matches both that of $\widetilde{\gamma}$ and $\widehat{\gamma}$ as $d=1$.[7]

2.4 Difficulties in Applying Two Alternative Estimators

It is known (Section 4.2.2 of Porter and Yu, 2011) that the DKE is asymptotically equivalent to the LSE and the PLE when q is a single covariate. In what follows, we define the LSE and the PLE when other covariates are present and discuss the difficulties that arise in deriving their asymptotic distributions.

Define the nonparametric LSE of γ in the general case as follows,

$$\widehat{\gamma}_{LSE}^N=\arg\min_\gamma \frac{1}{n}\sum_{i=1}^n\left[y_i\widehat{f}_i-\widehat{mf}_-^\gamma(x_i,q_i)1(q_i\le\gamma)-\widehat{mf}_+^\gamma(x_i,q_i)1(q_i>\gamma)\right]^2,$$

where $\widehat{f}_i=\frac{1}{n-1}\sum\limits_{j=1,j\neq i}^n K_{h,ij}$ with $K_{h,ij}=K_{h,ij}^x\cdot k_h(q_j-q_i)$ is the kernel estimator of $f_i\equiv f(x_i,q_i)$,

$$\begin{aligned}
\widehat{mf}_-^\gamma(x_i,q_i) &= \frac{1}{n-1}\sum_{j=1,j\neq i}^n y_jK_{h,ij}^x k_h^{\gamma-}(q_j-q_i,q_i),\\
\widehat{mf}_+^\gamma(x_i,q_i) &= \frac{1}{n-1}\sum_{j=1,j\neq i}^n y_jK_{h,ij}^x k_h^{\gamma+}(q_j-q_i,q_i),
\end{aligned}$$

with

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

In the construction of $\widehat{\gamma}_{LSE}^N$, we eliminate the random denominator as in $\widehat{\gamma}$. We next define the PLE as

$$\widehat{\gamma}_{PLE}=\arg\min_\gamma \frac{1}{n}\sum_{i=1}^n\left[y_i\widehat{f}_i-\mathbf{x}_i'\delta 1(q_i\le\gamma)\widehat{f}_i-\widehat{g}_f(x_i,q_i;\delta,\gamma)\right]^2, \tag{7}$$

where

$$\widehat{g}_f(x_i,q_i;\delta,\gamma)=\frac{1}{n-1}\sum_{j=1,j\neq i}^n\left(y_j-\mathbf{x}_j'\delta 1(q_j\le\gamma)\right)K_{h,ij}.$$

This density-weighted objective function of the PLE was suggested in Li (1996) without considering the threshold effects. In both the LSE and the PLE, γ is estimated by finding the best fit between y_i and an

[7] In (2), if we neglect the data of x, the relationship between y and q is $y=E[g(x,q)|q]+E[\mathbf{x}'\delta|q]1(q\le\gamma)+v$ with $v=e+g(x,q)-E[g(x,q)|q]+\left(\mathbf{x}'\delta-E[\mathbf{x}'\delta|q]\right)1(q\le\gamma)$ satisfying $E[v|q]=0$. From Porter and Yu (2011), in the limit distribution of the DKE, $z_{1i}=2E[\mathbf{x}'\delta|q=\gamma_0]v_i^-+\left(E[\mathbf{x}'\delta|q=\gamma_0]\right)^2$ and $z_{2i}=-2E[\mathbf{x}'\delta|q=\gamma_0]v_i^++\left(E[\mathbf{x}'\delta|q=\gamma_0]\right)^2$ with $v_i^\pm$ similarly defined as $e_i^\pm$, so $E[z_{\ell i}]=\left(E[\mathbf{x}'\delta|q=\gamma_0]\right)^2$. On the other hand, if we neglect $f(x_i,\gamma_0)f(x_i)$ in the limit distribution of the IDKE, $E[z_{\ell i}]=E[(\mathbf{x}'\delta)^2|q=\gamma_0]\ge\left(E[\mathbf{x}'\delta|q=\gamma_0]\right)^2$, i.e., the average jump size in $D(\cdot)$ of the IDKE is larger than that in $D(\cdot)$ of the DKE, which indicates that the IDKE is more efficient than the DKE.

estimator of $\mathbb{E}[y_i|x_i, q_i]$; the difference lies in that different estimators of $\mathbb{E}[y_i|x_i, q_i]$ are used.

The objective function of the IDKE is superior to that of the LSE and the PLE in two respects. First, according to Yu (2008, 2012a), only the information around the threshold point is informative for γ_0, so $\widehat{\Delta}_i(\gamma)$ in the objective function of the IDKE is constructed using only data in the neighborhood of γ. In contrast, the objective functions of the LSE and the PLE use information in other areas, and the resulting biases need to be handled carefully. The objective function of the IDKE therefore takes advantage of its local construction, whereas the global objective functions of the LSE and the PLE are influenced by the effects of information throughout the distribution. Second, since $\widehat{\Delta}_i(\gamma)$ in the objective function of the IDKE is linear in $k_+\left(\frac{q_j-\gamma}{h}\right)$ and $k_-\left(\frac{q_j-\gamma}{h}\right)$, it is easy to localize in the neighborhood of γ, which is key to deriving the convergence rate and the asymptotic distribution of $\widehat{\gamma}$. However, the objective functions of the LSE and PLE are complicated nonlinear functions of γ, which makes localization extremely hard. In addition, the objective function of the IDKE does not rely on the assumption that $\delta(x,q) = \mathbf{x}'\delta$, whereas that of the PLE does.

2.5 Estimation of δ

Given $\widehat{\gamma}$, we can estimate δ as if γ_0 were known. Due to the superconsistency of $\widehat{\gamma}$, the asymptotic distribution of our estimator $\widehat{\delta}$ should not be affected by the estimation of γ. In other words, the asymptotic distribution of $\widehat{\delta}$ is the same as when γ_0 is known. We provide two estimators of δ, both of which are based on the observation that

$$m_-(x) - m_+(x) \equiv \mathbb{E}[y|x, q = \gamma_0-] - \mathbb{E}[y|x, q = \gamma_0+] = \delta_{\alpha 0} + x'\delta_{x0} + \gamma_0\delta_{q0}. \tag{8}$$

The first estimator of δ is the IDKE. From (8), δ_{x0} and δ_{q0} are the slope differences of $\mathbb{E}[y|x,q]$ at the left and right neighborhoods of $q = \gamma_0$, so $\delta_{xq0} \equiv (\delta_{x0}', \delta_{q0})'$ can be identified using

$$\widehat{\delta}_{xq} = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{q_i-\widehat{\gamma}}{h}\right)\left(\widehat{b}_-(x_i) - \widehat{b}_+(x_i)\right) \Big/ \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{q_i-\widehat{\gamma}}{h}\right),$$

where $\widehat{b}_\pm(x_i)$ is the local polynomial estimator (LPE) of $(\partial\mathbb{E}[y_i|x_i, q_i = \gamma_0\pm]/\partial x', \partial\mathbb{E}[y_i|x_i, q_i = \gamma_0\pm]/\partial q)'$. Also, from (8),

$$\delta_{\alpha 0} = m_-(x) - m_+(x) - (x', \gamma_0)\,\delta_{xq0}$$

at any x, so $\delta_{\alpha 0}$ can be identified using

$$\widehat{\delta}_\alpha = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{q_i-\widehat{\gamma}}{h}\right)\left[\widehat{a}_-(x_i) - \widehat{a}_+(x_i) - (x_i', \widehat{\gamma})\,\widehat{\delta}(x_i)\right] \Big/ \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{q_i-\widehat{\gamma}}{h}\right),$$

where $\widehat{a}_\pm(x_i)$ is the LPE of $m_\pm(x_i)$, and $\widehat{\delta}(x_i) = \widehat{b}_-(x_i) - \widehat{b}_+(x_i)$. To be specific, the LPE $\left(\widehat{a}_+(x_i), \widehat{b}_+(x_i)'\right)'$ is the first $(d+1)$ elements of the solution to

$$\min_\beta \sum_{j=1, j\neq i}^{n} \left[y_j - (x_j' - x_i', q_j - \widehat{\gamma})^{S_p}\beta\right]^2 K_{h,ij}^{\widehat{\gamma}+},$$

where for a row vector $\xi \in \mathbb{R}^d$, $\xi^{S_p} = (\xi^{S(\nu)})_{\nu\in\{0,\cdots,p\}}$ is a row vector, $\xi^{S(\nu)} = (\xi^s)_{|s|=\nu}$ is a row vector of length $(\nu+d-1)!/\nu!(d-1)!$, $s = (s_1, \cdots, s_d)$ is a vector with all its elements being nonnegative integers, the norm of s is defined as $|s| \equiv s_1 + \cdots s_d$, and $\xi^s = \xi_1^{s_1}\cdots\xi_d^{s_d}/(s_1!\cdots, s_d!)$. For convenience, we assume that

$\{(s_1, \cdots, s_d)\}$ in the definition of ξ^{S_p} are ordered lexicographically. $\left(\widehat{a}_-(x_i), \widehat{b}_-(x_i)'\right)'$ is similarly defined with $K_{h,ij}^{\widehat{\gamma}+}$ replaced by $K_{h,ij}^{\widehat{\gamma}-}$, where $K_{h,ij}^{\gamma\pm}$ is defined in (6).

If γ_0 were known, this model can also be treated as a regression discontinuity design with covariates. In this case, we are interested in the treatment effect at $q = \gamma_0$, say,

$$\Delta_0 = \mathbb{E}\left[m_-(x) - m_+(x)\right],$$

which can be estimated as

$$\widehat{\Delta} = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{q_i - \widehat{\gamma}}{h}\right)\left[\widehat{a}_-(x_i) - \widehat{a}_+(x_i)\right] \Big/ \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{q_i - \widehat{\gamma}}{h}\right).$$

From Theorem 3 of Heckman et al. (1998), $\widehat{a}_\pm(x_i)$ and $\widehat{b}_\pm(x_i)$ are asymptotically linear, so the numerators of $\widehat{\delta} = \left(\widehat{\delta}_\alpha, \widehat{\delta}_{xq}'\right)'$ and $\widehat{\Delta}$ are asymptotically U-statistics. To ensure the validity of the linear approximation, we need the following conditions which strengthen assumptions G and H.

Assumption G$'$: $g(x, q)$ is $(p+1)$-times continuously differentiable on $\mathcal{X} \times \Gamma_\epsilon$ with $p > d$.

Assumption H$'$: $h \to 0$, $\sqrt{nh}h \to \infty$, $\sqrt{nh}h^{p+1} \to C \in [0, \infty)$, and $\sqrt{n}h^d / \ln n \to \infty$.

Note from the remarks following Assumption H that $\sqrt{n}h^d / \ln n \to \infty$ assures $nh^d \to \infty$. Also $\sqrt{nh}h = \frac{\sqrt{n}h^d}{\ln n} h^{\frac{3}{2}-d} \ln n \to \infty$ when $\sqrt{n}h^d / \ln n \to \infty$ and $d \geq 2$.

The following theorem gives the asymptotic distribution of $\widehat{\delta}$. For convenience of exposition, we introduce some notation. Let M_o^+ be the square matrix of size $\sum_{\nu=0}^{p} (\nu + d - 1)!/\nu!(d-1)!$ with the l-th row, t-th column "block" being

$$\int_0^\infty \int \left(u_x', u_q\right)^{S(l)\prime} \left(u_x', u_q\right)^{S(t)} K\left(u_x\right) k_+(u_q) du_x du_q, 0 \leq l, t \leq p.$$

Let B^+ be the $\sum_{\nu=0}^{p} (\nu + d - 1)!/\nu!(d-1)!$ by $(p+d)!/(p+1)!(d-1)!$ matrix whose l-th block is

$$\int_0^\infty \int \left(u_x', u_q\right)^{S(l)\prime} \left(u_x', u_q\right)^{S(p+1)} K\left(u_x\right) k_+(u_q) du_x du_q,$$

and let M_o^- and B^- be similarly defined with $\int_0^\infty$ and k_+ in M_o^+ and B^+ being replaced by $\int_{-\infty}^0$ and k_- respectively. Further, let

$$C_l^+(v_q) = \int k\left(u_q\right) \mathbf{e}_l' \left(M_o^+\right)^{-1} \left[\left(u_x', v_q\right)^{S_p}\right]' K\left(u_x\right) du_x du_q,$$

where $\mathbf{e}_l$ is a $\sum_{\nu=0}^{p} (\nu + d - 1)!/\nu!(d-1)!$ by 1 vector with the lth element being 1 and all other elements being 0, $l = 1, \cdots, d+1$, and $C_l^-(v_q)$ be similarly defined with M_o^+ in $C_l^+(v_q)$ replaced by M_o^-.

$$C^+(x, v_q) = \int k\left(u_q\right) \left(x', \gamma_0\right) \left(\mathbf{0}, I_d, \mathbf{0}\right) \left(M_o^+\right)^{-1} \left[\left(u_x', v_q\right)^{S_p}\right]' K\left(u_x\right) du_x du_q,$$

where $(\mathbf{0}, I_d, \mathbf{0})$ is a $d \times \sum_{\nu=0}^{p} (\nu + d - 1)!/\nu!(d-1)!$ matrix with the first zero matrix being a column vector and I_d being an identity matrix of size d. $C^-(x, v_q)$ is similarly defined with M_o^- in $C^+(x, v_q)$ replaced by M_o^-.

$$\sigma_\pm^2(x) = \mathbb{E}[e^2 | x, q = \gamma_0 \pm].$$

$g^{(p+1)}(x, \gamma_0)$ is a $(p+d)!/(p+1)!(d-1)!$ by 1 vector of the $(p+1)$th-order partial derivatives of $g(x,q)$ with respect to $(x', q)'$ at $q = \gamma_0$, where the elements of $g^{(p+1)}(x,q)$ are ordered in the same way as $\{(s_1, \cdots, s_d)\}_{s \in S(p+1)}$.

Theorem 2 *Under Assumptions E, F, G′, H′, I, K, and S,*

$$\sqrt{nh}h\left(\widehat{\delta}_\alpha - \delta_{\alpha 0} + h^p \mathbb{E}\left[(x', \gamma_0)(\mathbf{0}, I_d, \mathbf{0})\left[\left(M_o^-\right)^{-1} B^- - \left(M_o^+\right)^{-1} B^+\right] g^{(p+1)}(x, \gamma_0)\Big| q = \gamma_0\right]\right) \xrightarrow{d} N(0, \Sigma_\alpha),$$
$$\sqrt{nh}h\left(\widehat{\delta}_{x_l} - \delta_{x_l 0} - h^p \mathbf{e}'_{l+1}\left[\left(M_o^-\right)^{-1} B^- - \left(M_o^+\right)^{-1} B^+\right] \mathbb{E}[g^{(p+1)}(x, \gamma_0)\Big| q = \gamma_0]\right) \xrightarrow{d} N(0, \Sigma_{x_l}),$$
$$\sqrt{nh}h\left(\widehat{\delta}_q - \delta_{q0} - h^p \mathbf{e}'_{d+1}\left[\left(M_o^-\right)^{-1} B^- - \left(M_o^+\right)^{-1} B^+\right] \mathbb{E}[g^{(p+1)}(x, \gamma_0)\Big| q = \gamma_0]\right) \xrightarrow{d} N(0, \Sigma_q),$$

for $l = 1, \cdots, d-1$, *where*

$$\Sigma_\alpha = \mathbb{E}\left[\int \left[k_+^2(v_q)\sigma_+^2(x) C^+(x, v_q)^2 + k_-^2(v_q)\sigma_-^2(x) C^-(x, v_q)^2\right] dv_q \Big| q = \gamma_0\right] \Big/ f_q(\gamma_0),$$
$$\Sigma_{x_l} = \mathbb{E}\left[\int \left[k_+^2(v_q)\sigma_+^2(x) C_{l+1}^+(v_q)^2 + k_-^2(v_q)\sigma_-^2(x) C_{l+1}^-(v_q)^2\right] dv_q \Big| q = \gamma_0\right] \Big/ f_q(\gamma_0),$$
$$\Sigma_q = \mathbb{E}\left[\int \left[k_+^2(v_q)\sigma_+^2(x) C_{d+1}^+(v_q)^2 + k_-^2(v_q)\sigma_-^2(x) C_{d+1}^-(v_q)^2\right] dv_q \Big| q = \gamma_0\right] \Big/ f_q(\gamma_0).$$

According to this theorem, the bias and variance of $\widehat{\delta}$ are the integrated bias and variance of $(\widehat{a}_-(x_i) - \widehat{a}_+(x_i) - (x_i', \widehat{\gamma})\widehat{\delta}(x_i), \widehat{b}_-(x_i)' - \widehat{b}_+(x_i)')'$ for x_i in the neighborhood of $q = \gamma_0$. As shown in the proof, the convergence rate of $\widehat{\Delta}$ is $\sqrt{nh}$. Since $\widehat{\delta}_\alpha$ is based on $\delta_{\alpha 0} = m_-(x) - m_+(x) - (x', \gamma_0)\delta_{xq0}$, the slower convergence rate of $\widehat{\delta}_{xq}$ contaminates the convergence rate of $\widehat{\delta}_\alpha$. The theorem implies that the estimation of δ does not suffer the curse of dimensionality since the convergence rate is the same as the nonparametric slope estimator with a single covariate. This is understandable as all data in the h neighborhood of $q = \gamma_0$, or $O(nh)$ data points, are used in estimation.

For completeness, we state the asymptotic distribution of $\widehat{\Delta}$ in the following corollary. For this purpose, we change Assumption H′ to

Assumption H′′: $h \to 0$, $\sqrt{nh}h^{p+1} \to C \in [0, \infty)$, and $\sqrt{n}h^d / \ln n \to \infty$.

Compared with Assumption H′, Assumption H′′ neglects $\sqrt{nh}h \to \infty$. We need $nh \to \infty$ in the following corollary, but it is implied by $\sqrt{n}h^d / \ln n \to \infty$ as $d \geq 1$.

Corollary 2 *Under Assumptions E, F, G′, H′′, I, K, and S,*

$$\sqrt{nh}\left(\widehat{\Delta} - \Delta_0 - B_\Delta\right) \xrightarrow{d} N(0, \Sigma_\Delta),$$

where

$$\begin{aligned} B_\Delta = {} & h^{p+1}\mathbf{e}_1'\left[\left(M_o^-\right)^{-1} B_- - \left(M_o^+\right)^{-1} B^+\right] \mathbb{E}[g^{(p+1)}(x, \gamma_0)\Big| q = \gamma_0] \\ & + \sum\nolimits_{l=1}^{p+1} \frac{h^l}{l!}\left[\int k(v_q) {v_q}^l dv_q\right] \int (m_-(x) - m_+(x)) \frac{f_\gamma^{(l)}(x, \gamma_0)}{f_q(\gamma_0)} dx \\ & - \Delta_0 \sum\nolimits_{l=1}^{p+1} \frac{h^l}{l!}\left[\int k(v_q) {v_q}^l dv_q\right] \frac{f_\gamma^{(l)}(\gamma_0)}{f_q(\gamma_0)}, \end{aligned}$$

and

$$\Sigma_\Delta = \mathbb{E}\left[\int \left[k_+^2(v_q)\,\sigma_+^2(x)C_1^+(v_q)^2 + k_-^2(v_q)\,\sigma_-^2(x)C_1^-(v_q)^2\right]dv_q \middle| q=\gamma_0\right] \Big/ f_q(\gamma_0)$$
$$+\int k(v_q)^2\,dv_q\left(\mathbb{E}[(m_-(x)-m_+(x))^2\,|q=\gamma_0]-\Delta_0^2\right)\Big/ f_q(\gamma_0),$$

with $f_\gamma^{(l)}(x,\gamma_0)$ *being the* l*th order partial derivative of* $f(x,q)$ *with respect to* q *evaluated at* $q=\gamma_0$*, and* $f_\gamma^{(l)}(\gamma_0)$ *being the* l*th order derivative of* $f_q(\gamma)$ *with respect to* γ *evaluated at* $\gamma=\gamma_0$*.*

The convergence rate of the DKE of Δ_0 in Delgado and Hidalgo (2000) is $\sqrt{nh^d}$, which is much slower than $\sqrt{nh}$ especially when d is large. This is because we integrate the information of jumps at all the x_i's whereas the DKE uses only the information of the jump at some fixed x_o. Compared with $\widehat{\delta}$, the asymptotic bias and variance of $\widehat{\Delta}$ is a little more complicated. This is because

$$\sqrt{nh}\left(\widehat{\Delta}-\Delta_0\right)=\frac{\sqrt{nh}\left(\widehat{\Delta}_N-\overline{\Delta}_N\right)+\sqrt{nh}\left(\overline{\Delta}_N-\Delta_0\right)}{\widehat{f}_q(\widehat{\gamma})}.$$

where $\widehat{\Delta}_N$ is the numerator of $\widehat{\Delta}$,

$$\overline{\Delta}_N=\frac{1}{nh}\sum_{i=1}^n k\left(\frac{q_i-\widehat{\gamma}}{h}\right)(m_-(x_i)-m_+(x_i)),$$

and $\widehat{f}_q(\widehat{\gamma})=\frac{1}{nh}\sum_{i=1}^n k\left(\frac{q_i-\widehat{\gamma}}{h}\right)$. As a result, $\overline{\Delta}_N$ and $\widehat{f}_q(\widehat{\gamma})$ will also contribute to the asymptotic distribution of $\sqrt{nh}\left(\widehat{\Delta}-\Delta_0\right)$. The three terms of B_Δ are attributed to $\widehat{\Delta}_N-\overline{\Delta}_N$, $\overline{\Delta}_N-\Delta_0$ and $\widehat{f}_q(\widehat{\gamma})$, respectively. The first term of Σ_Δ is attributed to $\widehat{\Delta}_N-\overline{\Delta}_N$, and the second term is attributed to $\overline{\Delta}_N-\Delta_0$ and $\widehat{f}_q(\widehat{\gamma})$. The convergence rate of $\widehat{\Delta}$ is $\sqrt{nh}$ as expected, but its bias is $O(h)$. This large bias is due to $\overline{\Delta}_N-\Delta_0$ and $\widehat{f}_q(\widehat{\gamma})$. In the local linear case, i.e., $p=1$, Frölich (2010) suggests using a new kernel k^* in the construction of $\widehat{\Delta}$ to achieve a bias with rate $h^{p+1}=h^2$. This new kernel implicitly carries out a double boundary correction. Frölich considers the case with discontinuous $f(x,q)$ at $q=\gamma_0$. In our setup, a higher-order kernel $k(\cdot)$ in the construction of $\widehat{\Delta}$ can be used to achieve bias reduction.

The second estimator of δ is based on another implication of (8), namely that δ_0 is the coefficient from projecting $m_-(x)-m_+(x)$ on $\mathbf{x}$ in the neighborhood of $q=\gamma_0$. Empirically, we can project $\widehat{a}_-(x_i)-\widehat{a}_+(x_i)$ on $\mathbf{x}_i$ in a h neighborhood of $\widehat{\gamma}$ to estimate δ. However, $\widehat{a}_-(x)-\widehat{a}_+(x)$, as an estimate of $m_-(x)-m_+(x)$, is constructed at $q=\widehat{\gamma}$ so does not have variation in the direction of q. As a result, if we regress $\widehat{a}_-(x_i)-\widehat{a}_+(x_i)$ on $\mathbf{x}_i$ directly, the probability limit of the resulting estimator of δ_q is zero. To avoid this problem, we may regress $\widehat{a}_-(x_i)-\widehat{a}_+(x_i)$ only on $(1,x_i')'$. Specifically, define

$$(\overline{\delta}_\alpha,\widetilde{\delta}_x')'=\arg\min_{\underline{\delta}}\frac{1}{n}\sum_{i=1}^n k\left(\frac{q_i-\widehat{\gamma}}{h}\right)\left[\widehat{a}_-(x_i)-\widehat{a}_+(x_i)-(1,x_i')\,\underline{\delta}\right]^2. \tag{9}$$

Note that $\overline{\delta}_\alpha$ estimates $\delta_{\alpha0}+\gamma_0\delta_{q0}$, so we can estimate $\delta_{\alpha0}$ by

$$\widetilde{\delta}_\alpha=\overline{\delta}_\alpha-\widehat{\gamma}\widehat{\delta}_q,$$

where $\widehat{\delta}_q$ is the IDKE of δ_{q0}. Before stating the asymptotic distribution of $(\widetilde{\delta}_\alpha,\widetilde{\delta}_x')'$, we introduce some further notation. Define the $d \times d$ matrix

$$M = \begin{pmatrix} 1 & \mathbb{E}[x'|q=\gamma_0] \\ \mathbb{E}[x|q=\gamma_0] & \mathbb{E}\left[xx'|q=\gamma_0\right] \end{pmatrix},$$

and the (l,t) element of the $d \times d$ matrix Ψ as

$$\mathbb{E}\left[\overline{x}_l \overline{x}_t \int \left[k_+^2\left(v_q\right)\sigma_+^2(x)C_1^+(v_q)^2 + k_-^2\left(v_q\right)\sigma_-^2(x)C_1^-(v_q)^2\right] dv_q \right| q=\gamma_0\Bigg],$$

where $\overline{x}_l$ is the lth element of $(1,x')'$.

Theorem 3 *Under Assumptions E, F, G′, H″, I, K, and S,*

$$\sqrt{nh}\left(\widetilde{\delta}_{x_l} - \delta_{x_l0} - h^{p+1}\mathbf{e}_{l+1}'M^{-1}\mathbb{E}\left[(1,x')'\,\mathbf{e}_1\left[\left(M_o^-\right)^{-1}B^- - \left(M_o^+\right)^{-1}B^+\right]g^{(p+1)}(x,\gamma_0)\right|q=\gamma_0\right]\Bigg) \xrightarrow{d} N\left(0,\Omega_{x_l}\right)$$

for $l = 1,\cdots,d-1$, *where*

$$\Omega_{x_l} = \mathbf{e}_{l+1}'M^{-1}\Psi M^{-1}\mathbf{e}_{l+1}/\, f_q(\gamma_0).$$

When $\gamma_0 = 0$,

$$\sqrt{nh}\left(\widetilde{\delta}_{\alpha} - \delta_{\alpha0} - h^{p+1}\mathbf{e}_1'M^{-1}\mathbb{E}\left[(1,x')'\,\mathbf{e}_1\left[\left(M_o^-\right)^{-1}B^- - \left(M_o^+\right)^{-1}B^+\right]g^{(p+1)}(x,\gamma_0)\right|q=\gamma_0\right]\Bigg) \xrightarrow{d} N\left(0,\Omega_\alpha^{(1)}\right),$$

where

$$\Omega_\alpha^{(1)} = \mathbf{e}_1'M^{-1}\Psi M^{-1}\mathbf{e}_1/\, f_q(\gamma_0).$$

If Assumption H″ changes to H′ and $\gamma_0 \neq 0$, *then*

$$\sqrt{nh}h\left(\widetilde{\delta}_{\alpha} - \delta_{\alpha0} + h^p\gamma_0\mathbf{e}_{d+1}'\left[\left(M_o^-\right)^{-1}B^- - \left(M_o^+\right)^{-1}B^+\right]\mathbb{E}[g^{(p+1)}(x,\gamma_0)\big|\,q=\gamma_0]\right) \xrightarrow{d} N\left(0,\Omega_\alpha^{(2)}\right),$$

where

$$\Omega_\alpha^{(2)} = \gamma_0^2\Sigma_q$$

with Σ_q *defined in Theorem 2.*

Different from $\widehat{\delta}_{x_l}$, the convergence rate of $\widetilde{\delta}_{x_l}$ is $\sqrt{nh}$ rather than $\sqrt{nh}h$. Also, the convergence rate of $\widetilde{\delta}_\alpha$ depends on whether $\gamma_0 = 0$ or not. When $\gamma_0 = 0$, the convergence rate of $\widetilde{\delta}_\alpha$ is $\sqrt{nh}$ which differs from that of $\widehat{\delta}_\alpha$. When $\gamma_0 \neq 0$, the asymptotic distribution of $\widetilde{\delta}_\alpha$ is the same as $-\gamma_0\widehat{\delta}_q$, so the convergence rate is still $\sqrt{nh}h$. See Section 3.1 for more discussion on the differences between $\widehat{\delta}$ and $\widetilde{\delta}$. Finally, since consistent estimation of the biases and variances of the estimators of δ (which are necessary for statistical inference) is a standard econometric exercise, it is omitted here.

2.6 Intuition for the Identifiability of γ and δ

Although our analysis shows that γ and δ can be identified it may still appear mysterious that they are identifiable without instruments. An intuitive explanation is provided here. It is convenient to start by reviewing how instrumentation helps to identify a demand curve in classical simultaneous systems of supply and demand. We then explain how instrumentation is implicitly involved in the present threshold model setup.

Consider the following linear Marshallian stochastic demand/supply system

$$\begin{aligned} \text{Demand:} \quad & q_i = a + bp_i + u_i, \\ \text{Supply:} \quad & q_i = c + dp_i + v_i, \end{aligned}$$

where p_i and q_i are prices and quantities, respectively, u_i represents other factors that affect demand (such as income and consumer taste), v_i represents factors that affect supply (such as weather and union status), and a, b, c and d are parameters. It is well-known that a and b cannot be identified and are inconsistently estimated by least squares due to *simultaneous equations bias*. Conventionally, therefore, an explicit instrument z is introduced which shifts *only* the supply curve (e.g., weather conditions as in Angrist et al. (2000)) enabling equilibria to trace out the shape of the demand curve. This textbook argument is illustrated in the left panel of Figure 2. Given the linear structure of the demand curve, two values of z are enough to identify the whole straight line, which generates the famous *Wald estimator* (Wald, 1940).

If the system is nonparametric, e.g., the demand function takes the form of $q_i = g(p_i) + u_i$, then $g(\cdot)$ is generally considered to be much harder to identify due to the notorious ill-posed inverse problem. Most of the existing literature such as Newey et al. (1999), Ai and Chen (2003), Newey and Powell (2003), Hall and Horowitz (2005), and Darolles et al. (2011) use a nonparametric IV approach to help resolve this problem but with deleterious effects on the convergence rate; see Florens (2003) and Carraso et al. (2007) for a summary of the related literature. The nonparametric IV approach identifies $g(\cdot)$ *globally*, which means that some regularity conditions such as bounded supports and bounded densities on $(q_i, p_i)'$ are required to facilitate the theoretical development. Such regularities may not be innocuous in practice, as explained in Phillips and Su (2011). In contrast to the treatment of ill-posed inversion in nonparametric IV regression, Wang and Phillips (2009) and Phillips and Su (2011) show how the endogeneity problem may be resolved *locally* using characteristic nonstationary features of the data that implicitly provide instrumentation. That is, they show how to identify $g(\cdot)$ locally in some region of p where the data are informative. Intriguingly, when the system contains *local* shifters of the supply curve it transpires that no external instruments are required. In Wang and Phillips (2009), time series "nonstationarity" plays the role of the local shifter, and in Phillips and Su (2011), cross section locational shifts (such as geographical effects) play the same role. The middle panel of Figure 2 gives some graphical intuition exhibiting this identification scheme.

In threshold regression with endogeneity, the system contains a local shifter that helps to identify γ_0 in a similar fashion. This local shifter is the threshold indicator $1(q_i > \gamma)$, which plays a role analogous to the time series nonstationarity in Wang and Phillips (2009) and the location shifts in Phillips and Su (2011). The threshold indicator can identify γ_0 even in nonparametric threshold regression with endogeneity. To be explicit, suppose $y_i = g(q_i) + \varepsilon_i = g_1(q_i)1(q_i \leq \gamma) + g_2(q_i)1(q_i > \gamma) + \varepsilon_i$, where g_1 and g_2 are smooth functions with $g_1(\gamma_0) \neq g_2(\gamma_0)$, and $\mathbb{E}[\varepsilon|q] \neq 0$. For simplicity, we here neglect other covariates. In this setup, the objective function of the IDKE is equivalent to

$$\left| \frac{1}{n} \sum_{j=1}^{n} y_j k_h^+(q_j - \gamma) - \frac{1}{n} \sum_{j=1}^{n} y_j k_h^-(q_j - \gamma) \right|,$$

which is roughly

$$\left| \mathbb{E}[y1(q > \gamma)|q \in (\gamma - h, \gamma + h)] - \mathbb{E}[y1(q \leq \gamma)|q \in (\gamma - h, \gamma + h)] \right|.$$

In other words, we may use the indicator $1(q > \gamma)$ to shift y from the left neighborhood of γ to the right neighborhood, and check which shifter provides the largest variation in $\mathbb{E}[y]$. Carefully checking this objective

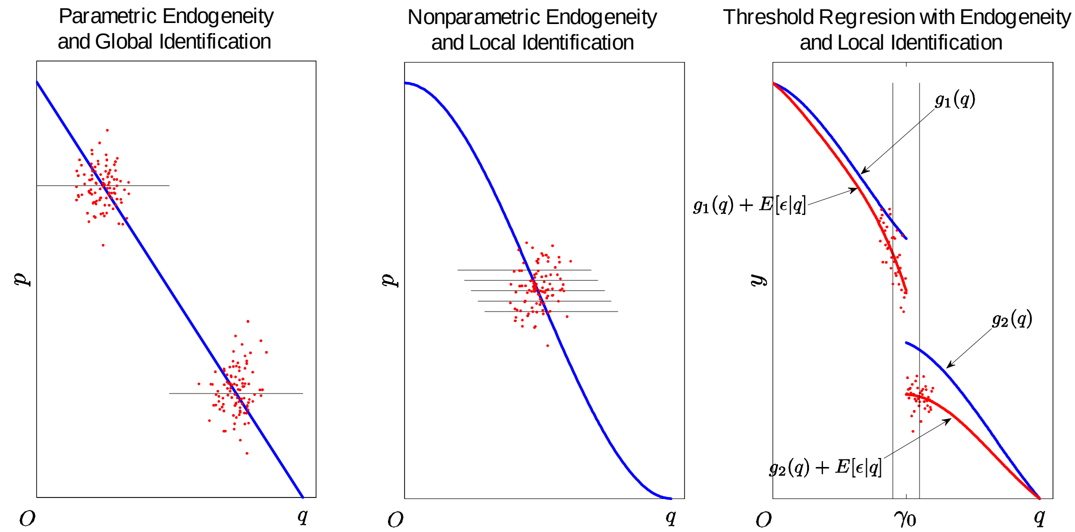

Figure 2: Graphical Intuition for the Identification of the Demand Curve under Endogeneity in Parametric, Nonparametric and Threshold Regression Models

function, we see that it is the numerator of the Wald estimator using only local-to-γ data.[8] In regression discontinuity designs (RDDs), Hahn et al. (2001) also find that the treatment effects estimator is numerically equivalent to the Wald estimator (see also Section 4.2 of Yu (2010) for an extensive discussion). However, the RDD literature concentrates on identifying the jump size, while we are interested in the jump location.[9] To identify the jump size $g_1(\gamma_0) - g_2(\gamma_0)$, we must assume $\mathbb{E}[\varepsilon|q]$ is continuous. This continuity assumption is key to identifying treatment effects in RDDs. In other words, the RDDs allow for endogeneity but require the endogeneity to be continuous (see Van der Klaauw (2002) for a convincing application with continuous endogeneity). In contrast, to identify the jump location, we do not need a continuity assumption as long as the discontinuity in endogeneity does not offset the original jump completely; see Section 5.1 for more discussions on this point. When there exist other covariates x_i, the local shifter $1(q_i > \gamma)$ is valid at any x_i, so integrating all the jump information can provide a stronger signal for the jump location. This integration is precisely what the IDKE seeks to accomplish.

To understand why the local shifter $1(q_i > \gamma_0)$ can identify the jump size, recall from Lee and Lemieux (2010) that this local shifter plays the role of *local randomization* if $\mathbb{E}[\varepsilon|q]$ is continuous. From Section II of Heckman (1996), randomization plays the role of balancing (rather than eliminating) endogeneity biases. In our setup, the bias $\mathbb{E}[\varepsilon|q = \gamma_0+]$ balances the bias $\mathbb{E}[\varepsilon|q = \gamma_0-]$, so the jump size can be identified even in the presence of endogeneity. However, as emphasized by Heckman, "structural parameters" such as $g_1(\cdot)$ and $g_2(\cdot)$ cannot be identified by this local randomization scheme without other instruments, which means that counterfactual analysis is hard in RDDs with endogeneity. When there are other covariates x_i, Section III of Heckman (1996) mentions that randomization can play the role of an instrumental variable for any x_i, so $m_-(x_i) - m_+(x_i)$ in (8) can be identified for any x_i. Following the discussion in Section 2.5, $\widehat{\delta}$ or $\widetilde{\delta}$ can be used to identify δ_0. The right panel of Figure 2 illustrates this intuition concerning the identification schemes for γ_0 and δ_0.

[8] Since in the neighborhood of γ, $E[q]$ does not have much variation, the denominator is not needed.

[9] The RDD literature usually assumes the jump location is known; see Porter and Yu (2011) for work on identifying treatment effects without this assumption.

3 The Roles of Instrumentation

When instruments are available, they can play multiple roles. To fully appreciate the various roles of instrumentation, we need to be clear about the best that can be achieved with and without instruments. In the first subsection below, we state some optimality results for β, δ and γ when instruments are absent. The following subsection explores some of the extra roles that instruments can play.

3.1 Optimality Results Without Instruments

The coefficient vector β cannot be identified without instrumentation since the effect of $\mathbf{x}'\beta$ and $\mathbb{E}[\varepsilon|x,q]$ are intermixed, just as the parameter β cannot be identified in the linear regression model $y = x'\beta + \varepsilon$ with endogenous regressors. On the other hand, the analysis of the previous section shows both δ and γ can be identified, with δ being estimable at a nonparametric rate whereas γ is estimable at the same rate as the parametric case. In this section, we first study the optimal rate of convergence for estimates of δ and then give the optimal estimation rate for γ from the existing literature.

To obtain the optimal rate of convergence for δ, 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 θ be a functional defined on $\mathcal{P}$. Given an estimator $\widehat{\theta}$ of θ and a loss function $L\left(\widehat{\theta}, \theta\right)$, the maximum expected loss over $P \in \mathcal{P}$ is defined to be

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

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

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

for some fixed $\epsilon > 0$, which will be used in this paper.[10] Under this loss, $R\left(\widehat{\theta}, \mathcal{P}\right)$ is the maximum probability that $\widehat{\theta}$ is not in the $\epsilon/2$ neighborhood of θ. The goal is to find an achievable lower bound for the minimax risk defined by

$$\inf_{\widehat{\theta}} R\left(\widehat{\theta}, \mathcal{P}\right) = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P\left[L\left(\widehat{\theta}, \theta(P)\right)\right]. \tag{10}$$

The right side generally converges to zero; the best rate of convergence of $R\left(\widehat{\theta}, \mathcal{P}\right)$ to zero is called the *optimal rate of convergence* or the *minimax rate of convergence*.

Since γ_0 can be estimated at rate n, its estimation does not affect the optimal rate of convergence of δ. We therefore assume that γ_0 is known in deriving the optimal rate of convergence of δ.[11] Now $P \in \mathcal{P}$ is characterized by δ and $g(x,q)$ as follows

$$\mathcal{P}(s,B) = \left\{P_{g,\delta} : \frac{dP_{g,\delta}}{d\mu} = f(x,q)\varphi_{x,q}\left(y - g(x,q) - \mathbf{x}'\delta 1(q \leq \gamma_0)\right), g(x,q) \in \mathcal{C}_s\left(B, \mathcal{X} \times \mathcal{N}\right), \|\delta\| \leq B\right\},$$

where μ is Lebesgue measure on $\mathbb{R}^{d+1}$, $\varphi_{x,q}$ is the conditional density of e given $(x',q)'$, and $\mathcal{C}_s\left(B, \mathcal{X} \times \mathcal{N}\right)$ is the class of s times continuously differentiable functions on $\mathcal{X} \times \mathcal{N}$ with all derivatives up to order s bounded by B and with $\mathcal{N}$ being a neighborhood of $q = \gamma_0$. The parameter of interest θ can be any element of δ, e.g., $\delta_\alpha\left(P_{g,\delta}\right) = \delta_\alpha$. The following theorem provides upper bounds for the rates of convergence.

[10] Quadratic loss is also popular, see, e.g., Fan (1993). Since the expected mean square error may not exist for the IDKE of δ, it is convenient to use the 0-1 loss function here.

[11] The problem with unknown γ_0 is harder than the problem with known γ_0, so the upper bounds in Theorem 4 below are also the upper bounds for the problem with unknown γ_0. Given that these upper bounds are achievable even if γ_0 were unknown, these bounds are also the optimal rates of convergence with unknown γ_0.

Theorem 4 *Under Assumptions E, F′, G′, and S, if $P \in \mathcal{P}(s,B)$ with $s = p+1$, then for $l = 1, \cdots, d-1$,*

$$\begin{aligned}
\varliminf_{n\to\infty} \inf_{\widehat{\delta}_{x_l}} \sup_{P\in\mathcal{P}(s,B)} P\left(\left|n^{\frac{s}{2s+1}}\left(\widehat{\delta}_{x_l} - \delta_{x_l}(P)\right)\right| > \frac{\epsilon}{2}\right) &\geq C, \\
\varliminf_{n\to\infty} \inf_{\widehat{\delta}_{q}} \sup_{P\in\mathcal{P}(s,B)} P\left(\left|n^{\frac{s-1}{2s+1}}\left(\widehat{\delta}_{q} - \delta_{q}(P)\right)\right| > \frac{\epsilon}{2}\right) &\geq C,
\end{aligned}$$

and

$$\begin{aligned}
\varliminf_{n\to\infty} \inf_{\widehat{\delta}_{\alpha}} \sup_{P\in\mathcal{P}(s,B)} P\left(\left|n^{\frac{s-1}{2s+1}}\left(\widehat{\delta}_{\alpha} - \delta_{\alpha}(P)\right)\right| > \frac{\epsilon}{2}\right) &\geq C \text{ if } \gamma_0 \neq 0, \\
\varliminf_{n\to\infty} \inf_{\widehat{\delta}_{\alpha}} \sup_{P\in\mathcal{P}(s,B)} P\left(\left|n^{\frac{s}{2s+1}}\left(\widehat{\delta}_{\alpha} - \delta_{\alpha}(P)\right)\right| > \frac{\epsilon}{2}\right) &\geq C \text{ if } \gamma_0 = 0
\end{aligned}$$

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

This theorem has interesting consequences. First, the main result is that we can estimate δ at most at a nonparametric rate. Second, estimation of δ does not suffer the curse of dimensionality. Specifically, an upper bound to the rate of convergence for δ_x is the same as for one-dimensional conditional mean estimation, and the upper bound for δ_q is the same as for one-dimensional slope estimation. As for δ_α, the upper bound depends on whether $\gamma_0 = 0$ or not: if $\gamma_0 \neq 0$, the upper bound is the same as in slope estimation; otherwise, it is the same as in level estimation. The upper bound for δ_q is not a surprise because δ_q is the slope difference in the neighborhood of $q = \gamma_0$. However, it may seems mysterious why δ_x, as the *slope* difference in the neighborhood of $q = \gamma_0$, has the same upper bound as in *level* estimation. The result may be understood as in an analogous way to average derivative estimation (ADE) (see, e.g., Stoker (1986), Powell et al. (1989), and Härdle and Stoker (1989) among others). Although the nonparametric derivative cannot be estimated at a $\sqrt{n}$ rate, the average derivative can be. In our case, only the data in a h neighborhood of γ_0 are used to estimate the average derivative, so the convergence rate should be $\sqrt{nh}$, and correspondingly, the optimal rate should be $\frac{s}{2s+1}$ (rather than $\frac{s-1}{2s+1}$). Actually, the present case is closer to the single index model of Ichimura (1993). Here the index is $\mathbf{x}'\delta$, so the slope differences in the left and right neighborhoods of $q = \gamma_0$ are the same at any x. This is also why we do not need the boundary condition that $f(x|q) = 0$ for q in a neighborhood of γ_0 and x on the boundary of its conditional support (see, e.g., Assumption 3 of Stoker (1986), Assumption 2 of Powell et al. (1989), Assumption 3.1 of Newey and Stoker (1993) or Assumption A.1.2 of Härdle and Stoker (1989) for counterparts in the average derivative estimation) to achieve this optimal rate. Without such boundary conditions, the average derivative cannot be estimated at a $\sqrt{n}$ rate; nevertheless, $\sqrt{n}$-consistency can still be achieved by the weighted semiparametric least squares estimator (WSLSE) of Ichimura (1993). See Yu (2014) for more discussions on this point.

With this intuition on the optimal rate for δ_x, the upper bound for δ_α is not hard to understand. Recall that $\delta_{\alpha 0} = \mathbb{E}\left[m_-(x) - m_+(x)\right] - \left(\mathbb{E}\left[x\right]' \delta_{x0} + \gamma_0\delta_{q0}\right)$. $\mathbb{E}\left[m_-(x) - m_+(x)\right]$, as a level difference, has the optimal rate $\frac{s}{2s+1}$, and δ_x has the optimal rate $\frac{s}{2s+1}$, so the optimal rate for δ_α is determined by whether $\gamma_0 = 0$ or not. If $\gamma_0 = 0$, its optimal rate is determined by the optimal rate of $\mathbb{E}\left[m_-(x) - m_+(x)\right]$ and δ_x, which is $\frac{s}{2s+1}$. Otherwise, its optimal rate is determined by the optimal rate of δ_q, which is $\frac{s-1}{2s+1}$ and is slower than the $\gamma_0 = 0$ case.

Checking the asymptotic distribution of $\widehat{\delta}$ and $\widetilde{\delta}$ in Theorem 2 and 3, we can see that the estimators $\widetilde{\delta}_\alpha$, $\widetilde{\delta}_x$ and $\widehat{\delta}_q$ each achieve the optimal rate for δ_α, δ_x and δ_q, respectively, provided the optimal bandwidth $h = O(n^{-1/(2s+1)})$ is used. It is interesting to notice that $\widehat{\delta}_x$ does not achieve the optimal rate of δ_x, whereas $\widetilde{\delta}_x$ does. This result parallels the efficiency comparison between the ADE and the WSLSE. Although both estimators are $\sqrt{n}$-consistent, the ADE is generally less efficient than the WSLSE; see, e.g., Section 5 of

Newey and Stoker (1993). This is because the ADE does not *fully* explore the linear index structure of the single index model. In our case, the IDKE of δ is like the ADE and does not use the information in the linear index structure $\mathbf{x}'\delta$. On the contrary, $\widetilde{\delta}_x$ fully exploits this linear index structure and so achieves the optimal rate of δ_x.[12] In contrast to the semiparametric case, in a nonparametric model the convergence rate of an estimator is inevitably slower if it does not fully exploit the linear index structure.

For γ, the optimality result is more subtle. In the parametric model, Yu (2012a) shows that the Bayes estimator is efficient in the minimix sense and is more efficient than the maximum likelihood estimator (MLE). Based on this result, Yu (2008) shows that the semiparametric empirical Bayes estimator (SEBE) can adaptively estimate γ_0 in the semiparametric case; in other words, the nonparametric components of the model do not affect the efficiency of γ_0, so that γ_0 can be estimated as if these components were known. Specifically, the following procedure is used to adaptively estimate γ_0 in the present case.

Algorithm G:

Step 1: Compute the IDKE $\left(\widehat{\gamma}, \widehat{\delta}'\right)'$, $\widehat{g}(x_i, q_i) = \frac{1}{(n-1)\widehat{f}_i} \sum_{j=1, j\neq i}^{n} K_{h,ij}\left(y_j - \mathbf{x}_j'\widehat{\delta}1(q_j \leq \widehat{\gamma})\right)$ and the corresponding residuals $\widehat{e}_i = y_i - \mathbf{x}_i'\widehat{\delta}1(q_i \leq \widehat{\gamma}) - \widehat{g}(x_i, q_i)$, $i = 1, \cdots, n$, where $\widehat{f}_i$ and $K_{h,ij}$ are defined in Section 2.4.

Step 2: Obtain a uniformly consistent estimator of the joint density of $\mathsf{w} \equiv (e, x', q)'$ by kernel smoothing, and denote the estimator as $\widehat{f}(\mathsf{w})$.

Step 3: Define the SEBE as

$$\widehat{\gamma}_o = \arg\min_t \int_\Gamma l_n(t-\gamma)\widehat{\mathcal{L}}_n(\gamma)\pi(\gamma)\,d\gamma.$$

where $l_n(t-\gamma) = l\left(n\left(t-\gamma\right)\right)$ is the loss function of γ, $\pi(\gamma)$ is the prior of γ, e.g., $\pi(\gamma)$ could be the uniform distribution on Γ, and

$$\begin{aligned}\widehat{\mathcal{L}}_n(\gamma) &= \prod_{i=1}^{n}\left[\widehat{f}\left(y_i - \mathbf{x}_i'\widehat{\delta}1(q_i \leq \widehat{\gamma}) - \widehat{g}(x_i, q_i), x_i, q_i\right)1(q_i \leq \gamma) + \widehat{f}\left(y_i - \widehat{g}(x_i, q_i), x_i, q_i\right)1(q_i > \gamma)\right] \\ &= \exp\left\{\sum_{i=1}^{n} 1(q_i \leq \gamma)\ln\left(\widehat{f}\left(y_i - \mathbf{x}_i'\widehat{\delta}1(q_i \leq \widehat{\gamma}) - \widehat{g}(x_i, q_i), x_i, q_i\right)\right) + \sum_{i=1}^{n} 1(q_i > \gamma)\ln\left(\widehat{f}\left(y_i - \widehat{g}(x_i, q_i), x_i, q_i\right)\right)\right\} \\ &\equiv \exp\left\{\widehat{L}_n(\gamma)\right\}\end{aligned}$$

is the estimated likelihood function.

The asymptotic distribution of $\widehat{\gamma}_o$ is $\arg\min_t \int_{\mathbb{R}} l\left(t-v\right)p^*(v)dv$, where $p^*(v) = \frac{\exp\{D_o(v)\}}{\int_{\mathbb{R}}\exp\{D_o(\widetilde{v})\}d\widetilde{v}}$, and $D_o(v)$ is similar to $D(v)$ in Theorem 1 except that now $\overline{z}_{1i} \equiv \ln\frac{f_{e|x,q}\left(e_i + \mathbf{x}_i'\delta_0 | x_i, q_i\right)}{f_{e|x,q}(e_i|x_i,q_i)}$ and $\overline{z}_{2i} \equiv \ln\frac{f_{e|x,q}\left(e_i - \mathbf{x}_i'\delta_0 | x_i, q_i\right)}{f_{e|x,q}(e_i|x_i,q_i)}$. Note also that the nonparametric posterior interval (NPI) based on $\widehat{\mathcal{L}}_n(\gamma)$ is a valid confidence interval for γ_0; see Section 4.1 of Yu (2009) for a summary of valid inference methods in threshold regression.

[12] Another estimator that fully exploits the linear index structure of the model is the PLE of δ (see its objective function (7)). We conjecture that this estimator also achieves the optimal rate of δ. However, a formal development of its asymptotic properties is beyond the scope of this paper; see Yu (2010) for such a development in the simple case of $d = 1$.

3.2 Optimality Results With Instruments

With instruments z in hand, we can estimate regular parameters $(\beta', \delta')'$ by means of the moment conditions

$$\mathbb{E}\left[z\varepsilon 1\left(q \leq \gamma_0\right)\right] = 0, \text{ and } \mathbb{E}\left[z\varepsilon 1\left(q > \gamma_0\right)\right] = 0, \tag{11}$$

where $z \in \mathbb{R}^{d_z}$ with $d_z \geq d+1$. Note that here we do not require $\mathbb{E}[\varepsilon|z, q] = 0$ as in Caner and Hansen (2004) to identify $(\beta', \delta')'$.[13] Also, it is irrelevant whether the reduced form is stable (i.e., the relationship between $\mathbf{x}$ and z is stable), which is important in the literature of 2SLS estimation. Since γ_0 can be consistently estimated by the IDKE, we can treat it as known in constructing the GMM objective function and estimates. Specifically,

$$\left(\widehat{\beta}'_{GMM}, \widehat{\delta}'_{GMM}\right)' = \arg\min_{\beta,\delta} n\overline{m}_n(\beta, \delta)' W_n \overline{m}_n(\beta, \delta), \tag{12}$$

where

$$\overline{m}_n(\beta, \delta) = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} z_i\left(y_i - \mathbf{x}_i'\beta - \mathbf{x}_i'\delta 1(q_i \leq \widehat{\gamma})\right) 1(q_i \leq \widehat{\gamma}) \\ z_i\left(y_i - \mathbf{x}_i'\beta - \mathbf{x}_i'\delta 1(q_i \leq \widehat{\gamma})\right) 1(q_i > \widehat{\gamma}) \end{pmatrix},$$

and W_n is a consistent estimator of the inverse of

$$\Omega = \mathbb{E}\left[\begin{pmatrix} zz'\varepsilon^2 1\left(q \leq \gamma_0\right) & \mathbf{0} \\ \mathbf{0} & zz'\varepsilon^2 1\left(q > \gamma_0\right) \end{pmatrix}\right] \equiv \begin{pmatrix} C & \mathbf{0} \\ \mathbf{0} & D \end{pmatrix}.$$

For example, W_n can be the inverse of the sample analog of Ω, say,

$$\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} z_i z_i' \widetilde{\varepsilon}_i^2 1\left(q \leq \widehat{\gamma}\right) & \mathbf{0} \\ \mathbf{0} & z_i z_i' \widetilde{\varepsilon}_i^2 1\left(q > \widehat{\gamma}\right) \end{pmatrix},$$

where $\widetilde{\varepsilon}_i = y_i - \mathbf{x}_i'\widetilde{\beta} - \mathbf{x}_i'\widetilde{\delta}1(q_i \leq \widehat{\gamma})$, and $\left(\widetilde{\beta}', \widetilde{\delta}'\right)'$ is the 2SLS estimator of $(\beta', \delta')'$ which is defined as the minimizer of (12) with

$$W_n^{-1} = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} z_i z_i' 1\left(q \leq \widehat{\gamma}\right) & \mathbf{0} \\ \mathbf{0} & z_i z_i' 1\left(q > \widehat{\gamma}\right) \end{pmatrix}.$$

It is easy to obtain

$$\begin{pmatrix} \widehat{\beta}_{GMM} \\ \widehat{\delta}_{GMM} \end{pmatrix} = \left(\widehat{G}'\widehat{\Omega}^{-1}\widehat{G}\right)^{-1} \widehat{G}'\widehat{\Omega}^{-1}\widehat{Z}'y,$$

where

$$\widehat{G} = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} z_i\mathbf{x}_i' 1\left(q \leq \widehat{\gamma}\right) & z_i\mathbf{x}_i' 1\left(q \leq \widehat{\gamma}\right) \\ z_i\mathbf{x}_i' 1\left(q > \widehat{\gamma}\right) & \mathbf{0} \end{pmatrix},$$

is a consistent estimator of

$$G = \begin{pmatrix} \mathbb{E}\left[z'\mathbf{x}1\left(q \leq \gamma_0\right)\right] & \mathbb{E}\left[z'\mathbf{x}1\left(q \leq \gamma_0\right)\right] \\ \mathbb{E}\left[z'\mathbf{x}1\left(q > \gamma_0\right)\right] & \mathbf{0} \end{pmatrix} \equiv \begin{pmatrix} A & A \\ B & \mathbf{0} \end{pmatrix},$$

and $\widehat{Z}$ and y denote matrices of stacked vectors $\left(z_i'1\left(q \leq \widehat{\gamma}\right), z_i'1\left(q > \widehat{\gamma}\right)\right)$ and y_i respectively. The following theorem gives the asymptotic distribution of $\left(\widehat{\beta}'_{GMM}, \widehat{\delta}'_{GMM}\right)'$.

[13] Since δ is already identified, we need only one of the two moment conditions in (11) to identify β.

Theorem 5 *Suppose $\widehat{\gamma}-\gamma_0=o_p(n^{-1/2})$, $\mathbb{E}\left[\|x\|^4\right]<\infty$, $\mathbb{E}[q^4]<\infty$, $\mathbb{E}[\varepsilon^4]<\infty$ and $\mathbb{E}[\|z\|^4]<\infty$; then*

$$\sqrt{n}\left(\begin{array}{c}\widehat{\beta}_{GMM}-\beta_0\\ \widehat{\delta}_{GMM}-\delta_0\end{array}\right)\overset{d}{\longrightarrow}N\left(0,\left(G'\Omega^{-1}G\right)^{-1}\right),$$

where the inverse of Ω and $G'\Omega^{-1}G$ are assumed to exist.

Compared to Theorem 2 and 3, the convergence rate of $\widehat{\delta}$ is improved from a nonparametric rate to $\sqrt{n}$. This is due to the fact that the moment conditions provide global information about δ, in contrast to the purely local identification information that is used when z is absent. Meanwhile, β, which is not identifiable without instruments, can now be identified. Note that we only assume $\widehat{\gamma}-\gamma_0=o_p(n^{-1/2})$ rather than $O_p(n^{-1})$ in the above theorem, an assumption that covers estimators of γ other than the IDKE.

From Hansen (1982), $\left(G'\Omega^{-1}G\right)^{-1}$ is the optimal asymptotic variance under moment conditions (11) with γ_0 known. Actually, according to Yu (2008), the GMM estimator is semiparametrically efficient even when γ_0 is unknown and the estimate $\widehat{\gamma}$ is used, as long as the loss function imposed on $(\beta',\delta')'$ and γ is additively separable. Alternatively, the empirical likelihood estimator of Qin and Lawless (1994) can be applied to achieve the semiparametric efficiency bound. Given the special forms of G and Ω, it can be shown that the asymptotic variance of $\widehat{\beta}_{GMM}$ is $\left(B'D^{-1}B\right)^{-1}$, and the asymptotic variance of $\widehat{\delta}_{GMM}$ is $\left(A'C^{-1}A\right)^{-1}\left[\left(A'C^{-1}A\right)^{-1}-\left(A'C^{-1}A+B'D^{-1}B\right)^{-1}\right]^{-1}\left(A'C^{-1}A\right)^{-1}$, so $\widehat{\beta}_{GMM}$ only exploits information in the data with $q_i>\widehat{\gamma}$ while $\widehat{\delta}_{GMM}$ uses information in all the data. These asymptotic variance matrices are consistently estimated using sample analogs, as is standard in the literature.

As to the efficient estimation of γ, we can still adaptively estimate it but now the joint density in Step 2 of Algorithm G also covers z. Specifically, we adjust Algorithm G as follows. In Step 1, we get a consistent estimator of ε_i (rather than e_i) as $\widehat{\varepsilon}_i=y_i-\mathbf{x}_i'\widehat{\beta}_{GMM}-\mathbf{x}_i'\widehat{\delta}_{GMM}1(q_i\leq\widehat{\gamma}_o)$.[14] In Step 2, we estimate the joint density of $(\varepsilon,x',q,z')'$ by kernel smoothing $\left\{(\widehat{\varepsilon}_i,x_i',q_i,z_i')'\right\}_{i=1}^n$ and still denote the estimator as $\widehat{f}$. In Step 3, we estimate γ_0 by $\arg\min_t\int_\Gamma l_n(t-\gamma)\widehat{\mathcal{L}}_n(\gamma)\pi(\gamma)\,d\gamma$, where $\widehat{L}_n(\gamma)$ in $\widehat{\mathcal{L}}_n(\gamma)$ is equal to

$$\sum_{i=1}^n 1(q_i\leq\gamma)\ln\left(\widehat{f}\left(y_i-\mathbf{x}_i'\widehat{\beta}_{GMM}-\mathbf{x}_i'\widehat{\delta}_{GMM}1(q_j\leq\widehat{\gamma}_o),x_i,q_i,z_i\right)\right)+\sum_{i=1}^n 1(q_i>\gamma)\ln\left(\widehat{f}\left(y_i-\mathbf{x}_i'\widehat{\beta}_{GMM},x_i,q_i,z_i\right)\right).$$

The asymptotic distribution of this estimator is similar to that of $\widehat{\gamma}_o$ except that now $\overline{z}_{1i}\equiv\ln\frac{f_{\varepsilon|x,q,z}\left(\varepsilon_i+\mathbf{x}_i'\delta_0\middle|x_i,q_i,z_i\right)}{f_{\varepsilon|x,q,z}(\varepsilon_i|x_i,q_i,z_i)}$ and $\overline{z}_{2i}\equiv\ln\frac{f_{\varepsilon|x,q,z}\left(\varepsilon_i-\mathbf{x}_i'\delta_0\middle|x_i,q_i,z_i\right)}{f_{\varepsilon|x,q,z}(\varepsilon_i|x_i,q_i,z_i)}$. So the information provided by z to γ improves its efficiency without affecting the convergence rate.

The following specific calculation illustrates the effect of z on the efficiency of γ estimation. Consider a simple threshold model

$$\begin{aligned} y &= \delta 1\left(q\leq\gamma\right)+\varepsilon, \\ \mathbb{E}\left[\varepsilon|q\right] &= g(q)\neq 0,\ \mathbb{E}[\varepsilon]=0. \end{aligned}\tag{13}$$

[14] $\widehat{e}_i$ may be used, but we expect that the performance based on $\widehat{\varepsilon}_i$ is better since the residuals are derived from a parametric (rather than semiparametric) model. Also, $\widehat{\gamma}_o$ is preferable to $\widehat{\gamma}$ since the former is more efficient than the later.

Suppose the joint distribution of $(\varepsilon, q, z)'$ is multivariate normal with mean $\mathbf{0}$ and variance matrix

$$\begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & \pi \\ 0 & \pi & 1 \end{pmatrix},$$

where π is defined in the reduced-form regression $q = \phi + z\pi + v$ with $\mathbb{E}[v|z] = 0$. A careful calculation shows that both z_{1i} and z_{2i} follow $N\left(-\frac{(1-\pi_0^2)\delta_0^2}{2(1-\pi_0^2-\rho_0^2)}, \frac{(1-\pi_0^2)\delta_0^2}{(1-\pi_0^2-\rho_0^2)}\right)$. Note that $\mathbb{E}[z_{1i}] < 0$ is a decreasing function of π_0^2, so the instrument z indeed improves the efficiency of γ estimation. Table 1 provides numerical results for this example based on the algorithms in Appendix D of Yu (2012a). The risk calculation in Table 1 is based on the asymptotic distribution rather than the finite-sample distribution, and RMSE entries are for the posterior mean and MAD for the posterior median. In Table 1, $\rho_0 = 0.5$, $\delta_0 = 1$, and $\gamma_0 = 0$. Evidently, as π_0 increases, z indeed provides more information about γ raising efficiency. Note that the case with $\pi_0 = 0$ corresponds to the risk of $\widehat{\gamma}_o$, where z does not provide extra information. Note further that z may provide information for γ without assuming $\mathbb{E}[\varepsilon|z] = 0$ or $Cov(z, x) \neq 0$ as long as z is not independent of $(\varepsilon, x', q)'$. The assumptions that $\mathbb{E}[\varepsilon|z] = 0$ and $Cov(z, x) \neq 0$ are used mainly to identify the parameters β and δ and achieve a $\sqrt{n}$ convergence rate.

In summary, instruments play different roles in relation to β, δ and γ as summarized in Table 2. From this table, the parameters β, δ and γ are affected in different ways by the presence of instrumentation, leading to differing convergence rates for the estimates of (β, γ) with and without instruments and efficiency improvements for estimates of γ.

	RMSE	MAD
$\pi_0 = 0$	9.109	6.093
$\pi_0 = 0.1$	9.017	6.085
$\pi_0 = 0.5$	8.143	5.473

Table 1: Efficiency Improvement in γ Estimation by z: $\rho_0 = 0.5$, $\delta_0 = 1$, and $\gamma_0 = 0$.

	Without Instruments	With Instruments
β	Unidentified	$\sqrt{n}$-consistency
δ	Nonparametric Consistency	$\sqrt{n}$-consistency
γ	n-consistency	Efficiency Improvement

Table 2: The Roles of Instruments to Different Parameters

4 Two Specification Tests

In this section, we discuss two specification tests of interest. The first test addresses potential endogeneity and the corresponding hypotheses $H^{(1)}$ are

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

This exogeneity test can be conducted prior to model estimation and we can use the techniques developed in Fan and Li (1996) and Zheng (1996) to test the null $H_0^{(1)}$. In the second test, the hypotheses $H^{(2)}$ are

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

If $H_0^{(1)}$ is not rejected, i.e., there is no evidence of endogeneity, then $H^{(2)}$ 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 $H_0^{(1)}$ 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 statistic are not pivotal, the simulation method of Hansen (1996) can be applied to obtain critical values. Details concerning these tests are given in the supplementary material of the paper. When there are no instruments, the Wald-type statistic is hard to implement since its asymptotic distribution is hard to derive given that $\widehat{\delta}$ can only be estimated at a nonparametric rate – see Section 3.3 of Porter and Yu (2011) for discussion.[15] However, the score-type test of Porter and Yu (2011) can be extended to this case with some technical complications. Importantly, the hypotheses $H^{(2)}$ relate to whether $m(x,q)$ is continuous, so $H_0^{(2)}$ encompasses more data generating processes 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 $\mathbf{x}'\beta + \mathbf{x}'\delta 1(q \leq \gamma)$ (see, e.g., Section 5.4 of Andrews (1993))[16], 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 (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.[17] 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 following discussion, H_0 indicates both $H_0^{(1)}$ and $H_0^{(2)}$, and H_1 indicates both $H_1^{(1)}$ and $H_1^{(2)}$, $1_q^\Gamma = 1(q \in \Gamma)$, $1_i^\Gamma = 1(q_i \in \Gamma)$, $m_i = m(x_i, q_i) = \mathbb{E}[y_i|x_i, q_i]$, $f_i = f(x_i, q_i)$, $K_{h,ij} = K_{h,ij}^x \cdot k_h(q_j - q_i)$, and $L_{b,ij} = L_{b,ij}^x \cdot l_b(q_j - q_i)$ with $l_b(\cdot)$ similarly defined as $k_h(\cdot)$. Denote the class of probability measures under $H_0^{(\ell)}$ as $\mathcal{H}_0^{(\ell)}$ and under $H_1^{(\ell)}$ as $\mathcal{H}_1^{(\ell)}$. Both $\mathcal{H}_0^{(\ell)}$ and $\mathcal{H}_1^{(\ell)}$ are characterized by $m(\cdot)$, so we acknowledge the dependence of the distribution of y given $(x', q)'$ upon $m(x,q)$ by denoting probabilities and expectations as P_m and $\mathbb{E}_m$, respectively. To unify notation, we define $u_i = y_i - \mathbb{E}[y_i|x_i, q_i] = y_i - m_i$ under both the null and alternative in both tests.

[15] 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.

[16] In this framework and assuming $m(x,q) = \mathbf{x}'\beta(q)$, the structural change tests focus on whether $\beta(q) = \beta$. See, e.g., Chen and Hong (2012), Kristensen (2012) and references therein for related tests in the time series context using nonparametric techniques. Actually, we can test whether $\beta(q)$ is continuous by extending the tests in Section 4.1, e.g., we can construct residuals $\widehat{e}_i$ in $I_n^{(2)}$ by estimating $\beta(q)$ using estimation techniques from the varying coefficient model (VCM) literature - see Robinson (1989, 1991), Cleveland et al. (1992) and Hastie and Tibshirani (1993).

[17] 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).

4.1 Test Construction and Asymptotics

For the first test, we use the test statistic

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

and, for the second, we use

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

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

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

in the first test, and

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

in the second test, where $\mathcal{C}_s(\cdot,\cdot)$ is defined in Section 3.1. Note 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 $\mathbb{E}[e\mathbb{E}[e|x,q]f(x,q)] = \mathbb{E}\left[\mathbb{E}[e|x,q]^2 f(x,q)\right] \geq 0$ in the first test and $\mathbb{E}[e\mathbb{E}[e|x,q]f(x,q)1_q^\Gamma] = \mathbb{E}\left[\mathbb{E}[e|x,q]^2 f(x,q)1_q^\Gamma\right] \geq 0$ in the second test where the equalities hold if and only if H_0 holds. So we can construct the statistic based on $\mathbb{E}[e\mathbb{E}[e|x,q]f(x,q)]$ in the first test and $\mathbb{E}[e\mathbb{E}[e|x,q]f(x,q)1_q^\Gamma]$ in the second test. Here, $f(x,q)$ is added in to avoid the random denominator problem in kernel estimation, and 1_q^Γ appears in the second test because 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

$$\widehat{e}_i = y_i - \widehat{y}_i = y_i - \left[\mathbf{x}_i'\widehat{\beta} + \mathbf{x}_i'\widehat{\delta} 1(q_i \leq \widehat{\gamma})\right], \tag{14}$$

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

$$\widehat{e}_i = y_i - \widehat{y}_i = (m_i - \widehat{m}_i) + (u_i - \widehat{u}_i), \tag{15}$$

where

$$\widehat{y}_i = \frac{1}{n-1} \sum\nolimits_{j \neq i} y_j L_{b,ij} \Big/ \widehat{f}_i \tag{16}$$

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

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

and $\widehat{m}_i$ and $\widehat{u}_i$ are defined in the same way as $\widehat{y}_i$ in (16) with y_j replaced by m_j and u_j, respectively. Under H_0, $\widehat{e}_i$ is a good estimate of u_i, while under H_1, $\widehat{e}_i$ includes a bias term which generates power. Now, $\mathbb{E}[e|x,q]f(x,q)$ at $(x_i', q_i)'$ is estimated by $\frac{1}{n-1}\sum_{j \neq i} \widehat{e}_j K_{h,ij}$ in the first test and $\frac{1}{n-1}\sum_{j \neq i} \widehat{e}_j K_{h,ij} 1_j^\Gamma$ in the second test. Hence, we may regard $I_n^{(1)}$ and $I_n^{(2)}$ as the sample analogs of $\mathbb{E}[e\mathbb{E}[e|x,q]f(x,q)]$ and $\mathbb{E}[e\mathbb{E}[e|x,q]f(x,q)1_q^\Gamma]$, respectively. The statistics are constructed under the null, mimicking the idea of score tests. More especially, the construction of $I_n^{(2)}$ does not involve $H_1^{(2)}$ at all (see Figure 2 of Porter and Yu (2011) for an intuitive illustration in a simple case), while the usual test statistics in the structural change literature typically involve $H_1^{(2)}$ in one way or another.

Before giving the asymptotic distributions of $I_n^{(1)}$ and $I_n^{(2)}$, it is convenient to specify regularity conditions on $f(x,q)$, $g(x,q)$, the distribution of u, the bandwidths b and h, and the kernel functions $l(\cdot)$.

Assumption F′: $f(x,q) \in \mathcal{C}_1(B, \mathcal{X} \times \mathcal{Q})$.
Assumption F″: $f(x,q) \in \mathcal{C}_\lambda(B, \mathcal{X} \times \Gamma_\epsilon)$ with $\lambda \geq 1$, and $0 < \underline{f} \leq f(x,q) \leq \overline{f} < \infty$ for $(x',q)' \in \mathcal{X} \times \Gamma_\epsilon$.
Assumption G″: $g(x,q) \in \mathcal{C}_s(B, \mathcal{X} \times \mathcal{Q})$ 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 $\mathcal{Q}^+ = (\gamma_0, \overline{q}]$.

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

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

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

Given $d > 1$, $h/b \to 0$ implies $h^{d/2}/b \to 0$, so $nh^d \to \infty$ implies that $nh^{d/2}b \to \infty$, where $nh^{d/2}b$ is the magnitude of $I_n^{(2)}$ under $H_1^{(2)}$. The quantity $nh^{d/2}b^{2\eta}$ is the bias of $I_n^{(2)}$ under $H_0^{(2)}$, so the assumption $nh^{d/2}b^{2\eta} \to 0$ guarantees that $I_n^{(2)}$ is centered at the origin. Under $H_0^{(1)}$, the bias of $I_n^{(1)}$ is $h^{d/2}$, so $h \to 0$ ensures that $I_n^{(1)}$ is also centered at the origin. The condition $h/b \to 0$ requires that h is smaller than b, which helps to generate power under $H_1^{(2)}$ and shrink the bias under $H_0^{(2)}$ to zero. Intuitively, if $h/b \to 0$, then the term $K_{h,ij}$ in $I_n^{(2)}$ makes the product $\widehat{e}_i\widehat{e}_j$ behave like a squared term 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 $h/b \to 0$.

Assumption L: $l_h(u,t)$ takes the form of (5) with order $p = s + \lambda - 1$.

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

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

Theorem 6 *Under Assumptions B1, F′, G″, K, S, and U, the following hold:*

(i)

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

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

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

with $\sigma^2(x,q) = \mathbb{E}[u^2|x,q]$ *can be consistently estimated by*

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

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

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

has significance level α*, where* z_α *is the* $1-\alpha$ *quantile of the standard normal distribution.*[18]

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

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

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

According to this result, $I_n^{(1)}$ does not have power if $\mathbb{E}\left[(m(x,q) - \overline{m}(x,q))^2 f(x,q)\right] = 0$. Consider the following special example to illustrate. Suppose $m(x,q)$ under $H_0^{(1)}$ is $\mathbf{x}'\beta + \mathbf{x}'\delta 1(q \leq \gamma)$, and the alternative is $m(x,q) = \mathbf{x}'\beta + \mathbf{x}'\delta 1(q \leq \gamma) + \mathbf{x}'\xi + \mathbf{x}'\zeta 1(q \leq \gamma)$, then obviously, $\mathbb{E}\left[(m(x,q) - \overline{m}(x,q))^2 f(x,q)\right] = 0$ 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). Possible cases that do generate power include (i) $m(x,q)$ takes the parametric form but has a different threshold point from $\overline{m}(x,q)$; (ii) $m(x,q)$ takes a nonparametric form.

Theorem 7 *Under Assumptions B2, F'', G'', K, L, S, and U, the following hold:*

(i)

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

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

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

and can be consistently estimated by

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

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

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

has significance level α*, where* z_α *is the* $1-\alpha$ *quantile of the standard normal distribution.*

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

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

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

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

These 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, $n^{-1/2}h^{-d/4}b^{-1/2}/n^{-1/2}h^{-d/4} = b^{-1/2} \to \infty$.

4.2 Bootstrapping Critical Values

As is evident from the proofs of theorems 6 and 7, the convergence rates of $T_n^{(1)}$ and $T_n^{(2)}$ to the standard normal is slow. The bias under $H_0^{(1)}$ is $h^{d/2}$ and under $H_0^{(2)}$ is $nh^{d/2}b^{2\eta}$. Both these rates are low for some standard choices of bandwidth. As argued in the literature of classical specification testing (see, e.g., Härdle and Mammen (1993), Li and Wang (1998), Stute et al. (1998), Delgado and Manteiga (2001), and Gu et al. (2007)), an improved approximation of the finite-sample distribution of $T_n^{(\ell)}$ can be obtained using the wild bootstrap (Wu, 1986; Liu, 1988). We therefore suggest that the following algorithm WB be used in both tests, with $\widehat{e}_i$ and $\widehat{y}_i$ having different definitions in the two tests.

Algorithm WB:

Step 1: For $i = 1, \cdots, n$, generate the two-point wild bootstrap residual $u_i^* = \widehat{e}_i\left(1-\sqrt{5}\right)/2$ with probability $\left(1+\sqrt{5}\right)/\left(2\sqrt{5}\right)$, and $u_i^* = \widehat{e}_i\left(1+\sqrt{5}\right)/2$ with probability $\left(\sqrt{5}-1\right)/\left(2\sqrt{5}\right)$, then $\mathbb{E}^*\left[u_i^*\right] = 0$, $\mathbb{E}^*\left[u_i^{*2}\right] = \widehat{e}_i^2$ and $\mathbb{E}^*\left[u_i^{*3}\right] = \widehat{e}_i^3$, where $\mathbb{E}^*[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_n]$ 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[19]

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

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

Step 3: Use the bootstrap samples to compute the statistics

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

and the estimated asymptotic variances

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

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

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

The studentized bootstrap statistics are $T_n^{(\ell)*} = I_n^{(\ell)*}/v_n^{(\ell)*}$. Here, the same b and h are used as in $I_n^{(\ell)}$ and $v_n^{(\ell)2}$ in Theorem 6 and 7.[20]

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

In Step 1, a popular way to simulate u_i^* in the second test is based on $\widehat{e}_i$'s centralized counterpart $\overline{\widehat{e}}_i = \widehat{e}_i - \overline{\widehat{e}}$ rather than $\widehat{e}_i$ itself, where $\overline{\widehat{e}} = \sum_{i=1}^{n} \widehat{e}_i 1_i^{\Gamma_b} \Big/ \sum_{i=1}^{n} 1_i^{\Gamma_b}$, $\Gamma_b = \left(\underline{\gamma} - b, \overline{\gamma} + b\right)$; see, e.g., Gijbels and Goderniaux (2004) and Su and Xiao (2008). Such a formulation can lead to $\sum_{i=1}^{n} \overline{\widehat{e}}_i 1_i^{\Gamma_b} \Big/ \sum_{i=1}^{n} 1_i^{\Gamma_b} = 0$,[21] which will not affect the asymptotic results but may affect the finite-sample performance of Algorithm WB especially under $H_1^{(2)}$.

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

Theorem 8 *Under the assumptions of Theorem 6 and 7,*

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

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

5 An Extension and Simplification

This section considers an extension and simplification of the earlier framework and analysis. We first examine the more general case where all elements of $(x', q)'$ are endogenous but $\mathbb{E}[\varepsilon|x, q]$ is not smooth at $q = \gamma_0$, and then look at the simpler case where some elements of $(x', q)'$ are exogenous.

5.1 $\mathbb{E}[\varepsilon|x, q]$ is Not Smooth at $q = \gamma_0$

When $\mathbb{E}[\varepsilon|x, q]$ is not smooth at $q = \gamma_0$, there are two cases. First, $\mathbb{E}[\varepsilon|x, q]$ is continuous but has a cusp at $q = \gamma_0$; second, $\mathbb{E}[\varepsilon|x, q]$ is discontinuous at $q = \gamma_0$. For example, consider the simple threshold model $y = \delta 1(q \le \gamma) + \varepsilon$, where $\varepsilon = \sigma_1 u 1(q \le \gamma) + \sigma_2 u 1(q > \gamma)$, and $\sigma_{10} \neq \sigma_{20}$. Also suppose $\mathbb{E}[u|q] = aq$ for a scalar $a \neq 0$. Then

$$\mathbb{E}[\varepsilon|q] = \sigma_{10} a q 1(q \le \gamma_0) + \sigma_{20} a q 1(q > \gamma_0).$$

[20] 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.

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

If $\gamma_0 = 0$, then $\mathbb{E}[\varepsilon|q]$ is continuous, but has a cusp at $q = \gamma_0$. If $\gamma_0 \neq 0$, then $\mathbb{E}[\varepsilon|q]$ is discontinuous at $q = \gamma_0$. In the general case where other covariates x are present, $\mathbb{E}[\varepsilon|x, q]$ may be a mixture of all three cases (smooth, continuous but having a cusp, and discontinuous) at $q = \gamma_0$ for different areas of x. To simplify the analysis, we discuss each case separately. Table 3 summarizes the identification and efficiency results with and without instruments in the latter two cases.

	$\mathbb{E}[\varepsilon\|x, q]$ Has a Cusp at $q = \gamma_0$		$\mathbb{E}[\varepsilon\|x, q]$ is Discontinuous at $q = \gamma_0$	
	Without Instruments	With Instruments	Without Instruments	With Instruments
β	Unidentified	$\sqrt{n}$-consistency	Unidentified	$\sqrt{n}$-consistency
δ_α, δ_q	Unidentified	$\sqrt{n}$-consistency	Unidentified	$\sqrt{n}$-consistency
δ_x	Nonparametric Consistency	$\sqrt{n}$-consistency	Unidentified	$\sqrt{n}$-consistency
γ	n-consistency	Efficiency Improvement	n-consistency	Efficiency Improvement

Table 3: The Roles of Instrumentation for Different Parameters when $\mathbb{E}[\varepsilon|x, q]$ is Not Smooth at $q = \gamma_0$

When $\mathbb{E}[\varepsilon|x, q]$ is continuous but has a cusp at $q = \gamma_0$, we find that $\lim_{q \to \gamma_0+} \partial\mathbb{E}[\varepsilon|x, q]/\partial x = \lim_{q \to \gamma_0-} \partial\mathbb{E}[\varepsilon|x, q]/\partial x$ by using a contradiction argument. So the estimators of δ_x in Section 2.5 are still applicable. On the other hand, δ_α and δ_q cannot be identified. This is because although $m_-(x) - m_+(x) = \delta_{\alpha 0} + x'\delta_{x0} + \gamma_0\delta_{q0}$ can be identified and thus the component $\delta_{\alpha 0} + \gamma_0\delta_{q0}$ can also be identified, $\delta_{\alpha 0}$ and δ_{q0} cannot be individually identified since δ_{q0} cannot be identified due to the cusp at γ_0.[22] When $\mathbb{E}[\varepsilon|x, q]$ is discontinuous at $q = \gamma_0$, we exclude the trivial case that $\mathbb{E}[\varepsilon|x, q]$ equals $-\mathbf{x}'\delta_0 1(q \leq \gamma_0)$ plus a smooth function of $(x', q)'$ as there will be no threshold effect in $m(x, q)$ at all in that case. If $m(x, q)$ indeed has a jump at $q = \gamma_0$,[23] no elements of δ can be identified, but γ can still be identified and estimated at the rate of n by the IDKE.

In testing $H_0^{(2)}$ versus $H_1^{(2)}$, our test statistic $T_n^{(2)}$ 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 $q = \gamma_0$.

5.2 Part of $(x', q)'$ is Exogenous

When part of $(x', q)'$ is exogenous, we can simplify our estimators in Section 2. Partition the variates $(x', q)'$ into $(x_1', x_2')'$, where x_1 is exogenous, and x_2 is endogenous and includes q. Importantly $\mathbb{E}[\varepsilon|x_1] = 0$ does not imply the mean independence condition $\mathbb{E}[\varepsilon|x, q] = \mathbb{E}[\varepsilon|x_2] \equiv g_2(x_2)$, that is, we cannot express $\mathbb{E}[y|x, q]$ as

$$\begin{aligned}\mathbb{E}[y|x, q] &= \beta_{1\alpha} 1(q \leq \gamma) + \beta_{2\alpha} 1(q > \gamma) + x_1'\beta_{21} + g(x_2) + (x_1'\delta_1 + x_2'\delta_2) 1(q \leq \gamma) \\ &= [\beta_{1\alpha} + x_1'\beta_{11} + g(x_2) + x_2'\delta_2] 1(q \leq \gamma) + [\beta_{2\alpha} + x_1'\beta_{21} + g(x_2)] 1(q > \gamma)\end{aligned}$$

which takes an additively separable form in x_1 and x_2, where β_ℓ and δ are partitioned according to the partition of $\mathbf{x} = (1, x_1', x_2')'$ as $\left(\beta_{\ell\alpha}, \beta_{\ell 1}', \beta_{\ell 2}'\right)$ and $\left(\delta_\alpha, \delta_1', \delta_2'\right)$, and $g(x_2) = x_2'\beta_{22} + g_2(x_2)$. In other words, the fact that only some of the regressors are endogenous does not provide extra identification information. So the estimation procedures given in Section 2 are still appropriate. But if the condition $\mathbb{E}[\varepsilon|x, q] = \mathbb{E}[\varepsilon|x_2]$ indeed holds almost surely, as is assumed by Newey et al. (1999) in the nonparametric estimation of triangular simultaneous equations models, then we can simplify the 'general endogenous case' estimation procedure.

[22] This is entirely analogous to the corresponding result in the linear model setting where both δ_α and δ_q cannot be identified in $y = \delta_\alpha + \delta_q q + \varepsilon$ if q is endogenous. If $\gamma_0 = 0$, then δ_α can be identified but this case is very special.

[23] More rigorously, $P(m_-(x) \neq m_+(x)) > 0$.

First, the IDKE of γ can be simplified. For each $\gamma \in \Gamma$, $\mathbb{E}[y|x_i, q = \gamma-]$ can be estimated as follows. the components β_{11} and $\beta_{1\alpha} + g(x_{2i}) + x_{2i}'\delta_2$ are estimated by extremum estimators $\widehat{\beta}_{11}$ and $\widehat{a}_i$ that are obtained from the following minimization problem

$$\min_{\beta_{11}, a_1, \cdots, a_n} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} K_{h,ij}^{\underline{x}_2} k_h^- (q_j - \gamma) \left[y_j - a_i - x_{1j}'\beta_{11} \right]^2, \tag{17}$$

where $\underline{x}_2$ is x_2 excluding q, and $K_{h,ij}^{\underline{x}_2}$ is similarly defined as $K_{h,ij}^{x}$ in (6). Note that β_{11} is the same for all $\underline{x}_{2i}$ in the objective function (17). In other words, the data in the left h neighborhood of $q = \gamma$ satisfies a partially linear model. The systematic part $\mathbb{E}[y|x_i, q = \gamma-]$ is then estimated as $x_{1i}'\widehat{\beta}_{11} + \widehat{a}_i$, which is denoted as $\widehat{m}_-(x_i; \gamma)$. The convergence rate of $\widehat{\beta}_{11}$ is expected to be $\sqrt{nh}$ if $\mathbb{E}\left[(x_1 - \mathbb{E}[x_1|\underline{x}_2, q = \gamma-]) (x_1 - \mathbb{E}[x_1|\underline{x}_2, q = \gamma-])' \right| q = \gamma-]$ > 0, and the convergence rate of $\widehat{a}_i$ is expected to be $\sqrt{nh^{d_2}}$, where $d_2 = \dim(x_2)$, and the positive-definiteness condition is a localized version of condition (3.5) in Robinson (1988). Similarly, $\mathbb{E}[y|x_i, q = \gamma+]$ can be estimated by $\widehat{m}_+(x_i; \gamma)$. Then, we can estimate γ_0 via the extremum problem

$$\widehat{\gamma} = \arg\max_{\gamma} \frac{1}{n} \sum_{i=1}^{n} \left[\widehat{m}_-(x_i; \gamma) - \widehat{m}_+(x_i; \gamma) \right]^2,$$

which is expected to be n-consistent.

Given $\widehat{\gamma}$, we can use the data with $q \le \widehat{\gamma}$ and $q > \widehat{\gamma}$ to estimate β_{11} and β_{21} using either the double residual regression method of Robinson (1988) or the pairwise difference estimator of Powell (1987, 2001). The resulting estimators are expected to be $\sqrt{n}$-consistent when

$$\mathbb{E}\left[(x_1 - \mathbb{E}[x_1|x_2]) (x_1 - \mathbb{E}[x_1|x_2])' 1(q \le \gamma_0) \right] > 0 \text{ and } \mathbb{E}\left[(x_1 - \mathbb{E}[x_1|x_2]) (x_1 - \mathbb{E}[x_1|x_2])' 1(q > \gamma_0) \right] > 0.$$ [24]

Note that here we use all the data with $q \le \widehat{\gamma}$ to estimate β_{11} but only the data in the left h neighborhood of $q = \gamma$ to estimate β_{11} in (17). This is because for an arbitrary $\gamma \in \Gamma$, $\mathbb{E}[y|x, q]$ may not take a partially linear form when $q \le \gamma$. For example, suppose $\gamma > \gamma_0$. Then for $\gamma_0 < q \le \gamma$, $\mathbb{E}[y|x, q] = \beta_{2\alpha} + x_1'\beta_{21} + g(x_2)$, while for $q \le \gamma_0$, $\mathbb{E}[y|x, q] = \beta_{1\alpha} + x_1'\beta_{11} + g(x_2) + x_2'\delta_2$. So, there is no uniformly partially linear form for all $q \le \gamma$. Nevertheless, $\mathbb{E}[y|x, q]$ must take a partially linear form in the left neighborhood of $q = \gamma$ although we are unsure a priori which one of the two forms it will take. In other words, $\widehat{\beta}_{11}$ in (17) may actually be estimating β_{21}. Given the estimates of β_{11} and β_{21}, which we still denote as $\widehat{\beta}_{11}$ and $\widehat{\beta}_{21}$ to simplify notation, we can construct

$$\widetilde{y} = y - x_1'\widehat{\beta}_{11} 1(q \le \widehat{\gamma}) - x_1'\widehat{\beta}_{21} 1(q > \widehat{\gamma}),$$

which satisfies

$$\mathbb{E}[\widetilde{y}|x_2] \approx \beta_{2\alpha} + g(x_2) + (\delta_\alpha + x_2'\delta_2) 1(q \le \gamma_0).$$

So here δ_α and δ_2 can be estimated using the procedures in Section 2.5 by only $\{(\widetilde{y}_i, x_{2i}', q_i)'\}_{i=1}^n$.

Often endogeneity affects only a single covariate, in which case x_2 is one-dimensional. In this case, the simplified estimators do not suffer the curse of dimensionality as do the general estimators. In the empirical application of Section 7, where x_2 is binary, we show that even kernel smoothing is not required. If we further assume that ε is independent of x_1 conditional on $(x_2', z')'$ when instruments z are available, we need only estimate the joint density of $(\varepsilon, x_2', z')'$ in Step 2 of the modified Algorithm G in Section 3.2.[25]

[24] This definition covers the case where q is included in x_2. If q is included in x_1, the corresponding conditions can be written as $E[\underline{m}\underline{m}'] > 0$, where $\underline{m} = \begin{pmatrix} (x_1 1(q \le \gamma_0) - E[x_1 1(q \le \gamma_0)|x_2]) 1(q \le \gamma_0) \\ (x_1 1(q > \gamma_0) - E[x_1 1(q > \gamma_0)|x_2]) 1(q > \gamma_0) \end{pmatrix}$.

[25] Note also that if e is independent of $(x', q)'$, then in Step 2 of Algorithm G, we need only estimate the density of e. Of

In testing $H_0^{(1)}$ versus $H_1^{(1)}$, our test statistic $T_n^{(1)}$ can be modified correspondingly. If we are sure that x_1 is exogenous in the sense that $\mathbb{E}[\varepsilon|x,q] = \mathbb{E}[\varepsilon|x_2]$ almost surely, then the null hypothesis changes to $\mathbb{E}[\varepsilon|x_2] = 0$. In this case, $K_{h,ij}$ in $T_n^{(1)}$ changes to $K_{h,ij}^{x_2} \cdot k_h(q_j - q_i)$, and the asymptotic distributions in Theorem 6 and the wild bootstrap procedure in Section 5.2 adjust accordingly. If we further assume that there is indeed a threshold effect at $q = \gamma_0$ (which can be verified by the test based on $T_n^{(2)}$), then we can estimate γ_0, β_{11} and β_{21} and construct $\widetilde{y}$ as above. Under the null, $\mathbb{E}[\widetilde{y}|x_2] \approx \beta_{2\alpha} + x_2'\beta_{22} + (\delta_\alpha + x_2'\delta_2) 1(q \le \gamma_0)$. So $T_n^{(1)}$ can be constructed based on $\{(\widetilde{y}_i, x_{2i}', q_i)'\}_{i=1}^n$ instead of $\{(y_i, x_i', q_i)'\}_{i=1}^n$. In this case, we can also test whether the threshold effect is conveyed only by x_1, i.e., whether $(\delta_\alpha, \delta_2')' = 0$, using $T_n^{(2)}$ computed by only $\{(\widetilde{y}_i, x_{2i}', q_i)'\}_{i=1}^n$.[26]

6 Simulations

This section presents two simulation studies designed to assess the adequacy of the limit theory. The first simulation compares the efficiency of the IDKE and DKE of γ, and the second compares the size and power properties of the test $T_n^{(2)}$ with the parametric testing procedure of Hansen (1996).

According to our earlier findings, the DKE is less efficient asymptotically than IDKE. Also, in applying the DKE the fixed point x_o used in the criterion function is hard to select since $\mathbb{E}[y|x, q = \gamma_0-] - \mathbb{E}[y|x, q = \gamma_0+]$ and $f_{x|q}(x|\gamma_0)$ have unknown forms. In implementing the simulation, we used for illustration the simple model $y = 1(q \le \gamma) + \varepsilon$, where $\gamma_0 = 0$, $\delta_0 = (1,0,0)'$, x and q are independent and each is uniformly distributed over $[-0.5, 0.5]$, and $\varepsilon|(x,q) \sim N(-q, 0.2^2)$. The threshold effect does not depend on x, and so the DKE of Delgado and Hidalgo (2000) can be applied. We set $x_o = 0$, and $\Gamma = [-0.2, 0.2]$. Three bandwidths are used based on $Cn^{-1/6}$ with proportionality constants $C = 0.3$, 0.5 and 0.7.[27] 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. The rescaled Epanechnikov kernel is used, viz.,

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

which degenerates to the Epanechnikov kernel when $r = 1$, and $k_+(x,r) = k_-(-x,r)$. This kernel function guarantees that $k_\pm(0,r) > 0$. Note that the kernel functions in Table 1 of Müller (1991) do not satisfy this condition and so they are not used in this simulation.

n	200				800			
Estimators	$\widehat{\gamma}$		$\widetilde{\gamma}$		$\widehat{\gamma}$		$\widetilde{\gamma}$	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
$C = 0.3$	-5.144	8.296	-7.853	10.309	-0.498	1.891	-5.473	8.575
$C = 0.5$	-1.632	3.937	-4.100	6.720	-0.262	0.665	-1.906	4.125
$C = 0.7$	-1.258	3.059	-2.750	5.158	-0.252	0.579	-0.958	2.192

Table 4: Bias and RMSE of $\widehat{\gamma}$ and $\widetilde{\gamma}$ (in 10^{-2}): $x_o = 0$
(Based on 500 Repetitions)

course, ε cannot be independent of $(x', q)'$, but it is quite possible that ε is independent of $(x', q)'$ conditional on z as in the control function approach. In this case, we need only estimate the joint density of $(\varepsilon, z')'$ in Step 2 of the modified Algorithm G in Section 3.2.

[26] $T_n^{(1)}$ and $T_n^{(2)}$ can be constructed by treating γ_0 as known to be $\widehat{\gamma}$ or unknown but restricted to a small interval around $\widehat{\gamma}$ rather than Γ to improve test power.

[27] $C = 0.3$ roughly approximates the standard deviation (0.289) of the uniform distribution on $[-0.5, 0.5]$. $1/6 = 1/(2s+d)$ with $s = d = 2$. There are roughly $N = n \times (2Cn^{-1/6}) \times Cn^{-1/6} = 2C^2n^{2/3}$ data points in a h neighborhood of (x_i, γ). When $c = 0.3$ and $n = 200$, $N \approx 6$. When $c = 0.7$ and $n = 800$, $N \approx 84$.

Tests	$T_n^{(2)}$					Hansen (1996)			
	Size (%)		Power (%)			Size (%)		Power (%)	
n	200	800	200	800		200	800	200	800
$C = 0.3$	4.4	3.4	77.6	100	SupW	100	100	100	100
$C = 0.4$	4.6	4.4	94.8	100	AveW	100	100	100	100
$C = 0.5$	3.6	5.6	99.8	100	SupLM	100	100	100	100
$C = 0.6$	3.6	5.4	100	100	AveLM	100	100	100	100

Table 5: Size and Power of $T_n^{(2)}$ and Hansen (1996): Significance Level = 5%
(Based on 500 Repetitions)

We consider 500 random samples of size $n = 200$ and 800. The simulation results are summarized in Table 4. The following conclusions are drawn. First, the IDKE performs better than the DKE in terms of both bias and RMSE for all bandwidths and sample sizes. For this simple setup, a larger bandwidth seems preferable. For the bandwidth specification $Cn^{-1/6} \approx 0.3$ when $C = 0.7$ and $n = 200$, which roughly corresponds to the parametric estimation, noticing that the distance between $\overline{\gamma}$ $(= 0.2)$ and the right boundary of q's support (0.5) is 0.3. Understandably, parametric estimation is more efficient.

To illustrate why the IDKE is more efficient than the DKE, Figure 3 shows typical objective functions of the IDKE and DKE. There are local maximizers in both objective functions. But since the DKE is determined only by the information in the neighborhood of the chosen point x_o, this estimator turns out to be determined by a global-maximizer (in this case a pseudo-maximizer) that lies further from the true value in the parameter space than the local maximizer. In contrast, the IDKE incorporates jump information from other areas of the sample space $\mathcal{X}$, and turns out to be determined by the maximizer that is closer to the true value. Second, comparing the RMSE of $\widehat{\gamma}$ and $\widetilde{\gamma}$ for $n = 200$ and 800, it is apparent that the convergence rate of $\widehat{\gamma}$ is much faster than $\widetilde{\gamma}$. Taking the ratio of the RMSEs for $n = 200$ and $n = 800$, the convergence rate of $\widetilde{\gamma}$ is clearly slower than n, whereas for $\widehat{\gamma}$ the convergence rate seems close to $O(n)$ [28]. Another interesting phenomenon is that all biases are negative. This is mainly due to the bias problem in the construction of the objective functions for $\widehat{\gamma}$ and $\widetilde{\gamma}$, as mentioned in Section 2.2. But if the local linear smoother is used, then the algorithm was found to be unstable in our simulations because the denominator matrix tends to be singular.

For specification testing, we retain the same setup. Since $\beta_{x0} = \delta_{x0} = 0$, we neglect data for x, and the model becomes $y = -q + \delta 1(q \leq \gamma) + e$ with $e \sim N(0, 0.2^2)$. We study type-I error by setting $\delta_0 = 0$ and power by setting $\delta_0 = 1$ and $\gamma_0 = 0$. We use Algorithm WB with $B = 399$ to obtain the critical values for $T_n^{(2)}$ and let the bandwidth $b = Cn^{-1/5}$ with $C = 0.3, 0.4, 0.5, 0.6$ and $h = b^{2.1}$.[29] Suppose Hansen (1996) misspecifies the model as $y = \beta + \delta 1(q \leq \gamma) + e$. In simulating his critical values, the approximate parameter space Γ_A is the set of q_i's in Γ, and the replication number is $J = 500$. For illustration purpose, we only report simulation results for four of his test statistics. Table 5 summarizes the simulation results for all tests of significance level 5%. The results show that the type-I errors for the Hansen (1996) tests are all large (100%), as expected, while the type-I errors of our test match the target 5% quite well for all bandwidths considered. The power properties of $T_n^{(2)}$ are very good, showing that the test identifies models with $\delta_0 \neq 0$ with high probability even when $n = 200$. The powers of $T_n^{(2)}$ and those of the Hansen (1996) tests are not comparable since the latter are based on a misspecified model with distorted size.

[28] For example, when $C = 0.3$ we have $8.296/1.891 = 4.387$ for $\hat{\gamma}$ and $10.309/8.575 = 1.202$ for $\widetilde{\gamma}$.

[29] When $d = 1$, the assumption $h/b \to 0$ in Assumption B2 should change to $h^{1/2}/b \to 0$ and $h = b^{2.1}$ satisfies this assumption.

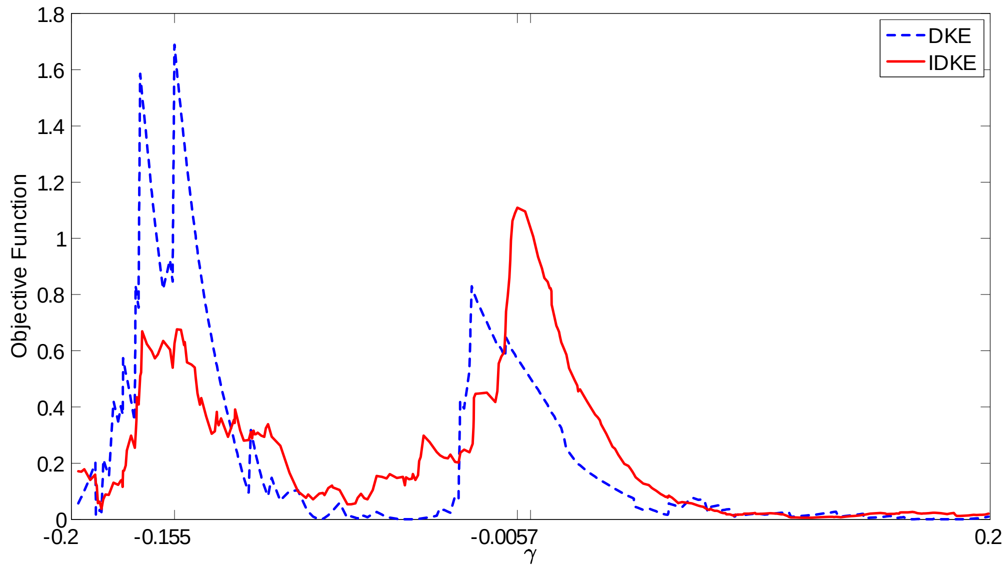

Figure 3: Objective Functions of the DKE and IDKE

7 Empirical Application

In the early 1980s, the United States introduced several tax-deferred savings options designed to increase individual savings for retirement, the most popular being Individual Retirement Accounts (IRAs) and 401(k) plans. IRAs and 401(k) plans are similar in that both allow the individual to deduct contributions to retirement accounts from taxable income and they both permit tax-free accrual of interest. The key difference between these schemes is that employers provide 401(k) plans and may match some percentage of the employee 401(k) contributions. Therefore, only workers in firms that offer such programs are eligible, whereas IRAs are open to all.[30]

An interesting question that has attracted attention in the literature is whether contributions to tax-deferred retirement plans represent additional savings or simply crowd out other types of savings. A central difficulty that complicates empirical investigation of this question is the presence of saver heterogeneity coupled with nonrandom selection into the participation states. Individuals who participate in tax-advantaged retirement savings plans are likely to already have a strong preference for savings, implying that they would have saved more than those who do not participate even in the absence of such schemes. The econometric consequence is that conventional least squares regression may overestimate the effects of these plans. A common solution to this endogeneity problem is to select an instrument and apply 2SLS. As suggested by Poterba et al. (1994, 1995, 1998), 401(k) eligibility is exogenous given some observables (most importantly, income).[31] Their suggestion is based on the observation that 401(k) eligibility is decided by employers, and unobserved preferences for savings may play a minor role in the determination of eligibility once we control for the effects of observables. Following this suggestion, we use 401(k) eligibility as an instrument for participation in 401(k) programs. The same approach is used by Abadie (2003) and Chernozhukov and Hansen (2004) in estimating local average treatment effects (LATEs) and the quantile treatment effects, respectively.

[30] See the Employee Benefit Research Institute (1997) for a detailed description of tax-deferred retirement programs, their history and regulations.

[31] See Engen et al. (1996) for a different point of view. These authors contend that eligibility should not be treated as exogenous.

We use the same data set as Abadie (2003), comprising 9275 observations from the Survey of Income and Program Participation (SIPP) of 1991. This sample is often referred to as the 1991 SIPP, and is used extensively in the literature to examine the effect of 401(k) plans on wealth; see, inter alia, Benjamin (2003), Engen and Gale (2000), Engen et al. (1996), and Poterba et al. (1994, 1995, 1998). As discussed in Chernozhukov and Hansen (2004), the sample is confined to households in which the reference person is 25-64 years old (with spouse if present) and at least one family member is employed and no member is self-employed. Annual family income is required to fall in the \$10,000-\$200,000 interval. Outside this interval, 401(k) eligibility in the sample is rare. See Table 1 of Abadie (2003) for descriptive statistics of the data set.

There is no literature considering possible threshold effects in the evaluation of treatment effects under endogeneity. Our threshold treatment model is motivated by the 2SLS estimates of the treatment effects for different income categories. Table 6 summarizes the OLS and 2SLS estimates of the effect of 401(k) participation for the full sample and the six income categories (as similarly specified in Table 3 of Chernozhukov and Hansen, 2004). The model is formulated as

$$y = D\alpha + X'\beta + \varepsilon, \mathbb{E}\left[\varepsilon|z, X\right] = 0,$$

where y is net financial assets, D is 401(k) participation status, z is 401(k) eligibility, and X includes a constant and five covariates (family income, age, age squared, marital status and family size) just as in Abadie (2003).

The findings that emerge from Table 6 are as follows. From the first stage regression results reported in column 3, it is evident that the instrument z is not weak either for the full sample or for the subsamples within each income category. Second, there is an obvious upward bias in the OLS estimates (except for Category I). Third and most importantly for the present study, there are obvious threshold effects evident in the 2SLS estimates: Category V and VI clearly differ from the other four categories; and Category III and IV (especially IV) differ from the first two categories. The 2SLS estimate using the full sample is close to the 2SLS estimate for Category III but differs from the 2SLS estimates for all other categories. Based on these findings, we specify the model as[32]

$$y = \begin{cases} D\alpha_1 + X'\beta_1 + \varepsilon, & inc \leq \gamma_1, \\ D\alpha_2 + X'\beta_2 + \varepsilon, & \gamma_1 < inc \leq \gamma_2, \\ D\alpha_3 + X'\beta_3 + \varepsilon, & inc > \gamma_2, \end{cases} \tag{18}$$

where inc, the family income, is the threshold variable. The three regimes correspond to low-income, middle-income and high-income individuals.

Model (18) is very special since the only endogenous variable D is binary. As in Section 5.2, suppose ε is mean independent of X given D, that is, $\mathbb{E}[\varepsilon|D, X] = \mathbb{E}[\varepsilon|D]$. Then because D is binary, $\mathbb{E}[\varepsilon|D]$ must be a linear function of D.[33] In other words, the relationship between y and $(D, X')'$ satisfies

$$y = \begin{cases} D\widetilde{\alpha}_1 + \widetilde{\beta}_{10} + \underline{X}'\underline{\beta}_1 + e, & inc \leq \gamma_1, \\ D\widetilde{\alpha}_2 + \widetilde{\beta}_{20} + \underline{X}'\underline{\beta}_2 + e, & \gamma_1 < inc \leq \gamma_2, \\ D\widetilde{\alpha}_3 + \widetilde{\beta}_{30} + \underline{X}'\underline{\beta}_3 + e, & inc > \gamma_2, \end{cases} \tag{19}$$

where $\underline{X}$ ($\underline{\beta}_\ell$) is X (β_ℓ) excluding the constant (the intercept), $\widetilde{\alpha}_\ell$ and $\widetilde{\beta}_{\ell 0}$, $\ell = 1, 2, 3$, may differ from those

[32] In the notation of (1), $x = (D, \underline{x}')'$, $q = inc$, where $\underline{x}$ is X excluding the constant and inc.

[33] This result is not correct when D is continuous or can take more than two values when it is discrete. Note that Perron and Yamamoto (2012b) use OLS to estimate the structural change points even when D is continuous and the resulting estimates are generally inconsistent.

in (18), but $\underline{\beta}_\ell$, $\ell = 1, 2, 3$, are the same as in (18). The new error term satisfies $\mathbb{E}[e|D, X] = 0$. Given this structure, the LSEs of γ_1 and γ_2 are consistent although the LSEs of α_ℓ, $\ell = 1, 2, 3$, are inconsistent. We use the sequential estimation procedure of Bai (1997) to consistently estimate γ_1 and γ_2. Given a consistent estimator of γ_1 and γ_2, α_ℓ and β_ℓ can be consistently estimated by the 2SLS procedure developed here, and a consistent estimate of ε follows. A testable restriction of $\mathbb{E}[\varepsilon|D, X] = \mathbb{E}[\varepsilon|D]$ can be based on the difference between the LSE of $\underline{\beta}_\ell$ and the 2SLS estimator of $\underline{\beta}_\ell$. We will conduct such tests after estimation.

Sample	n	First Stage	OLS	2SLS		OLS	2SLS
Full Sample	9275	0.6883 (0.0080)	13527.05 (1809.59)	9418.83 (2152.08)	D	13527.05 (1809.59)	9418.83 (2152.08)
I: \$10 − 20K	1848	0.6433 (0.0253)	5486.07 (1476.71)	5716.16 (1629.46)	Constant	−23549.00 (2177.26)	−23298.74 (2166.58)
II: \$20 − 30K	2093	0.6120 (0.0193)	8029.73 (1422.41)	4507.68 (2243.38)	Family Income (in thousand \$)	976.93 (83.34)	997.19 (83.82)
III: \$30 − 40K	1693	0.6677 (0.0178)	12626.59 (2525.26)	9348.88 (2715.16)	Age − 25	−376.17 (236.89)	−345.95 (238.01)
IV: \$40 − 50K	1204	0.7194 (0.0187)	14780.65 (2433.97)	11297.49 (3563.82)	$(\text{Age} - 25)^2$	38.70 (7.66)	37.85 (7.69)
V: \$50 − 75K	1572	0.7452 (0.0147)	24309.73 (3332.90)	23107.01 (3911.53)	Married	−8369.47 (1829.24)	−8355.87 (1828.98)
VI: > \$75K	765	0.8341 (0.0174)	27948.78 (10463.97)	25965.50 (12987.00)	Family Size	−785.65 (410.62)	−818.96 (410.39)

Table 6: OLS and 2SLS Estimates of the Effect of 401(k) Participation for Six Income Categories [first five columns] and All Coefficients for the Full Sample [last three columns]

Notes: n is the sample size for each row, column "First Stage" contains the coefficients of 401(k) eligibility in the first stage regression, and standard errors are reported in parentheses.

Given the LSE of γ_1 and γ_2, we can use the modified Algorithm G in Section 3.2 to improve efficiency in estimation of γ_1 and γ_2. To simplify the estimation of the likelihood function, assume $\varepsilon \perp X|(D, z)$ where "$\perp$" denotes independence (c.f., Dawid, 1979) and variables to the right of "$|$" are the conditioning variables.[34] Then as argued in Section 5.2, we need only estimate $f(\varepsilon|D, z)$ to construct the nonparametric posterior interval (NPI) for γ. In other words, only three univariate density functions are estimated.[35] The bandwidths in the density estimation are selected by the method proposed in Botev et al. (2010). For computational convenience we combine Regimes I and II in (18) to construct the NPI for γ_1, and combine Regimes II and III to construct the NPI for γ_2, rather than constructing the NPI for γ_1 and γ_2 simultaneously. All implementation details and code are available upon request.

Another estimator of γ_1 and γ_2 is the 2SLS estimator of Caner and Hansen (2004), as mentioned in the Introduction. That estimator is inconsistent unless a consistent estimator of $\mathbb{E}[D|z, X]$ rather than the linear projection of D on $(z, X')'$ is used in the second stage (see Yu, 2012b). To illustrate, we use both the linear projection of D on $(z, X')'$ and the Probit fit of D on $(z, X')'$ in the second stage to show the differences in the corresponding 2SLS estimators. All the estimators of γ_1 and γ_2 mentioned above and the corresponding three regimes are summarized in Table 7. Some of the findings in Table 7 are summarized as follows. First,

[34] $E[\varepsilon|D, X] = E[\varepsilon|D]$ does not imply $\varepsilon \perp X|D$, but when one more variable z is put in the conditional set, $\varepsilon \perp X|D, z$ is more likely to hold.

[35] Note that $z = 0$ and $D = 1$ are an impossible outcome since only eligible individuals can open a 401(k) account.

Regime I is the same for all estimators. Second, compared to the Caner-Hansen 2SLS estimator, the LSE of γ_2 is closer to the posterior mean (or median) which is most efficient. When the Probit fit of D on $(z, X')'$ is used in the second stage of Caner-Hansen 2SLS estimation, the resulting estimate is the same as the LSE. This result corroborates the finding in Yu (2012b) that a consistent estimator of $\mathbb{E}[D|z, X]$ is preferable to linear projection of D on $(z, X')'$ in that procedure. Third, the NPIs for both γ_1 and γ_2 are narrow (each interval covers only 12 data points), which indicates that regime splitting by the posterior mean (or median) is precise here.

	γ_1	γ_2	n in Regime I	n in Regime II	n in Regime III
OLS	42.870	69.006	6112	2151	1012
CH (2004)+Linear	42.870	68.225	6112	2116	1047
CH (2004)+Probit	42.870	69.006	6112	2151	1012
Posterior Mean	42.866	71.326	6112	2260	903
Posterior Median	42.869	71.349	6112	2262	901
NPI	$[42.810, 42.876]$	$[71.087, 71.358]$			

Table 7: Estimates of γ_1 and γ_2, the NPI and Numbers of Data Points in Each Regime

Table 8 reports the OLS and 2SLS estimates of $(\alpha_\ell, \beta'_\ell)'$ in the three regimes split according to the posterior median. (Results based on the posterior mean are similar and are omitted here). First, the 2SLS estimates of α_ℓ in all three regimes are significantly different from zero at all conventional significance levels. This result implies that participation in the 401(k) plans indeed increases savings for all individuals across different levels of income, and that the putative crowding-out effect on savings is not significant. Second, the savings of the high-income individuals increase the most, i.e., the greatest advantage of 401(k) plans is taken by rich people. Third, the OLS and 2SLS estimates of $\underline{\beta}_\ell$ are similar. Rigorous tests cannot reject the null that they are equal in all three regimes, which supports the assumption that $\mathbb{E}[\varepsilon|D, X] = \mathbb{E}[\varepsilon|D]$. Fourth, the OLS and 2SLS estimates of α_ℓ are quite different, which confirms that D is endogenous. Fifth, the $\underline{\beta}_\ell$'s in the three regimes are all quite different. In other words, saving behavior of these three groups of individuals differs empirically. More specifically, we note the following: (i) family income has a larger (positive) impact on savings for richer people; (ii) differing from people in Regime I and II, age has a large positive impact on savings for people in Regime III; (iii) married persons generally have less savings than unmarried persons, and the extent is larger for richer people; (iv) family size does not have much impact on savings for high-income individuals, whereas it has a significantly negative impact for low-income and middle-income individuals. All these results are intuitively reasonable. Importantly, compared to the last three columns of Table 6, the 2SLS estimates using the full sample obscure the differences in the roles of covariates (especially the participation in 401(k) plans) on savings amongst various income groups.

These findings have significant policy implications. The intended purpose of IRAs and 401(k) plans was to encourage savings for retirement rather than encourage investment by avoiding taxation. IRAs have already witnessed large balances since their introduction, which triggers limitations on deductable levels of income. Specifically, the amount of IRA contributions deductable from current-year taxes is partially reduced for levels of income beyond a threshold, and eliminated entirely beyond another threshold.[36][37] Such limitations

[36] This rule applies if the contributor and/or the contributor's spouse is covered by an employer-based retirement plan; see IRS Publication 590 for the details.

[37] This policy can be justified by repeating the analysis above with the IRA participation status added to X. The coefficients of D are qualitatively similar to those in Table 8. Also, the coefficients of the IRA participation status are statistically significant and show threshold effects among the three regimes. We did not conduct such an analysis in the main text because the IRA participation status is also endogenous, while the (comprehensive) IRA eligibility (not like the 401(k) eligibility) is trivial and is not a valid instrument.

do not exist for 401(k) plans, although there is a maximum deductible level.[38] The analysis above shows that the limitation structure of two thresholds on income are also applicable to 401(k) plans. This finding may help to determine suitable threshold levels in managing 401(k) plans.

Since the analysis above rests on the assumption that there are threshold effects, we conduct two specification tests to assess evidence for this assumption. The corresponding hypotheses are

$$\begin{aligned} H_0 &: \left(\alpha_1, \beta_1'\right)' = \left(\alpha_2, \beta_2'\right)' = \left(\alpha_3, \beta_3'\right)', \\ H_1 &: \text{at least two of } \left(\alpha_\ell, \beta_\ell'\right)', \ \ell = 1, 2, 3, \text{ are not equal.} \end{aligned}$$

The first test is based on (19) with no instruments available. We adapt both the (sup and average) Wald test and the score test to this environment, the derivations being provided in the Supplementary Material. All four tests reject the null strongly with p-values equal to zero. The second test is based on (18) with z as the instrument. Again, all four tests reject the null with zero p-values. These results strongly validate the presence of threshold effects in the data and serve to support the empirical analysis given above.

Finally, it deserves mention that OLS estimation of γ_1 and γ_2 and 2SLS estimation of α_ℓ are suited to the case where only the selection effect is present, not to cases where essential heterogeneity is also present. Notwithstanding this shortcoming, an objective function for the LATE as in Abadie (2003) which incorporates threshold effects can be used to estimate the γ and α parameters provided we use the model (19) for compliers. A formal extension of our analysis to this framework is beyond the scope of the current work.

	Regime I: $inc \leq 42.869$		Regime II: $42.869 < inc \leq 71.349$		Regime III: $inc > 71.349$	
	OLS	2SLS	OLS	2SLS	OLS	2SLS
D	9811.47	7258.49	19663.49	18164.69	29982.27	26214.79
	(1141.41)	(1342.37)	(2428.96)	(3092.96)	(9373.62)	(11641.56)
Constant	−7238.00	−7321.94	−16469.57	−16507.50	−165023.82	−163662.09
	(1013.07)	(1014.93)	(11204.50)	(11183.96)	(39491.72)	(40063.86)
Family Income	418.12	441.63	731.03	741.16	1967.02	1970.89
(in thousand $)	(47.56)	(50.48)	(168.01)	(162.89)	(451.03)	(448.38)
Age − 25	−47.94	−36.512	−551.01	−532.28	2882.54	2892.55
	(138.58)	(137.85)	(620.08)	(615.95)	(1910.19)	(1918.83)
$(\text{Age} - 25)^2$	17.58	17.25	65.34	64.87	4.68	4.18
	(4.72)	(4.70)	(20.66)	(20.55)	(54.48)	(54.94)
Married	−1446.37	−1532.38	−12534.08	−12558.78	−15314.22	−14876.92
	(1084.75)	(1089.54)	(5587.10)	(5585.97)	(17556.90)	(17614.99)
Family Size	−1152.91	−1160.58	−2198.98	−2213.39	8.09	−57.14
	(245.35)	(245.41)	(892.00)	(893.10)	(3665.470)	(3652.44)

Table 8: OLS and 2SLS Estimates of $\left(\alpha_\ell, \beta_\ell'\right)'$ in the Three Regimes Split by the Posterior Median

Note: standard errors are reported in parentheses.

[38] See http://www.irs.gov/uac/2013-Pension-Plan-Limitations, but this maximum deductible level is much higher than our $\widehat{\gamma}_2$.

8 Conclusion

Just as in conventional linear regression, endogeneity of the covariates complicates threshold regression. In both models, the complications are commonly addressed by the use of instrumentation. The present paper studies estimation and specification testing in threshold regression under endogeneity with a focus on what can be achieved without instruments.

As we have shown, it turns out that threshold points can be identified at an $O(n)$ rate and parameters of threshold effects can be identified at a nonparametric rate even when instruments are absent. This somewhat surprising finding is the direct result of the nonstationary discontinuity structure induced by threshold effects, which provides identifying information. Thus, important parameters in threshold regression are identifiable and estimable under endogeneity without instrumentation. When instruments are available, they deliver identification for the remaining structural coefficients in the usual way but play different roles for the threshold parameter and related coefficients by improving efficiency or raising convergence rates.

Our results show that it is possible to test for threshold effects in the absence of instrumentation even when endogeneity is present. We also develop a test for endogeneity, which is important in empirical work 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 improved finite sample performance a wild bootstrap procedure is suggested to obtain critical values. Our simulation results confirm the relevance of the asymptotic theory in finite samples and our empirical findings confirm the usefulness of these new procedures in detecting important threshold effects in IRA/401(k) retirement programs on savings.

As indicated earlier in the paper, both estimation and testing procedures can be extended to more general models and these can be simplified in cases where only a subset of the covariates is endogenous. There are many other relevant issues that deserve study and we conclude by outlining some of these here.

1. Assumptions H, B1 and B2 do not provide specific criteria for bandwidth selection besides the constraints on rates. Porter and Yu (2011) suggest using cross validation to select h and b in the simple case with $d = 1$. Their procedure may be extended to the more general context of the present paper at the cost of more complex analysis. In specification testing, Gao et al. (2008) suggest using an adaptive testing procedure where the test statistic is first maximized over a range of bandwidths and then a wild bootstrap procedure is used to obtain the critical values. Their test statistic takes an averaged Wald-type test statistic form. So the extension of their arguments to the degenerate U-statistic case of the present paper would need investigation.

2. The simulation studies reported here are limited in view of the time-intensive nature of the calculations. A large-scale simulation study that provides further information on the performance of the procedures and the effects of bandwidth selection would be useful.

3. The model considered here is based on threshold effects in the conditional mean. Two obvious extensions that are relevant in applications are threshold models involving conditional variances and conditional quantiles. The former extension is potentially useful in financial econometrics – see Section 7 of Tong (2011) for a review of the related time series literature. As for the latter, a parametric endogenous quantile regression model without threshold effects was considered in Chernozhukov and Hansen (2006) and applied in Chernozhukov and Hansen (2004). Also, Yu (2013b) showed how to integrate quantile difference information to improve efficiency in threshold estimation in models with no endogeneity. Combining the ideas in these literatures with those of this paper seems promising and useful for many microeconometric applications where thresholding effects are suspected.

4. This paper is based on the fixed-threshold-effect framework of Chan (1993). Using the IDKE procedure to estimate threshold points in the small-threshold-effect framework of Hansen (2000) would be a useful extension of our theory. In a fixed design model with only one covariate, Müller and Song (1997) have shown that the DKE has a similar asymptotic distribution to that of the parametric case.

5. The limit theory developed here is for i.i.d. data. Extension of our findings to stationary and ergodic time series data will be useful in many applications in macroeconomics and finance. For simple time series specifications this extension seems quite straightforward but if the covariates x and q involve lagged dependent variables, the extension is not trivial in view of the complications involved in dynamic fully nonparametric threshold autoregressions.

6. Finally, the limit theory considers only a single threshold point. This simplification in the theory was made to facilitate access to an already complex body of theory and notation. Extending our analysis to the multiple threshold case (e.g., along the lines of Bai and Perron, 1998) does not involve any fundamentally new difficulties. In fact, we already consider the two threshold points case in our application.

References

Abadie, A., 2003, Semiparametric Instrumental Variable Estimation of Treatment Response Models, *Journal of Econometrics*, 113, 231-263.

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

Ai, C. and X. Chen, 2003, Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions, *Econometrica*, 71, 1795-1843.

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

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

Angrist, J.D., K. Graddy and G.W. Imbens, 2000, The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish, *Review of Economic Studies*, 67, 499-527.

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

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

Benjamin, D.J., 2003, Does 401(k) Eligibility Increase Saving? Evidence from Propensity Score Subclassification, *Journal of Public Economics*, 87, 1259-1290.

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

Boldea, O., A.R. Hall and S. Han, 2012, Asymptotic Distribution Theory for Break Point Estimators in Models Estimated via 2SLS, *Econometric Reviews*, 31, 1-33.

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

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

Carrasco, M., J.P. Florens and E. Renault, 2007, Linear Inverse Problems in Structural Econometrics: Estimation Based on Spectral Decomposition and Regularization, *Handbook of Econometrics*, Vol. 6B, J. Heckman and E. Leamer, eds., Elsevier Science B.V., Ch. 77, 5633-5751.

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

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

Chen, B. and Y. Hong, 2012, Testing for Smooth Structural Changes in Time Series Models via Nonparametric Regression, *Econometrica*, 80, 1157-1183.

Chernozhukov, V. and C. Hansen, 2004, The Effects of 401(k) Participation on the Wealth Distribution: An Instrumental Quantile Regression Analysis, *Review of Economics and Statistics*, 86, 735-751.

Chernozhukov, V. and C. Hansen, 2006, Instrumental Quantile Regression Inference for Structural and Treatment Effect Models, *Journal of Econometrics*, 132, 491-525.

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

Darolles, S., Y. Fan, J.P. Florens and E. Renault, 2011, Nonparametric Instrumental Regression, *Econometrica*, 79, 1541-1565.

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

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

Dawid, A.P., 1979, Conditional Independence in Statistical Theory, *Journal of the Royal Statistical Society, Series B*, 41, 1-31.

De Jong, P., 1987, A Central Limit Theorem for Generalized Quadratic Forms, *Probability Theory and Related Fields*, 75, 261-277.

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

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

Employee Benefit Research Institute, 1997, Fundamentals of Employee Benefit Programs, Washington D.C.: EBRI.

Engen, E.M. and W.G. Gale, 2000, The Effects of 401(k) Plans on Household Wealth: Differences across Earnings Groups, NBER working paper #8032.

Engen, E.M., W.G. Gale and J.K. Scholz, 1996, The Illusory Effects of Saving Incentives on Saving, *Journal of Economic Perspectives*, 10, 113-138.

Fan, J., 1993, Local Linear Regression Smoothers and Their Minimax Efficiency, *Annals of Statistics*, 21, 196-216.

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

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

Florens, J.P., 2003, Inverse Problems and Structural Econometrics: The Example of Instrumental Variables, M. Dewatripont, L.P. Hansen, and S.J. Turnovsky, eds., *Advances in Economics and Econometrics: Theory and Applications*, Eighth World Congress, Vol. II, Cambridge: Cambridge University Press, 284-311.

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

Frölich, M., 2010, Regression Discontinuity Design with Covariates, IZA Discussion Paper 3024.

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

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

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

Hahn, J., P. Todd and W. Van der Klaauw, 2001, Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design, *Econometrica*, 69, 201–209.

Hall, P., 1984, Central Limit Theorem for Integrated Square Error of Multivariate Nonparametric Density Estimators, *Journal of Multivariate Analysis*, 14, 1-16.

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

Hall, P. and J.L. Horowitz, 2005, Nonparametric Methods for Inference in the Presence of Instrumental Variables, *Annals of Statistics*, 33, 2904-2929.

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

Hansen, B.E., 2000, Sample Splitting and Threshold Estimation, *Econometrica*, 68, 575-603.

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

Hansen, L.P., 1982, Large Sample Properties of Generalized Method of Moments Estimators, *Econometrica*, 50, 1029-1054.

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

Härdle, W. and T.M. Stoker, 1989, Investigating Smooth Multiple Regression by the Method of Average Derivatives, *Journal of the American Statistical Association*, 84, 986-995.

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

Heckman, J.J., 1979, Sample Selection Bias as a Specification Error, *Econometrica*, 47, 153-161.

Heckman, J.J., 1996, Randomization as an Instrumental Variable, *Review of Economics and Statistics*, 78, 336-341.

Heckman, J.J., H. Ichimura and P. Todd, 1998, Matching as an Econometric Evaluation Estimator, *Review of Economic Studies*, 65, 261–294.

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

Hsiao, C., Q Li and J.S. Racine, 2007, A Consistent Model Specification Test with Mixed Discrete and Continuous Data, *Journal of Econometrics*, 140, 802-826.

Ichimura, H., 1993, Semiparametric Least Squares Estimation of Single Index Models (SLS) and Weighted SLS Estimation of Single Index Models, *Journal of Econometrics*, 58, 72-120.

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

Kourtellos, A., T. Stengos and C.M. Tan, 2009, Structural Threshold Regression, mimeo, Department of Economics, University of Cyprus.

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

Lee, D.S. and T. Lemieux, 2010, Regression Discontinuity Designs in Economics, *Journal of Economic Literature*, 48, 281-355.

Li, Q., 1996, On the Root-N-Consistent Semiparametric Estimation of Partially Linear Models, *Economics Letters*, 51, 277-285.

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

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

Mammen, E., 1992, When Does Bootstrap Work: Asymptotic Results and Simulations, *Lecture Notes in Statistics*, 77 Springer, Berlin.

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

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

Newey, W.K., 1994, Kernel Estimation of Partial Means and a General Variance Estimator, *Econometric Theory*, 10, 233-253.

Newey, W.K. and D.L. McFadden, 1994, Large Sample Estimation and Hypothesis Testing, *Handbook of Econometrics*, Vol. 4, R.F. Eagle and D.L. McFadden, eds., New York: Elsevier Science B.V., Ch. 36, 2111-2245.

Newey, W.K. and J.L. Powell, 2003, Instrumental Variables Estimation for Nonparametric Models, *Econometrica*, 71, 1565-1578.

Newey, W.K., J.L. Powell and F. Vella, 1999, Nonparametric Estimation of Triangular Simultaneous Equations Models, *Econometrica*, 67, 565-603.

Newey, W. and T.M. Stoker, 1993, Efficiency of Weighted Average Derivative Estimators and Index Models, *Econometrica*, 61, 1199-1223.

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

Perron, P. and Y. Yamamoto, 2012a, A Note on Estimating and Testing for Multiple Structural Changes in Models with Endogenous Regressors via 2SLS, forthcoming in *Econometric Theory*.

Perron, P. and Y. Yamamoto, 2012b, Using OLS to Estimate and Test for Structural Changes in Models with Endogenous Regressors, forthcoming in *Journal of Applied Econometrics*.

Phillips, P. and L. Su, 2011, Non-parametric Regression Under Location Shifts, *Econometrics Journal*, 14, 457-486.

Pollard, D., 1993, Asymptotics for a Binary Choice Model, Preprint, Department of Statistics, Yale University.

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

Porter, J. and P. Yu, 2011, Regression Discontinuity with Unknown Discontinuity Points: Testing and Estimation, Mimeo, Department of Economics, University of Auckland.

Poterba, J.M., S.F. Venti and D.A. Wise, 1994, 401(k) Plans and Tax-deferred Savings, *Studies in the Economics of Aging*, D.A. Wise, eds., Chicago: University of Chicago Press, 105-142.

Poterba, J.M., S.F. Venti and D.A. Wise, 1995, Do 401(k) Contributions Crowd Out Other Personal Saving?, *Journal of Public Economics*, 58, 1-32.

Poterba, J.M., S.F. Venti and D.A. Wise, 1998, Personal Retirement Saving Programs and Asset Accumulation: Reconciling the Evidence, Frontiers in the Economics of Aging, D.A. Wise, eds., Chicago: University of Chicago Press, 23-124.

Potter, S.M., 1995, A Nonlinear Approach to US GNP, *Journal of Applied Econometrics*, 10, 109-125.

Powell, J.L., 1987, Semiparametric Estimation of Bivariate Latent Variable Models, Working Paper No. 8704, Revised April 1989, SSRI, University of Wisconsin, Madison.

Powell, J.L., 2001, Semiparametric Estimation of Censored Selection Models, C. Hsiao, K. Morimune, and J.L. Powell, eds., *Nonlinear Statistical Modeling*, New York: Cambridge University Press.

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

Qin, J. and J. Lawless, 1994, Empirical Likelihood and General Estimating Equations, *Annals of Statistics*, 22, 300-325.

Qiu, P., C. Asano and X. Li, 1991, Estimation of Jump Regression Function, *Bulletin of Informatics and Cybernetics*, 24, 197-212.

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

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

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

Stoker, T.M., 1986, Consistent Estimation of Scaled Coefficients, *Econometrica*, 1461-1481.

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

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

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

Sun, Y.X., 2005, Adaptive Estimation of the Regression Discontinuity Model, mimio, Department of Economics, University of California, San Diego.

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

Tong, H., 2011, Threshold Models in Time Series Analysis - 30 Years On, *Statistics and Its Interface*, 4, 107-118.

Van der Klaauw, W., 2002, Estimating the Effect of Financial Aid Offers on College Enrollment: a Regression-Discontinuity Approach, *International Economic Review*, 43, 1249-1287.

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

Wald, A., 1940, The Fitting of Straight Lines if Both Variables are Subject to Error, *Annals of Mathematical Statistics*, 11, 284-300.

Wang, Q. and P. Phillips, 2009, Structural Nonparametric Cointegrating Regression, *Econometrica*, 77, 1901-1948.

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

Yu, P., 2008, Adaptive Estimation of the Threshold Point in Threshold Regression, forthcoming, *Journal of Econometrics*.

Yu, P., 2009, The Bootstrap in Threshold Regression, forthcoming, *Econometric Theory*.

Yu, P., 2010, Understanding Estimators of Treatment Effects in Regression Discontinuity Designs, forthcoming, *Econometric Reviews*.

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

Yu, P., 2012b, Inconsistency of 2SLS Estimators in Threshold Regression with Endogeneity, forthcoming, *Economics Letters*.

Yu, P., 2013a, Threshold Regression Under Misspecification, mimeo, Department of Economics, University of Auckland.

Yu, P., 2013b, Integrated Quantile Threshold Regression and Distributional Threshold Effects, mimeo, Department of Economics, University of Auckland.

Yu, P., 2014, A Note on Average Derivative Estimation, mimeo, Department of Economics, University of Auckland.

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

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

Appendix A: Proofs

In the following proofs, some steps are omitted for brevity whenever they are available in the literature and references are provided. This simplification makes the proofs cleaner and more readable. Derivations that differ from the existing literature are given in full detail.

Proof of Theorem 1. Proposition 1 proves the consistency of $\widehat{\gamma}$, and Proposition 2 proves $\widehat{\gamma} - \gamma_0 = O_p(n^{-1})$, so we can apply the argmax continuous mapping theorem (see, e.g., Theorem 3.2.2 of Van der Vaart and Wellner (1996)) to establish the asymptotic distribution of $n(\widehat{\gamma} - \gamma_0)$. From Proposition 3, for v in any compact set of $\mathbb{R}$,

$$
\begin{aligned}
& nh\left(\widehat{Q}_n\left(\gamma_0 + \frac{v}{n}\right) - \widehat{Q}_n(\gamma_0)\right)\Big/ 2k_+(0) \\
= \quad & -\sum_{i=1}^{n} \overline{z}_{1i} 1\left(\gamma_0 - \frac{v}{n} < q_i \leq \gamma_0\right) - \sum_{i=1}^{n} \overline{z}_{2i} 1\left(\gamma_0 < q_i \leq \gamma_0 + \frac{v}{n}\right) + o_p(1),
\end{aligned}
$$

where $\overline{z}_{1i}$ and $\overline{z}_{2i}$ are defined in the main text. Now, we can obtain the asymptotic distribution of $n(\widehat{\gamma} - \gamma_0)$ by applying the same argument as in the proofs of Theorem 1 and 2 in Yu (2012a). The only difference lies in the definitions of $\overline{z}_{1i}$ and $\overline{z}_{2i}$. ■

Proof of Corollary 1. The proofs of the consistency of $\widetilde{\gamma}$ and $nh^{d-1}(\widetilde{\gamma}-\gamma_0) = O_p(1)$ are similar to Theorem 1, so are omitted here. We concentrate on deriving the weak limit of the localized process $nh^d\left(\widehat{\Delta}_o^2(\gamma)-\widehat{\Delta}_o^2(\gamma_0)\right)$ for γ in an $\left(nh^{d-1}\right)^{-1}$ neighborhood of γ_0 .

Let $a_n = nh^{d-1}(= o(h))$, then

$$nh^d\left(\widehat{\Delta}_o^2\left(\gamma_0+\frac{v}{a_n}\right)-\widehat{\Delta}_o^2(\gamma_0)\right) = \left(\widehat{\Delta}_o\left(\gamma_0+\frac{v}{a_n}\right)+\widehat{\Delta}_o(\gamma_0)\right)nh^d\left(\widehat{\Delta}_o\left(\gamma_0+\frac{v}{a_n}\right)-\widehat{\Delta}_o(\gamma_0)\right).$$

It is easy to show that $\widehat{\Delta}_o\left(\gamma_0+\frac{v}{a_n}\right)-\mathbb{E}\left[\widehat{\Delta}_o\left(\gamma_0+\frac{v}{a_n}\right)\right]\xrightarrow{p}0$ for v in any compact set. Without loss of generality, let $\gamma>\gamma_0$ or $v>0$. Then

$$\begin{aligned}
&\mathbb{E}\left[\widehat{\Delta}_o(\gamma)\right]\\
=\;&\int_{-1}^{0}\int K^x(u_x,x_o)k_-(u_q)g(x_o+u_xh,\gamma+u_qh)f(x_o+u_xh,\gamma+u_qh)du_xdu_q\\
&+\int_{-1}^{\frac{\gamma_0-\gamma}{h}}\int K^x(u_x,x_o)k_-(u_q)\left(1,(x_o+u_xh)',\gamma+u_qh\right)\delta_0 f(x_o+u_xh,\gamma+u_qh)du_xdu_q\\
&-\int_{0}^{1}\int K^x(u_x,x_o)k_+(u_q)g(x_o+u_xh,\gamma+u_qh)f(x_o+u_xh,\gamma+u_qh)du_xdu_q\\
=\;&(1,x_o',\gamma_0)\,\delta_0 f(x_o,\gamma_0)+O(h).
\end{aligned}$$

Now, we need only consider the behavior of $nh^d\left(\widehat{\Delta}_o\left(\gamma_0+\frac{v}{a_n}\right)-\widehat{\Delta}_o(\gamma_0)\right)$. Proposition 4 shows that

$$nh^d\left(\widehat{\Delta}_o\left(\gamma_0+\frac{v}{a_n}\right)-\widehat{\Delta}_o(\gamma_0)\right)\Rightarrow D_o(v),$$

where $\Rightarrow$ signifies the weak convergence on a compact set of v,

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

is a cadlag process with $D_o(0)=0$,

$$z_{1i}=\left(-2e_i^- -(1,x_o',\gamma_0)\,\delta_0\right)K(U_i^-)k_-(0)\,,\; z_{2i}=\left(2e_i^+ -(1,x_o',\gamma_0)\,\delta_0\right)K(U_i^+)k_+(0)\,,$$

and the distributions of $e_i^-, e_i^+, U_i^-, U_i^+$ are defined in the corollary. So

$$nh^d\left(\widehat{\Delta}_o^2\left(\gamma_0+\frac{v}{a_n}\right)-\widehat{\Delta}_o^2(\gamma_0)\right)\Rightarrow\overline{D}(v),$$

where $\overline{D}(v)$ takes a similar form to $D_o(v)$, but now

$$z_{1i}=2\left(-2\,(1,x_o',\gamma_0)\,\delta_0 e_i^- -\delta_0'\,(1,x_o',\gamma_0)'\,(1,x_o',\gamma_0)\,\delta_0\right)K(U_i^-)f(x_o,\gamma_0)k_-(0)\,,$$

and

$$z_{2i}=2\left(-2\,(1,x_o',\gamma_0)\,\delta_0 e_i^+ -\delta_0'\,(1,x_o',\gamma_0)'\,(1,x_o',\gamma_0)\,\delta_0\right)K(U_i^+)f(x_o,\gamma_0)k_+(0)\,.$$

Given the weak limit of $nh^d\left(\widehat{\Delta}_o^2\left(\gamma_0+\frac{v}{a_n}\right)-\widehat{\Delta}_o^2(\gamma_0)\right)$, we can apply the argmax continuous mapping theorem (Theorem 3.2.2 in Van der Vaart and Wellner, 1996) to obtain the asymptotic distribution of $\widetilde{\gamma}$. We need to check four conditions, just as in the proof of Theorem 2 of Yu (2012a). Since these checks are all similar, we omit the details here and only note that $\arg\max_v \overline{D}(v)=\arg\min_v D(v)$, given that $k_-(0)=k_+(0)>0$ and $f(x_o,\gamma_0)>0$. ■

Proof of Theorem 2. We first derive the formula for $\widehat{\delta}$. Following Appendix A.1 of Heckman et al. (1998), we have

$$\begin{aligned}\left(\widehat{a}_+(x_i),\widehat{b}'_+(x_i)\right)' &= (I_{d+1},\mathbf{0})\left(X_i'W_i^+X_i\right)^{-1}X_i'W_i^+Y\\ &= (I_{d+1},\mathbf{0})\,H^{-1}\left(H^{-1}X_i'W_i^+X_iH^{-1}\right)^{-1}H^{-1}X_i'W_i^+Y\\ &= (I_{d+1},\mathbf{0})\,H^{-1}\left(Z_i'W_i^+Z_i\right)^{-1}Z_i'W_i^+Y\\ &= (I_{d+1},\mathbf{0})\,H^{-1}\left(\frac{1}{n}\sum_{j=1,j\neq i}^{n}z_j^iz_j^{i\prime}w_j^{i+}\right)^{-1}\left(\frac{1}{n}\sum_{j=1,j\neq i}^{n}z_j^iw_j^{i+}y_j\right)\\ &\equiv (I_{d+1},\mathbf{0})\,H^{-1}\left(M_i^+\right)^{-1}r_i^+,\end{aligned}$$

where

$$X_i=\begin{pmatrix}(x_1-x_i,q_1-\widehat{\gamma})^{S_p}\\ \vdots\\ (x_{i-1}-x_i,q_{i-1}-\widehat{\gamma})^{S_p}\\ \mathbf{0}\\ (x_{i+1}-x_i,q_{i+1}-\widehat{\gamma})^{S_p}\\ \vdots\\ (x_n-x_i,q_n-\widehat{\gamma})^{S_p}\end{pmatrix}_{n\times\sum_{\nu=0}^{p}(\nu+d-1)!/\nu!(d-1)!},\ Y=\begin{pmatrix}y_1\\ \vdots\\ y_{i-1}\\ 0\\ y_{i+1}\\ \vdots\\ y_n\end{pmatrix}_{n\times 1},$$

$$\begin{aligned}H &= \mathrm{diag}\left\{1,hI_d,\cdots,hI_{(p+d-1)!/p!(d-1)!}\right\},Z_i=X_iH^{-1},z_j^{i\prime}=(x_j-x_i,q_j-\widehat{\gamma})^{S_p}H^{-1},\\ W_i^+ &= \mathrm{diag}\left\{K_h^x(x_1-x_i,x_i)\,k_h^+(q_1-\widehat{\gamma}),\cdots,K_h^x(x_n-x_i,x_i)\,k_h^+(q_n-\widehat{\gamma})\right\}=\mathrm{diag}\left\{w_1^{i+},\cdots,w_n^{i+}\right\}_{n\times n},\end{aligned}$$

and $\left(\widehat{a}_-(x_i),\widehat{b}'_-(x_i)\right)'$ are similarly defined with

$$W_i^-=diag\left\{K_h^x(x_1-x_i,x_i)\,k_h^-(q_1-\widehat{\gamma}),\cdots,K_h^x(x_n-x_i,x_i)\,k_h^-(q_n-\widehat{\gamma})\right\}$$

replacing W_i^+. Next

$$\begin{pmatrix}\widehat{\Delta}\\ \widehat{\delta}_{xq}\end{pmatrix}=\frac{1}{nh}\sum_{i=1}^{n}k\left(\frac{q_i-\widehat{\gamma}}{h}\right)\begin{pmatrix}\widehat{a}_-(x_i)-\widehat{a}_+(x_i)\\ \widehat{b}_-(x_i)-\widehat{b}_+(x_i)\end{pmatrix}\Bigg/\frac{1}{nh}\sum_{i=1}^{n}k\left(\frac{q_i-\widehat{\gamma}}{h}\right),$$

and

$$\begin{pmatrix}\widehat{\delta}_\alpha\\ \widehat{\delta}_{xq}\end{pmatrix}=\begin{pmatrix}\widehat{\Delta}-\frac{1}{nh}\sum_{i=1}^{n}k\left(\frac{q_i-\widehat{\gamma}}{h}\right)(x_i',\widehat{\gamma})\left(\widehat{b}_-(x_i)-\widehat{b}_+(x_i)\right)\Big/\frac{1}{nh}\sum_{i=1}^{n}k\left(\frac{q_i-\widehat{\gamma}}{h}\right)\\ \widehat{\delta}_{xq}\end{pmatrix}.$$

The first step in deriving the asymptotic distribution of $\widehat{\delta}$ is to show that $\widehat{\gamma}$ can be replaced by γ_0. In other words,

$$\sqrt{nh}h\begin{pmatrix}\widehat{\delta}_{\alpha}-\widehat{\delta}^{0}_{\alpha}\\ \widehat{\delta}_{xq}-\widehat{\delta}^{0}_{xq}\end{pmatrix}\xrightarrow{p}0,$$

where estimators with superscript 0 denotes the original estimators but with $\widehat{\gamma}$ replaced by γ_0. Of course, we need only show that $\begin{pmatrix}\sqrt{nh}\left(\widehat{\Delta}-\widehat{\Delta}^{0}\right)\\ \sqrt{nh}h\left(\widehat{\delta}_{xq}-\widehat{\delta}^{0}_{xq}\right)\end{pmatrix}\xrightarrow{p}0$ since $\sqrt{nh}h\begin{pmatrix}\widehat{\delta}_{\alpha}-\widehat{\delta}^{0}_{\alpha}\\ \widehat{\delta}_{xq}-\widehat{\delta}^{0}_{xq}\end{pmatrix}$ is just a linear combination of $\begin{pmatrix}\sqrt{nh}\left(\widehat{\Delta}-\widehat{\Delta}^{0}\right)\\ \sqrt{nh}h\left(\widehat{\delta}_{xq}-\widehat{\delta}^{0}_{xq}\right)\end{pmatrix}$. Proposition 5 gives this result. Now,

$$\begin{aligned}
&\begin{pmatrix}\sqrt{nh}\left(\widehat{\Delta}^{0}-\overline{\Delta}^{0}\right)\\ \sqrt{nh}h\left(\widehat{\delta}^{0}_{xq}-\delta_{xq0}\right)\end{pmatrix}\\
=\;&\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}k\left(\frac{q_i-\gamma_0}{h}\right)\begin{pmatrix}\widehat{a}^{0}_{-}(x_i)-\widehat{a}^{0}_{+}(x_i)-\left(a^{0}_{-}(x_i)-a^{0}_{+}(x_i)\right)\\ h\left(\widehat{b}^{0}_{-}(x_i)-\widehat{b}^{0}_{+}(x_i)-\delta_{xq0}\right)\end{pmatrix}\Bigg/\frac{1}{nh}\sum_{i=1}^{n}k\left(\frac{q_i-\gamma_0}{h}\right)\\
=\;&\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}k\left(\frac{q_i-\gamma_0}{h}\right)\begin{pmatrix}\left(\widehat{a}^{0}_{-}(x_i)-a^{0}_{-}(x_i)\right)-\left(\widehat{a}^{0}_{+}(x_i)-a^{0}_{+}(x_i)\right)\\ h\left[\left(\widehat{b}^{0}_{-}(x_i)-b^{0}_{-}(x_i)\right)-\left(\widehat{b}^{0}_{+}(x_i)-b^{0}_{+}(x_i)\right)\right]\end{pmatrix}\Bigg/\frac{1}{nh}\sum_{i=1}^{n}k\left(\frac{q_i-\gamma_0}{h}\right),
\end{aligned}$$

where $\left(a^{0}_{+}(x_i),b^{0}_{+}(x_i)'\right)$ is defined by $a^{0}_{+}(x_i)=m_{+}(x_i)\equiv\lim_{\gamma\to\gamma_0+}m(x_i,\gamma)$ and by $b^{0}_{+}(x_i)=\nabla m_{+}(x_i)\equiv\lim_{\gamma\to\gamma_0+}\left(\partial m(x_i,\gamma)/\partial x_i',\,\partial m(x_i,\gamma)/\partial\gamma\right)'$, with a similar definition for $\left(a^{0}_{-}(x_i),b^{0}_{-}(x_i)'\right)$, and

$$\overline{\Delta}^{0}=\frac{1}{nh}\sum_{i=1}^{n}k\left(\frac{q_i-\gamma_0}{h}\right)\left(a^{0}_{-}(x_i)-a^{0}_{+}(x_i)\right)\Bigg/\frac{1}{nh}\sum_{i=1}^{n}k\left(\frac{q_i-\gamma_0}{h}\right).$$

Note that, under our specification, $b^{0}_{-}(x_i)-b^{0}_{+}(x_i)=\delta_{xq0}$ and $a^{0}_{-}(x_i)-a^{0}_{+}(x_i)-(x_i',\gamma_0)(b^{0}_{-}(x_i)-b^{0}_{+}(x_i))=\delta_{\alpha0}$ for any x_i. Also,

$$\begin{aligned}
&\sqrt{nh}h\left(\widehat{\delta}^{0}_{\alpha}-\delta_{\alpha0}\right)\\
=\;&\sqrt{nh}h\left(\widehat{\Delta}^{0}-\overline{\Delta}^{0}\right)-\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}k\left(\frac{q_i-\gamma_0}{h}\right)(x_i',\gamma_0)\left[h\left(\widehat{b}^{0}_{-}(x_i)-b^{0}_{-}(x_i)\right)-h\left(\widehat{b}^{0}_{+}(x_i)-b^{0}_{+}(x_i)\right)\right]\Bigg/\frac{1}{nh}\sum_{i=1}^{n}k\left(\frac{q_i-\gamma_0}{h}\right).
\end{aligned}$$

We first derive the asymptotic distribution of $\begin{pmatrix}\sqrt{nh}\left(\widehat{\Delta}^{0}-\overline{\Delta}^{0}\right)\\ \sqrt{nh}h\left(\widehat{\delta}^{0}_{xq}-\delta_{xq0}\right)\end{pmatrix}$ and then consider $\sqrt{nh}h\left(\widehat{\delta}^{0}_{\alpha}-\delta_{\alpha0}\right)$.

Given assumptions E, F$'$, G$'$, and H$'$, we can apply the arguments in Theorem 3 of Heckman et al. (1998) to obtain

$$\begin{aligned}
&\begin{pmatrix}\sqrt{nh}\left(\widehat{\Delta}^{0}-\overline{\Delta}^{0}\right)\\ \sqrt{nh}h\left(\widehat{\delta}^{0}_{xq}-\delta_{xq0}\right)\end{pmatrix}\\
=\;&-\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}k\left(\frac{q_i-\gamma_0}{h}\right)(I_{d+1},\mathbf{0})\left\{\begin{array}{c}h^{p+1}\left[\left(\overline{M}^{+}_{i0}\right)^{-1}\overline{r}^{m+}_{i0}-\left(\overline{M}^{-}_{i0}\right)^{-1}\overline{r}^{m-}_{i0}\right]\\ +\left[\left(\overline{M}^{+}_{i0}\right)^{-1}r^{e+}_{i0}-\left(\overline{M}^{-}_{i0}\right)^{-1}r^{e-}_{i0}\right]+R^{+}_{i}-R^{-}_{i}\end{array}\right\}\Bigg/\frac{1}{nh}\sum_{i=1}^{n}k\left(\frac{q_i-\gamma_0}{h}\right),
\end{aligned}$$

where $\overline{M}_{i0}^{+}$ is the square matrix of size $\sum_{\nu=0}^{p}(\nu+d-1)!/\nu!(d-1)!$ with the l-th row, t-th column "block" being, for $0 \leq l, t \leq p$,

$$\int_0^\infty \int \left(u_x', u_q\right)^{S(l)\prime} \left(u_x', u_q\right)^{S(t)} K^x\left(u_x, x_i\right) k_+(u_q) f(x_i + u_x h, \gamma_0 + u_q h) du_x du_q,$$

$\overline{r}_{i0}^{m+}$ is a $\sum_{\nu=0}^{p}(\nu+d-1)!/\nu!(d-1)!$ by 1 vector with the t-th block being the transpose of

$$\int \left(u_x', u_q\right)^{S(t)} \left[\left(u_x', u_q\right)^{S(p+1)} m_+^{(p+1)}(x_i)\right] K^x\left(u_x, x_i\right) k_+(u_q) f(x_i, \gamma_0) du_x du_q,$$

and $m_+^{(p+1)}(x)$ being a $(p+d)!/(p+1)!(d-1)! \times 1$ vector of the partial derivatives of $m(x,q)$ at $q = \gamma_0+$,

$$r_{i0}^{e+} = \frac{1}{n} \sum_{j=1, j\neq i}^{n} z_{j0}^i w_{j0}^{i+} e_j,$$

with z_{j0}^i and w_{j0}^{i+} being z_j^i and w_j^{i+} but having $\widehat{\gamma}$ replaced by γ_0,

$$\frac{1}{\sqrt{nh}} \sum_{i=1}^{n} k\left(\frac{q_i - \gamma_0}{h}\right) \left(I_{d+1}, \mathbf{0}\right) R_i^+ = o_p(1),$$

and the objects with superscript $-$ are similarly defined. It turns out that the terms associated with $\overline{r}_{i0}^{m\pm}$ will contribute to the bias and the terms associated with $\overline{r}_{i0}^{e\pm}$, which is a U-statistic, will contribute to the variance. Given that $\frac{1}{nh}\sum\limits_{i=1}^{n} k\left(\frac{q_i-\gamma_0}{h}\right) \xrightarrow{p} f_q(\gamma_0)$, we need only concentrate on the numerator.

First, analyze the bias.

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{q_i-\gamma_0}{h}\right)\left(I_{d+1}, \mathbf{0}\right)\left[\left(\overline{M}_{i0}^{+}\right)^{-1} \overline{r}_{i0}^{m+} - \left(\overline{M}_{i0}^{-}\right)^{-1} \overline{r}_{i0}^{m-}\right]\right] \\
= \quad & \left(I_{d+1}, \mathbf{0}\right) \int\int \left[\left(\overline{M}_{i0}^{+}\right)^{-1} \overline{r}_{i0}^{m+} - \left(\overline{M}_{i0}^{-}\right)^{-1} \overline{r}_{i0}^{m-}\right] f(x_i|q_i) dx_i k_h\left(q_i - \gamma_0\right) f(q_i) dq_i \\
\to \quad & \left(I_{d+1}, \mathbf{0}\right)\left[\left(M_o^+\right)^{-1} B^+ - \left(M_o^-\right)^{-1} B^-\right] \mathbb{E}[g^{(p+1)}(x,\gamma_0)\big| q=\gamma_0] f_q(\gamma_0),
\end{aligned}$$

where $M_o^{\pm}$ and $B^{\pm}$ are defined in the main text, and $m_+^{(p+1)}(x_i) = m_-^{(p+1)}(x_i) = g^{(p+1)}(x_i, \gamma_0)$ under Assumption G$'$. Note here that the kernel K^x is replaced by K because the data in the h neighborhood of the boundary of $\mathcal{X}$ can be neglected asymptotically. Also, we can calculate that the variance of this term is $O\left(\frac{1}{nh}\right) = o(1)$, so it converges in probability to its expectation. Second, analyze the variance. Taking the lth element of $\begin{pmatrix} \sqrt{nh}\left(\widehat{\Delta}^0 - \overline{\Delta}^0\right) \\ \sqrt{nh}h\left(\widehat{\delta}_{xq}^0 - \delta_{xq0}\right) \end{pmatrix}$, we consider

$$\frac{1}{\sqrt{nh}} \sum_{i=1}^{n} k\left(\frac{q_i-\gamma_0}{h}\right) \mathbf{e}_l' \left[\left(\overline{M}_{i0}^{+}\right)^{-1} r_{i0}^{e+} - \left(\overline{M}_{i0}^{-}\right)^{-1} r_{i0}^{e-}\right],$$

which is a second-order U-statistic. From Lemma 8.4 of Newey and McFadden (1994), this U-statistic is asymptotically equivalent to $\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} m_n(x_i, q_i, e_i)$, where

$$\begin{aligned} m_n(x_j,q_j,e_j) &= \mathbb{E}\left[k\left(\frac{q_i-\gamma_0}{h}\right)\mathbf{e}_l'\left[\left(\overline{M}_{i0}^{+}\right)^{-1} z_{j0}^i w_{j0}^{i+} e_j - \left(\overline{M}_{i0}^{-}\right)^{-1} z_{j0}^i w_{j0}^{i-} e_j\right]\middle| x_j,q_j,e_j\right] \\ &= e_j \int k\left(\frac{q_i-\gamma_0}{h}\right)\mathbf{e}_l'\left[\left(\overline{M}_{i0}^{+}\right)^{-1} z_{j0}^i w_{j0}^{i+} - \left(\overline{M}_{i0}^{-}\right)^{-1} z_{j0}^i w_{j0}^{i-}\right] f(x_i,q_i)dx_i dq_i, \end{aligned}$$

We apply the Liapunov central limit theorem to derive the asymptotic distribution. It is standard to check that the Liapunov condition is satisfied, so we concentrate on calculating the asymptotic variance as follows.

$$\begin{aligned} & \frac{1}{h}\mathbb{E}\left[e_j^2\left(\int k\left(\frac{q_i-\gamma_0}{h}\right)\mathbf{e}_l'\left[\left(\overline{M}_{i0}^{+}\right)^{-1} z_{j0}^i w_{j0}^{i+} - \left(\overline{M}_{i0}^{-}\right)^{-1} z_{j0}^i w_{j0}^{i-}\right] f(x_i,q_i)dx_i dq_i\right)^2\right] \\ \approx\ & \frac{1}{h}\mathbb{E}\left[e_j^2\left(\int k\left(\frac{q_i-\gamma_0}{h}\right)\mathbf{e}_l'\left(\overline{M}_{i0}^{+}\right)^{-1} z_{j0}^i w_{j0}^{i+} f(x_i,q_i)dx_i dq_i\right)^2\right] \\ & +\frac{1}{h}\mathbb{E}\left[e_j^2\left(\int k\left(\frac{q_i-\gamma_0}{h}\right)\mathbf{e}_l'\left(\overline{M}_{i0}^{-}\right)^{-1} z_{j0}^i w_{j0}^{i-} f(x_i,q_i)dx_i dq_i\right)^2\right] \\ & -\frac{2}{h}\mathbb{E}\left[\begin{array}{c} e_j^2\left(\int k\left(\frac{q_i-\gamma_0}{h}\right)\mathbf{e}_l'\left(\overline{M}_{i0}^{+}\right)^{-1} z_{j0}^i w_{j0}^{i+} f(x_i,q_i)dx_i dq_i\right) \\ \left(\int k\left(\frac{q_i-\gamma_0}{h}\right)\mathbf{e}_l'\left(\overline{M}_{i0}^{-}\right)^{-1} z_{j0}^i w_{j0}^{i-} f(x_i,q_i)dx_i dq_i\right) \end{array}\right] \\ \equiv\ & T1+T2+T3. \end{aligned}$$

We analyze $T1$, $T2$ and $T3$ in turn.

$$\begin{aligned} T1 &\approx \frac{1}{h}\mathbb{E}\left[e_j^2\left(k_+\left(\frac{q_j-\gamma_0}{h}\right)\int k\left(u_q\right)\mathbf{e}_l'\left(M_o^{+}\right)^{-1}\left[\left(u_x',\frac{q_j-\gamma_0}{h}\right)^{S_p}\right]' K\left(u_x\right)du_x du_q\right)^2\right] \\ &\approx \int\int \sigma_+^2(x_j)\left(k_+\left(v_q\right)\int k\left(u_q\right)\mathbf{e}_l'\left(M_o^{+}\right)^{-1}\left[\left(u_x',v_q\right)^{S_p}\right]' K\left(u_x\right)du_x du_q\right)^2 f(x_j,\gamma_0)dx_j dv_q \\ &= \mathbb{E}\left[\int k_+^2\left(v_q\right)\sigma_+^2(x_j)C_l^+(v_q)^2 dv_q | q_j=\gamma_0\right] f_q(\gamma_0). \end{aligned}$$

Similarly,

$$T2 \approx \mathbb{E}\left[\int k_-^2\left(v_q\right)\sigma_-^2(x_j)C_l^-(v_q)^2 dv_q | q_j=\gamma_0\right] f_q(\gamma_0),$$

and $T3=0$ since $k_+\left(\frac{q_j-\gamma_0}{h}\right)k_-\left(\frac{q_j-\gamma_0}{h}\right)=0$. In summary,

$$\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} m_n(x_i,q_i,e_i) \xrightarrow{d} N\left(0,\mathbb{E}\left[\int\left[k_+^2\left(v_q\right)\sigma_+^2(x)C_l^+(v_q)^2 + k_-^2\left(v_q\right)\sigma_-^2(x)C_l^-(v_q)^2\right]dv_q\middle| q=\gamma_0\right] f_q(\gamma_0)\right),$$

and the asymptotic distribution of $\sqrt{nh}h\left(\widehat{\delta}_{xq}^0-\delta_{xq0}\right)$ follows as in the theorem.

We next derive the asymptotic distribution of $\sqrt{nh}h\left(\widehat{\delta}_{\alpha}^0-\delta_{\alpha 0}\right)$. Given that $\sqrt{nh}\left(\widehat{\Delta}^0-\overline{\Delta}^0\right)=O_p(1)$ under Assumption H$'$, the term $\sqrt{nh}h\left(\widehat{\Delta}^0-\overline{\Delta}^0\right)$ can be neglected, and $\sqrt{nh}h\left(\widehat{\delta}_{\alpha}^0-\delta_{\alpha 0}\right)$ has the same asymptotic distribution as

$$-\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}k\left(\frac{q_i-\gamma_0}{h}\right)(x_i',\gamma_0)\left[h\left(\widehat{b}_-^0(x_i)-b_-^0(x_i)\right)-h\left(\widehat{b}_+^0(x_i)-b_+^0(x_i)\right)\right]\Big/\frac{1}{nh}\sum_{i=1}^{n}k\left(\frac{q_i-\gamma_0}{h}\right).$$

For the bias, note that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{nh}\sum_{i=1}^{n}k\left(\frac{q_i-\gamma_0}{h}\right)(x_i',\gamma_0)(\mathbf{0},I_d,\mathbf{0})\left[\left(\overline{M}_{i0}^{+}\right)^{-1}\overline{r}_{i0}^{m+}-\left(\overline{M}_{i0}^{-}\right)^{-1}\overline{r}_{i0}^{m-}\right]\right]\\
=&\int\int(x_i',\gamma_0)(\mathbf{0},I_d,\mathbf{0})\left[\left(\overline{M}_{i0}^{+}\right)^{-1}\overline{r}_{i0}^{m+}-\left(\overline{M}_{i0}^{-}\right)^{-1}\overline{r}_{i0}^{m-}\right]f(x_i|q_i)dx_ik_h(q_i-\gamma_0)f(q_i)dq_i\\
\to&\ \mathbb{E}\left[(x',\gamma_0)(\mathbf{0},I_d,\mathbf{0})\left[\left(M_o^{+}\right)^{-1}B^{+}-\left(M_o^{-}\right)^{-1}B_{-}\right]g^{(p+1)}(x,\gamma_0)\Big|q=\gamma_0\right]f_q(\gamma_0).
\end{aligned}$$

For the variance, the corresponding U-projection $m_n(x_i,q_i,e_i)$ is

$$e_j\int k_h\left(\frac{q_i-\gamma_0}{h}\right)(x_i',\gamma_0)(\mathbf{0},I_d,\mathbf{0})\left[\left(\overline{M}_{i0}^{+}\right)^{-1}z_{j0}^{i}w_{j0}^{i+}-\left(\overline{M}_{i0}^{-}\right)^{-1}z_{j0}^{i}w_{j0}^{i-}\right]f(x_i,q_i)dx_idq_i.$$

We can proceed in a similar fashion to the above in deriving the asymptotic variance. For example, the corresponding form to $T1$ is

$$\begin{aligned}
T1\ \approx&\ \frac{1}{h}\mathbb{E}\left[e_j^2\left(k_h^{+}(q_j-\gamma_0)\int k(u_q)(x_j',\gamma_0)(\mathbf{0},I_d,\mathbf{0})\left(M_o^{+}\right)^{-1}\left[\left(u_x',\frac{q_j-\gamma_0}{h}\right)^{S_p}\right]'K(u_x)du_xdu_q\right)^2\right]\\
\approx&\int\int\sigma_+^2(x_j)\left(k_+(v_q)\int k(u_q)(x_j',\gamma_0)(\mathbf{0},I_d,\mathbf{0})\left(M_o^{+}\right)^{-1}\left[(u_x'\ v_q)^{S_p}\right]'K(u_x)du_xdu_q\right)^2f(x_j,\gamma_0)dx_jdv_q\\
=&\ \mathbb{E}\left[\int k_+^2(v_q)\sigma_+^2(x_j)C^{+}(x_j,v_q)^2dv_q|q_j=\gamma_0\right]f_q(\gamma_0).
\end{aligned}$$

■

Proof of Corollary 2. The asymptotic distribution of $\sqrt{nh}\left(\widehat{\Delta}^0-\Delta_0\right)$ is more involved since it includes variations from two components as in

$$\sqrt{nh}\left(\widehat{\Delta}^0-\Delta_0\right)=\sqrt{nh}\left(\widehat{\Delta}^0-\overline{\Delta}^0\right)+\sqrt{nh}\left(\overline{\Delta}^0-\Delta_0\right).$$

First note that

$$\begin{aligned}
\sqrt{nh}\left(\widehat{\Delta}^0-\Delta_0\right)=&\ \sqrt{nh}\left(\frac{\widehat{\Delta}_N^0}{\widehat{f}_q(\gamma_0)}-\frac{\Delta_0f_q(\gamma_0)}{f_q(\gamma_0)}\right)\\
\approx&\ \frac{\sqrt{nh}\left[\widehat{\Delta}_N^0-\Delta_0f_q(\gamma_0)\right]-\sqrt{nh}\Delta_0\left[\widehat{f}_q(\gamma_0)-f_q(\gamma_0)\right]}{f_q(\gamma_0)}\\
=&\ \frac{\sqrt{nh}\left[\widehat{\Delta}_N^0-\overline{\Delta}_N^0\right]+\sqrt{nh}\left[\overline{\Delta}_N^0-\Delta_0f_q(\gamma_0)\right]-\sqrt{nh}\Delta_0\left[\widehat{f}_q(\gamma_0)-f_q(\gamma_0)\right]}{f_q(\gamma_0)}
\end{aligned}$$

where $\widehat{\Delta}_N^0$ and $\overline{\Delta}_N^0$ are the numerators of $\widehat{\Delta}^0$ and $\overline{\Delta}^0$, and $\widehat{f}_q(\gamma_0)=\frac{1}{nh}\sum_{i=1}^{n}k\left(\frac{q_i-\gamma_0}{h}\right)$. From the earlier analysis in the proof of Theorem 2, $\widehat{\Delta}_N^0 - \overline{\Delta}_N^0$ satisfies

$$\begin{aligned} & \sqrt{nh}\left(\widehat{\Delta}_N^0 - \overline{\Delta}_N^0 - h^{p+1}\mathbf{e}_1\left[\left(M_o^-\right)^{-1}B^- - \left(M_o^+\right)^{-1}B^+\right]\mathbb{E}[g^{(p+1)}(x,\gamma_0)\Big| q=\gamma_0]f_q(\gamma_0)\right) \\ \approx\ & \frac{1}{\sqrt{nh}}\sum_{i=1}^n k\left(\frac{q_i-\gamma_0}{h}\right)\mathbf{e}_1'\left[\left(\overline{M}_{i0}^-\right)^{-1}r_{i0}^{e-} - \left(\overline{M}_{i0}^+\right)^{-1}r_{i0}^{e+}\right] \\ \approx\ & \frac{1}{\sqrt{nh}}\sum_{j=1}^n e_j\int k_h\left(\frac{q_i-\gamma_0}{h}\right)\mathbf{e}_1'\left[\left(\overline{M}_{i0}^-\right)^{-1}z_{j0}^i w_{j0}^{i-} - \left(\overline{M}_{i0}^+\right)^{-1}z_{j0}^i w_{j0}^{i+}\right]f(x_i,q_i)dx_i dq_i, \end{aligned}$$

and also

$$\begin{aligned} & \sqrt{nh}\left(\overline{\Delta}_N^0 - \mathbb{E}\left[\overline{\Delta}_N^0\right] - \Delta_0\left[\widehat{f}_q(\gamma_0) - \mathbb{E}\left[\widehat{f}_q(\gamma_0)\right]\right]\right) \\ =\ & \frac{1}{\sqrt{nh}}\sum_{i=1}^n\left\{k\left(\frac{q_i-\gamma_0}{h}\right)\left(a_-^0(x_i)-a_+^0(x_i)\right) - \mathbb{E}\left[k\left(\frac{q_i-\gamma_0}{h}\right)\left(a_-^0(x_i)-a_+^0(x_i)\right)\right]\right\} \\ & -\frac{\Delta_0}{\sqrt{nh}}\sum_{i=1}^n\left\{k\left(\frac{q_i-\gamma_0}{h}\right) - \mathbb{E}\left[k\left(\frac{q_i-\gamma_0}{h}\right)\right]\right\}. \end{aligned}$$

It is not hard to see that these two influence functions are uncorrelated, so the variance of $\sqrt{nh}\left(\widehat{\Delta}^0 - \Delta_0\right)$ is the sum of the variances of these two parts. The variance of the first part is derived in the proof of Theorem 2. As to the second part, note that

$$\begin{aligned} & \frac{1}{h}\mathbb{E}\left[k\left(\frac{q_i-\gamma_0}{h}\right)^2\left(a_-^0(x_i)-a_+^0(x_i)\right)^2\right] \\ =\ & \frac{1}{h}\int k\left(\frac{q_i-\gamma_0}{h}\right)^2\left(a_-^0(x_i)-a_+^0(x_i)\right)^2 f(x_i,q_i)dx_i dq_i \\ \approx\ & \int k(v_q)^2\left(a_-^0(x_i)-a_+^0(x_i)\right)^2 f(x_i,\gamma_0)dx_i dv_q \\ =\ & \int k(v_q)^2 dv_q\mathbb{E}[\left(a_-^0(x_i)-a_+^0(x_i)\right)^2|q_i=\gamma_0]f_q(\gamma_0). \end{aligned}$$

Similarly,

$$\begin{aligned} \frac{1}{h}\mathbb{E}\left[k\left(\frac{q_i-\gamma_0}{h}\right)^2\left(a_-^0(x_i)-a_+^0(x_i)\right)\right] &\approx \int k(v_q)^2 dv_q\Delta_0 f_q(\gamma_0), \\ \frac{1}{h}\mathbb{E}\left[k\left(\frac{q_i-\gamma_0}{h}\right)^2\right] &\approx \int k(v_q)^2 dv_q f_q(\gamma_0), \end{aligned}$$

so the variance of the second part is approximately

$$\begin{aligned} & \int k(v_q)^2 dv_q\mathbb{E}[\left(a_-^0(x_i)-a_+^0(x_i)\right)^2|q_i=\gamma_0]f_q(\gamma_0) + \Delta_0^2\int k(v_q)^2 dv_q f_q(\gamma_0) - 2\Delta_0\int k(v_q)^2 dv_q\Delta_0 f_q(\gamma_0) \\ =\ & \int k(v_q)^2 dv_q\left(\mathbb{E}[\left(a_-^0(x_i)-a_+^0(x_i)\right)^2|q_i=\gamma_0] - \Delta_0^2\right)f_q(\gamma_0). \end{aligned}$$

For the bias of the second part, note that

$$
\begin{aligned}
&\mathbb{E}\left[\overline{\Delta}_N^0\right] - \Delta_0 f_q(\gamma_0) \\
=& \int k_h(q_i-\gamma_0)\left(a_-^0(x_i) - a_+^0(x_i)\right) f(x_i,q_i)dx_i dq_i - \Delta_0 f_q(\gamma_0) \\
=& \int k(v_q)\left(a_-^0(x_i) - a_+^0(x_i)\right) f(x_i,\gamma_0+v_q h)dx_i dv_q - \Delta_0 f_q(\gamma_0) \\
\approx& \int k(v_q)\left(a_-^0(x_i) - a_+^0(x_i)\right) \sum\nolimits_{l=1}^{p+1} \frac{1}{l!} f_\gamma^{(l)}(x_i,\gamma_0)(v_q h)^l dx_i dv_q \\
=& \sum\nolimits_{l=1}^{p+1} \frac{h^l}{l!}\left[\int k(v_q) v_q{}^l dv_q\right] \int \left(a_-^0(x_i) - a_+^0(x_i)\right) f_\gamma^{(l)}(x_i,\gamma_0)dx_i
\end{aligned}
$$

where $f_\gamma^{(l)}(x_i,\gamma_0)$ is the lth order partial derivative of $f(x_i,\gamma)$ with respect to γ evaluated at $\gamma=\gamma_0$, and

$$
\begin{aligned}
\mathbb{E}\left[\widehat{f}_q(\gamma_0)\right] - f_q(\gamma_0) &= \int k_h(q_i-\gamma_0) f(q_i)dq_i - f_q(\gamma_0) \\
&= \int k(v_q) f(\gamma_0+v_q h)dv_q - f_q(\gamma_0) \\
&\approx \int k(v_q)\sum\nolimits_{l=1}^{p+1} \frac{1}{l!} f_\gamma^{(l)}(\gamma_0)(v_q h)^l dv_q \\
&= \sum\nolimits_{l=1}^{p+1} \frac{h^l}{l!}\left[\int k(v_q) v_q{}^l dv_q\right] f_\gamma^{(l)}(\gamma_0),
\end{aligned}
$$

where $f_\gamma^{(l)}(\gamma_0)$ is the lth order derivative of $f_q(\gamma)$ with respect to γ evaluated at $\gamma=\gamma_0$. In sum, the asymptotic distribution of $\sqrt{nh}\left(\widehat{\Delta}^0 - \Delta_0\right)$ is as stated in the theorem. ■

Proof of Theorem 3. First derive the formula for $\left(\overline{\delta}_\alpha, \widetilde{\delta}_x'\right)'$. From (9),

$$
\left(\overline{\delta}_\alpha, \widetilde{\delta}_x'\right)' = \left(\frac{1}{n}\sum_{i=1}^n (1,x_i')'(1,x_i') k_h(q_i-\widehat{\gamma})\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n (1,x_i')' k_h(q_i-\widehat{\gamma})\left(\widehat{a}_-(x_i) - \widehat{a}_+(x_i)\right)\right).
$$

By similar analysis to the proof of Theorem 2, $\widehat{\gamma}$ in $\left(\overline{\delta}_\alpha, \widetilde{\delta}_x'\right)'$ can be replaced by γ_0 without affecting its asymptotic distribution. Also, $\widehat{a}_-(x_i) - \widehat{a}_+(x_i)$ can be replaced by its linear approximation with no asymptotic impact. In summary,

$$
\begin{aligned}
&\sqrt{nh}\left(\left(\overline{\delta}_\alpha, \widetilde{\delta}_x'\right)' - \left(\delta_{\alpha 0}+\gamma_0\delta_{q0}, \delta_{x0}'\right)'\right) \\
\approx& \left(\frac{1}{n}\sum_{i=1}^n (1,x_i')'(1,x_i') k_h(q_i-\gamma_0)\right)^{-1} \cdot \left(\frac{1}{\sqrt{nh}}\sum_{i=1}^n (1,x_i')' k\left(\frac{q_i-\gamma_0}{h}\right)\right. \\
&\left.\mathbf{e}_1\left\{h^{p+1}\left[\left(\overline{M}_{i0}^-\right)^{-1}\overline{r}_{i0}^{m-} - \left(\overline{M}_{i0}^+\right)^{-1}\overline{r}_{i0}^{m+}\right] + \left[\left(\overline{M}_{i0}^-\right)^{-1} r_{i0}^{e-} - \left(\overline{M}_{i0}^+\right)^{-1} r_{i0}^{e+}\right]\right\}\right),
\end{aligned}
$$

where $\overline{M}_{i0}^\pm$, $\overline{r}_{i0}^{m\pm}$ and $r_{i0}^{e\pm}$ are defined in the proof of Theorem 2.

By standard methods, the denominator converges in probability to $M\cdot f_q(\gamma_0)$, where M is defined in the main text, so we concentrate on the numerator. First, consider the bias term. From the proof of Theorem 2,

$$\frac{1}{nh}\sum_{i=1}^{n}(1,x_i)'\,k\left(\frac{q_i-\gamma_0}{h}\right)\mathbf{e}_1\left[\left(\overline{M}_{i0}^{-}\right)^{-1}\overline{r}_{i0}^{m-}-\left(\overline{M}_{i0}^{+}\right)^{-1}\overline{r}_{i0}^{m+}\right]$$
$$\xrightarrow{p}\mathbb{E}\left[(1,x')'\,\mathbf{e}_1\left[\left(M_o^{-}\right)^{-1}B^{-}-\left(M_o^{+}\right)^{-1}B^{+}\right]g^{(p+1)}(x,\gamma_0)\Big|\,q=\gamma_0\right]f_q(\gamma_0).$$

Next consider the variance. We need to calculate the covariance between the lth and tth element of the numerator, $l,t=1,\cdots,d$. Taking the $(l+1)$th element of the numerator, $l=1,\cdots.d-1$, we consider

$$\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}x_{li}k\left(\frac{q_i-\gamma_0}{h}\right)\mathbf{e}_1'\left[\left(\overline{M}_{i0}^{+}\right)^{-1}r_{i0}^{e+}-\left(\overline{M}_{i0}^{-}\right)^{-1}r_{i0}^{e-}\right],$$

which is a second-order U-statistic. From Lemma 8.4 of Newey and McFadden (1994), this U-statistic is asymptotically equivalent to $\frac{1}{\sqrt{nh}}\sum\limits_{i=1}^{n}m_n^l(x_i,q_i,e_i)$, where

$$\begin{aligned} m_n^l(x_j,q_j,e_j) &= \mathbb{E}\left[x_{li}k\left(\frac{q_i-\gamma_0}{h}\right)\mathbf{e}_1'\left[\left(\overline{M}_{i0}^{+}\right)^{-1}z_{j0}^{i}w_{j0}^{i+}e_j-\left(\overline{M}_{i0}^{-}\right)^{-1}z_{j0}^{i}w_{j0}^{i-}e_j\right]\Bigg|\,x_j,q_j,e_j\right] \\ &= e_j\int x_{li}k\left(\frac{q_i-\gamma_0}{h}\right)\mathbf{e}_1'\left[\left(\overline{M}_{i0}^{+}\right)^{-1}z_{j0}^{i}w_{j0}^{i+}-\left(\overline{M}_{i0}^{-}\right)^{-1}z_{j0}^{i}w_{j0}^{i-}\right]f(x_i,q_i)dx_idq_i. \end{aligned}$$

It is not hard to show that

$$\frac{1}{nh}\sum_{i=1}^{n}m_n^l(x_i,q_i,e_i)m_n^t(x_i,q_i,e_i)\xrightarrow{d}\mathbb{E}\left[x_lx_t\int\left[k_+^2(v_q)\,\sigma_+^2(x)C_1^{+}(v_q)^2+k_-^2(v_q)\,\sigma_-^2(x)C_1^{-}(v_q)^2\right]dv_q\Bigg|\,q=\gamma_0\right]f_q(\gamma_0).$$

Then, applying the Liapunov central limit theorem, the asymptotic distribution of $\sqrt{nh}\left(\overline{\delta}_\alpha-\delta_{\alpha0}-\gamma_0\delta_{q0}\right)$ and $\sqrt{nh}\left(\widetilde{\delta}_{x_l}-\delta_{x_l0}\right)$, $l=1,\cdots,d-1$, follows as in the theorem.

When $\gamma_0=0$, $\sqrt{nh}\left(\widetilde{\delta}_\alpha-\delta_{\alpha0}\right)=\sqrt{nh}\left(\overline{\delta}_\alpha-\widehat{\gamma}\widehat{\delta}_q-\delta_{\alpha0}\right)=\sqrt{nh}\left(\overline{\delta}_\alpha-\delta_{\alpha0}\right)-\sqrt{nh}O_p(n^{-1})O_p(\left(\sqrt{nh}h\right)^{-1}+h^p)=\sqrt{nh}\left(\overline{\delta}_\alpha-\delta_{\alpha0}\right)+o_p(1)$, so $\sqrt{nh}\left(\widetilde{\delta}_\alpha-\delta_{\alpha0}\right)$ have the same asymptotic distribution as $\sqrt{nh}\left(\overline{\delta}_\alpha-\delta_{\alpha0}\right)$. When $\gamma_0\neq0$, the convergence rate of $\widetilde{\delta}_\alpha-\delta_{\alpha0}$ is $\sqrt{nh}h$. It is obvious that $\sqrt{nh}h\left(\overline{\delta}_\alpha-\widehat{\gamma}\widehat{\delta}_q-\delta_{\alpha0}\right)-\sqrt{nh}h\left(\overline{\delta}_\alpha-\gamma_0\widehat{\delta}_q-\delta_{\alpha0}\right)=\sqrt{nh}hO_p(n^{-1})O_p(\left(\sqrt{nh}h\right)^{-1}+h^p)=o_p(1)$. Also,

$$\begin{aligned} \sqrt{nh}h\left(\overline{\delta}_\alpha-\gamma_0\widehat{\delta}_q-\delta_{\alpha0}\right) &= \sqrt{nh}h\left(\overline{\delta}_\alpha-\delta_{\alpha0}-\gamma_0\delta_{q0}\right)-\gamma_0\sqrt{nh}h\left(\widehat{\delta}_q-\delta_{q0}\right) \\ &= o_p(1)-\gamma_0\sqrt{nh}h\left(\widehat{\delta}_q-\delta_{q0}\right). \end{aligned}$$

So $\sqrt{nh}h\left(\widetilde{\delta}_\alpha-\delta_{\alpha0}\right)$ has the same asymptotic distribution as $-\gamma_0\sqrt{nh}h\left(\widehat{\delta}_q-\delta_{q0}\right)$. ■

Proof of Theorem 4. Assume the densities of $(x',q)'$ and e are known. Since the minimax risk for a larger class of probability models must not be smaller than that for a smaller class of probability models, the lower bound for a particular distributional assumption also holds for a wider class of distributions. To simplify the calculation, assume e_i is iid $N(0,1)$ and $(x_i',q_i)'$ is iid uniform on $\mathcal{X}\times\mathcal{N}$, where $\mathcal{N}$ is specified as $[-\zeta,\zeta]$. Such a specification also appears in Fan (1993) where it is called the assumption of richness of joint densities. We will use the technique in Sun (2005) to develop our results. This technique is also implicitly used in Stone (1980) and the essential part of the technique can be cast in the language of Neyman-Pearson testing.

Let P, Q be probability measures defined on the same measurable space $(\Omega, \mathcal{A})$ with the affinity between the two measures defined as usual to be

$$\pi(P, Q) = \inf\left(\mathbb{E}_P[\phi] + \mathbb{E}_Q[1-\phi]\right),$$

where the infimum is taken over the measurable function ϕ such that $0 \leq \phi \leq 1$. In other words, $\pi(P, Q)$ is the smallest sum of type I and type II errors of any test between P and Q. It is a natural measure of the difficulty of distinguishing P and Q. Suppose μ is a measure dominating both P and Q with corresponding densities p and q. It follows from the Neyman-Pearson lemma that the infimum is achieved by setting $\phi = 1(p \leq q)$ and then

$$\begin{aligned}\pi(P, Q) &= \int 1(p \leq q) p d\mu + \int 1(p > q) q d\mu \\ &= 1 - \frac{1}{2}\int |p - q|\, d\mu \equiv 1 - \frac{1}{2}\|P - Q\|_1,\end{aligned}$$

where $\|\cdot\|_1$ is the L_1 distance between two probability measures. Now consider a pair of probability models $P, Q \in \mathcal{P}(s, B)$ such that $|\delta_\alpha(P) - \delta_\alpha(Q)| \geq \epsilon$. For any estimator $\widehat{\delta}$, we have

$$1\left(\left\|\widehat{\delta}_\alpha - \delta_\alpha(P)\right\| > \epsilon/2\right) + 1\left(\left\|\widehat{\delta}_\alpha - \delta_\alpha(Q)\right\| > \epsilon/2\right) \geq 1.$$

Let

$$\phi = \frac{1\left(\left|\widehat{\delta}_\alpha - \delta_\alpha(P)\right| > \epsilon/2\right)}{1\left(\left|\widehat{\delta}_\alpha - \delta_\alpha(P)\right| > \epsilon/2\right) + 1\left(\left|\widehat{\delta}_\alpha - \delta_\alpha(Q)\right| > \epsilon/2\right)}.$$

Then $0 \leq \phi \leq 1$ and

$$\begin{aligned}\sup_{\mathbb{P} \in \mathcal{P}(s,B)} \mathbb{P}\left(\left|\widehat{\delta}_\alpha - \delta_\alpha(\mathbb{P})\right| > \epsilon/2\right) &\geq \frac{1}{2}\left\{P\left(\left|\widehat{\delta}_\alpha - \delta_\alpha(P)\right| > \epsilon/2\right) + Q\left(\left|\widehat{\delta}_\alpha - \delta_\alpha(Q)\right| > \epsilon/2\right)\right\} \\ &\geq \frac{1}{2}\mathbb{E}_P[\phi] + \frac{1}{2}\mathbb{E}_Q[1-\phi].\end{aligned}$$

Therefore

$$\inf_{\widehat{\delta}_\alpha} \sup_{\mathbb{P} \in \mathcal{P}(s,B)} \mathbb{P}\left(\left|\widehat{\delta}_\alpha - \delta_\alpha(\mathbb{P})\right| > \epsilon/2\right) \geq \frac{1}{2}\pi(P, Q)$$

for any P and Q such that $|\delta_\alpha(P) - \delta_\alpha(Q)| \geq \epsilon$. So we need only search for the pair (P, Q) which minimize $\pi(P, Q)$ subject to the constraint $|\delta_\alpha(P) - \delta_\alpha(Q)| \geq \epsilon$. To obtain a lower bound with a sequence of independent observations, let $(\Omega, \mathcal{A})$ be the product space and $\mathcal{P}(s, B)$ be the family of product probabilities on such a space. Then for any pair of finite-product measures $P = \prod_{i=1}^n P_i$ and $Q = \prod_{i=1}^n Q_i$, the minimax risk satisfies

$$\inf_{\widehat{\delta}_\alpha} \sup_{\mathbb{P} \in \mathcal{P}(s,B)} \mathbb{P}\left(\left|\widehat{\delta}_\alpha - \delta_\alpha(\mathbb{P})\right| > \epsilon/2\right) \geq \frac{1}{2}\left(1 - \frac{1}{2}\left\|\prod_{i=1}^n P_i - \prod_{i=1}^n Q_i\right\|_1\right)$$

provided that $|\delta_\alpha(P) - \delta_\alpha(Q)| \geq \epsilon$. From Pollard (1993), if $dQ_i/dP_i = 1 + \Delta_i(\cdot)$, then

$$\left\|\prod_{i=1}^n P_i - \prod_{i=1}^n Q_i\right\|_1 \leq \exp\left(\sum_{i=1}^n \nu_i^2\right) - 1,$$

where $\nu_i^2 = \mathbb{E}_{P_i}[\Delta_i^2(\cdot)]$ is finite. So

$$\inf_{\widehat{\delta}_\alpha} \sup_{\mathbb{P}\in\mathcal{P}(s,B)} \mathbb{P}\left(\left|\widehat{\delta}_\alpha - \delta_\alpha(\mathbb{P})\right| > \epsilon/2\right) \geq \frac{1}{2}\left(\frac{3}{2} - \exp\left(\sum_{i=1}^n \nu_i^2\right)\right) \tag{20}$$

provided that $|\delta_\alpha(P) - \delta_\alpha(Q)| \geq \epsilon$.

It remains to find probabilities P and Q that are difficult to distinguish by the data set $\{(x_i', q_i, y_i)\}_{i=1}^n$. First assume $\gamma_0 \neq 0$. Without loss of generality, let $\gamma_0 > 0$. Under P, the data is generated according to

$$y_i = g_P(x_i, q_i) + (\delta_{\alpha P} + x_i'\delta_{xP} + q_i\delta_{qP})\, 1(q_i \leq \gamma_0) + e_i,$$

and under Q, g_P and δ_P are changed to g_Q and δ_Q, respectively. We now specify g and δ for each model. For P, let $g_P = 0$ and $\delta_P = 0$; for Q, let

$$g_Q(x, q) = -\xi\eta^s\varphi_q\left(\frac{q-\gamma_0}{\eta}\right),\ \delta_{\alpha Q} = -\xi\gamma_0\eta^{s-1},\ \delta_{xQ} = 0, \text{ and } \delta_{qQ} = \xi\eta^{s-1},$$

where ξ is a positive constant, $\eta = n^{-1/(2s+1)}$, φ_q is an infinitely differentiable function in q satisfying (i) $\varphi_q(v) = 0$ for $v \geq 0$, (ii) $\varphi_q(v) = v$, for $v \leq -\zeta$, and (iii) $v - \varphi_q(v) \in (0, 1)$ for $v \in (-\zeta, 0)$. It is not hard to check that $g_Q(x, q) \in C(s, B)$ for some $B > 0$, so it remains to compute the L_1 distance between the two measures. Let the density of Q_i with respect to P_i be $1 + \Delta_i(\cdot)$, then

$$\Delta_i(x_i, q_i, y_i) = \begin{cases} \phi(y_i - g_Q(x_i, q_i) - \delta_{\alpha Q} - q_i\delta_{qQ})/\phi(y_i) - 1, & \text{if } q_i \in [\gamma_0 - \zeta\eta, \gamma_0], \\ 0, & \text{otherwise} \end{cases}$$

where $\phi(\cdot)$ is the standard normal pdf. Therefore,

$$\begin{aligned}
\mathbb{E}_{P_i}[\Delta_i^2] &= \int_{\gamma_0-\zeta\eta}^{\gamma_0}\int_0^1\cdots\int_0^1\int_{-\infty}^{\infty}\left[\phi(y - g_Q(x,q) - \delta_{\alpha Q} - q\delta_{qQ})/\phi(y) - 1\right]^2 \phi(y) f(x,q)\,dydxdq \\
&= \frac{1}{2\zeta}\int_{\gamma_0-\zeta\eta}^{\gamma_0}\int_0^1\cdots\int_0^1\int_{-\infty}^{\infty}\phi(y - g_Q(x,q) - \delta_{\alpha Q} - q\delta_{qQ})^2/\phi(y)\,dydxdq \\
&\quad -\frac{1}{\zeta}\int_{\gamma_0-\zeta\eta}^{\gamma_0}\int_0^1\cdots\int_0^1\int_{-\infty}^{\infty}\phi(y - g_Q(x,q) - \delta_{\alpha Q} - q\delta_{qQ})\,dydxdq + \frac{\eta}{2} \\
&= \frac{1}{2\zeta}\int_{\gamma_0-\zeta\eta}^{\gamma_0}\int_0^1\cdots\int_0^1\int_{-\infty}^{\infty}\phi(y - g_Q(x,q) - \delta_{\alpha Q} - q\delta_{qQ})^2/\phi(y)\,dydxdq - \frac{\eta}{2}.
\end{aligned}$$

Plugging in the standard normal pdf yields

$$\begin{aligned}
\mathbb{E}_{P_i}[\Delta_i^2] &= \frac{1}{2\zeta}\int_{\gamma_0-\zeta\eta}^{\gamma_0}\int_0^1\cdots\int_0^1\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{2(y - g_Q(x,q) - \delta_{\alpha Q} - q\delta_{qQ})^2}{2} + \frac{y^2}{2}\right\} dydxdq - \frac{\eta}{2} \\
&= \frac{1}{2\zeta}\int_{\gamma_0-\zeta\eta}^{\gamma_0}\int_0^1\cdots\int_0^1 \exp\left\{\left[g_Q(x,q) + \delta_{\alpha Q} + q\delta_{qQ}\right]^2\right\} dxdq - \frac{\eta}{2} \\
&= \frac{1}{2\zeta}\int_{\gamma_0-\zeta\eta}^{\gamma_0}\exp\left\{\xi^2\eta^{2s}\left[\frac{q-\gamma_0}{\eta} - \varphi_q\left(\frac{q-\gamma_0}{\eta}\right)\right]^2\right\} dq - \frac{\eta}{2} \\
&\leq \frac{\eta}{2}\exp\left(\xi^2\eta^{2s}\right) - \frac{\eta}{2} = \frac{\eta}{2}\left(\exp\left(\xi^2\eta^{2s}\right) - 1\right) = \frac{\xi^2}{2}\eta^{2s+1}(1 + o(1)) \leq \frac{\xi^2}{2n},
\end{aligned}$$

when n is large enough.

When ξ is small enough, say $\xi^2/2 \le \log(5/4)$, we have

$$\exp\left(\sum_{i=1}^n \nu_i^2\right) \le \exp\left(\frac{\xi^2}{2}\right) < \frac{5}{4}.$$

It follows from (20) that

$$\inf_{\widehat{\delta}_\alpha} \sup_{\mathbb{P}\in\mathcal{P}(s,B)} \mathbb{P}\left(\left|\widehat{\delta}_\alpha - \delta_\alpha(\mathbb{P})\right| > \frac{\epsilon}{2} n^{-\frac{s-1}{2s+1}}\right) \ge \frac{1}{2}\left(\frac{3}{2} - \frac{5}{4}\right) = \frac{1}{8} \ge C,$$

on choosing $C \le 1/8$, where $\frac{\epsilon}{2} n^{-\frac{s-1}{2s+1}}$ appears because $|\delta_\alpha(P) - \delta_\alpha(Q)| = \gamma_0 \xi n^{-\frac{s-1}{2s+1}} \ge \epsilon n^{-\frac{s-1}{2s+1}}$ for a small ϵ.

When $\gamma_0 = 0$, we choose

$$g_Q(x,q) = -\xi\eta^s \varphi_q\left(\frac{q}{\eta}\right),\ \delta_{\alpha Q} = \xi\eta^s,\ \delta_{xQ} = 0, \text{ and } \delta_{qQ} = 0,$$

where φ_q is an infinitely differentiable function in q satisfying (i) $\varphi_q(v) = 0$ for $v \ge 0$, (ii) $\varphi_q(v) = 1$, for $v \le -\zeta$, and (iii) $\varphi_q(v) \in (0,1)$ for $v \in (-\zeta, 0)$, then

$$\mathbb{E}_{P_i}[\Delta_i^2] = \frac{1}{2\zeta}\int_{-\zeta\eta}^{0} \exp\left\{\xi^2\eta^{2s}\left[1 - \varphi_q\left(\frac{q}{\eta}\right)\right]^2\right\} dq - \frac{\eta}{2} \le \frac{\eta}{2}\exp\left(\xi^2\eta^{2s}\right) - \frac{\eta}{2},$$

and following similar steps to those above we have $\inf_{\widehat{\delta}_\alpha} \sup_{\mathbb{P}\in\mathcal{P}(s,B)} \mathbb{P}\left(\left|\widehat{\delta}_\alpha - \delta_\alpha(\mathbb{P})\right| > \frac{\epsilon}{2} n^{-\frac{s}{2s+1}}\right) \ge C$ for some ϵ and C.

The above argument also shows that the optimal rate of convergence for δ_q is $n^{-\frac{s-1}{2s+1}}$. As for δ_x, we need only choose another pair of probabilities P and Q. To simplify notation, let $d-1=1$ so that x is only one-dimensional. Let P be the same as above, and

$$g_Q(x,q) = -\xi\eta^s \varphi_q\left(\frac{q-\gamma_0}{\eta}\right) x,\ \delta_{\alpha Q} = 0,\ \delta_{xQ} = \xi\eta^s, \text{ and } \delta_{qQ} = 0,$$

where φ_q is an infinitely differentiable function in q satisfying (i) $\varphi_q(v) = 0$ for $v \ge 0$, (ii) $\varphi_q(v) = 1$, for $v \le -\zeta$, and (iii) $\varphi_q(v) \in (0,1)$ for $v \in (-\zeta, 0)$. Then

$$\mathbb{E}_{P_i}[\Delta_i^2] = \frac{1}{2\zeta}\int_{\gamma_0-\zeta\eta}^{\gamma_0}\int_0^1 \exp\left\{\xi^2\eta^{2s}x^2\left[1 - \varphi_q\left(\frac{q}{\eta}\right)\right]^2\right\} dxdq - \frac{\eta}{2} \le \frac{\eta}{2}\exp\left(\xi^2\eta^{2s}\right) - \frac{\eta}{2},$$

and it follows that $\inf_{\widehat{\delta}_x} \sup_{\mathbb{P}\in\mathcal{P}(s,B)} \mathbb{P}\left(\left|\widehat{\delta}_x - \delta_x(\mathbb{P})\right| > \frac{\epsilon}{2} n^{-\frac{s}{2s+1}}\right) \ge C$ for some ϵ and C. ■

Proof of Theorem 5. Note that

$$\sqrt{n}\begin{pmatrix} \widehat{\beta}_{GMM} - \beta_0 \\ \widehat{\delta}_{GMM} - \delta_0 \end{pmatrix} = \left(\widehat{G}'\widehat{\Omega}^{-1}\widehat{G}\right)^{-1}\widehat{G}'\widehat{\Omega}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \begin{pmatrix} z_i 1\left(q_i \le \widehat{\gamma}\right) \\ z_i' 1\left(q_i > \widehat{\gamma}\right) \end{pmatrix}\left(\varepsilon_i + \mathbf{x}_i'\delta_0 1\left(\widehat{\gamma} < q_i \le \gamma_0\right)\right).$$

By the consistency of $\widehat{\gamma}$ and Glivenko-Cantelli, $\widehat{G} \xrightarrow{p} G$. Following the proof of Theorem 3 of Caner and Hansen (2004), we can show that $\widehat{\Omega} \xrightarrow{p} \Omega$ under the moment restrictions on x, q, ε and z. We still need to show that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix} z_i 1(q_i \le \widehat{\gamma}) \\ z_i' 1(q_i > \widehat{\gamma}) \end{pmatrix} \mathbf{x}_i'\delta_0 1(\widehat{\gamma} < q_i \le \gamma_0)$$

$$= \begin{pmatrix} \mathbf{0} \\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i'\mathbf{x}_i'\delta_0 1(\widehat{\gamma} < q_i \le \gamma_0) \end{pmatrix} \xrightarrow{p} 0,$$

and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix} z_i 1(\gamma_0 < q_i \le \widehat{\gamma}) \\ z_i' 1(\widehat{\gamma} < q_i \le \gamma_0) \end{pmatrix} \varepsilon_i \xrightarrow{p} 0.$$

For these two results, consistency of $\widehat{\gamma}$ is not enough; we need $n^{1/2}(\widehat{\gamma} - \gamma_0) \xrightarrow{p} 0$. But in this case, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i'\mathbf{x}_i'\delta_0 1(\widehat{\gamma} < q_i \le \gamma_0) = o_p\left(\frac{1}{n}\sum_{i=1}^{n} z_i'\mathbf{x}_i'\delta_0\right) = o_p(1)$, and the second result holds similarly. Given these two results, standard arguments yield the asymptotic distribution of the GMM estimator. ■

Proof of Theorem 6. Because

$$\begin{aligned} \widehat{e}_i &= y_i - \mathbf{x}_i'\widehat{\beta} - \mathbf{x}_i'\widehat{\delta}1(q_i \le \widehat{\gamma}) \\ &= u_i + \left[m_i - \mathbf{x}_i'\widehat{\beta} - \mathbf{x}_i'\widehat{\delta}1(q_i \le \widehat{\gamma})\right] \\ &\equiv u_i + D_i, \end{aligned}$$

we decompose $I_n^{(1)}$ as

$$\begin{aligned} I_n^{(1)} &= \frac{nh^{d/2}}{n(n-1)}\sum_i\sum_{j\neq i}\left[D_iD_j + u_iu_j + 2u_iD_j\right]K_{h,ij} \\ &\equiv I_{1n}^{(1)} + I_{2n}^{(1)} + I_{3n}^{(1)}. \end{aligned}$$

We complete the proof by examining $I_{1n}^{(1)}, I_{2n}^{(1)}, I_{3n}^{(1)}$, and showing that $v_n^{(1)2} = \Sigma^{(1)} + o_p(1)$ under $H_0^{(1)}$ and the local alternative and $v_n^{(1)2} = O_p(1)$ under $H_1^{(1)}$. Throughout this proof, $z_i = (x_i', q_i, u_i)'$ and $\mathbb{E}_i[\cdot] = \mathbb{E}[\cdot|x_i, q_i]$.

It is shown in Proposition 6 that $I_{1n}^{(1)} = O_p(h^{d/2})$ under $H_0^{(1)}$ and converges to δ under the local alternative. It can also be shown that $I_{3n}^{(1)} = O_p(h^{d/2})$ under $H_0^{(1)}$ and is dominated by I_{1n} under the alternative, see, e.g., Zheng (1996). Proposition 8 shows that $I_{2n}^{(1)} \xrightarrow{d} N\left(0, \Sigma^{(1)}\right)$, and Proposition 9 shows the results related to $v_n^{(1)2}$. So the proof is complete. ■

Proof of Theorem 7. First, decompose $I_n^{(2)}$ by using (15):

$$\begin{aligned} I_n^{(2)} &= \frac{nh^{d/2}}{n(n-1)}\sum_i\sum_{j\neq i} 1_i^{\Gamma}1_j^{\Gamma}\left\{(m_i - \widehat{m}_i)(m_j - \widehat{m}_j) + u_iu_j + \widehat{u}_i\widehat{u}_j\right. \\ &\quad \left. + 2u_i(m_j - \widehat{m}_j) - 2\widehat{u}_i(m_j - \widehat{m}_j) - 2u_i\widehat{u}_j\right\}K_{h,ij} \\ &\equiv I_{1n}^{(2)} + I_{2n}^{(2)} + I_{3n}^{(2)} + 2I_{4n}^{(2)} - 2I_{5n}^{(2)} - 2I_{6n}^{(2)}. \end{aligned}$$

We complete the proof by examining $I_{1n}^{(2)}, \cdots, I_{6n}^{(2)}$, and showing that $v_n^{(2)2} = \Sigma^{(2)} + o_p(1)$ under both $H_0^{(2)}$ and $H_1^{(2)}$. Throughout this proof, $z_i = (x_i', q_i, u_i)'$ and $\mathbb{E}_i[\cdot] = \mathbb{E}[\cdot|x_i, q_i]$. We show that $I_{2n}^{(2)}$ contributes to the asymptotic distribution under the null, and $I_{1n}^{(2)}$ contributes to the power under the local alternative. All other terms will not contribute to the asymptotic distribution under either the null or the alternative; that proof just extends Propositions 3, 4, 5 and 6 in Appendix B of Porter and Yu (2011), so is omitted here. The remaining part of the proof concentrates on $I_{1n}^{(2)}$ and $I_{2n}^{(2)}$, and we only briefly mention the results for the other terms since these are obtained in a similar fashion.

First, $I_{2n}^{(2)}$, $I_{3n}^{(2)}$ and $I_{6n}^{(2)}$ are invariant under $H_0^{(2)}$ and $H_1^{(2)}$. It can be shown that $I_{3n}^{(2)}$ and $I_{6n}^{(2)}$ are both $o_p(1)$. Proposition 8 shows that $I_{2n}^{(2)} \xrightarrow{d} N\left(0, \Sigma^{(2)}\right)$.

Under $H_0^{(2)}$, Proposition 7 shows that $I_{1n}^{(2)} = o_{P_m}(1)$, and it can also be shown that I_{4n} and I_{5n} are both $o_{P_m}(1)$ uniformly in $m(\cdot) \in H_0$.

Under $H_1^{(2)}$, it can be shown that $I_{4n}^{(2)}$ and $I_{5n}^{(2)}$ are dominated by $I_{1n}^{(2)}$, and Proposition 7 shows that $I_{1n}^{(2)} = O_p\left(nh^{d/2}b\right)$ under $H_1^{(2)}$. The local power can be easily obtained from the proof of Proposition 7.

Finally, Proposition 10 shows that $v_n^{(2)2} = \Sigma^{(2)} + o_p(1)$ under both $H_0^{(2)}$ and $H_1^{(2)}$. So the proof is complete. ■

Proof of Theorem 8. This proof is similar but more tedious than the proofs of Theorem 6 and 7. Note that $\Phi(z)$ is a continuous function. By Pólya's theorem, it suffices to show that for any fixed value of $z \in \mathbb{R}$, $\left|P\left(T_n^{(\ell)*} \le z|\mathcal{F}_n\right) - \Phi(z)\right| = o_p(1)$.

For the first test, let

$$D_i^* = \mathbf{x}_i'\widehat{\beta} + \mathbf{x}_i'\widehat{\delta}1\left(q_i \le \widehat{\gamma}\right) - \mathbf{x}_i'\widehat{\beta}^* - \mathbf{x}_i'\widehat{\delta}^* 1\left(q_i \le \widehat{\gamma}^*\right),$$

where $\left(\widehat{\beta}^*, \widehat{\delta}^*, \widehat{\gamma}^*\right)$ is the least squares estimator using the data $\{y_i^*, x_i, q_i\}_{i=1}^n$. Then

$$\begin{aligned} I_n^{(1)*} &= \frac{nh^{d/2}}{n(n-1)} \sum_i \sum_{j \ne i} \left[D_i^* D_j^* + u_i^* u_j^* + 2u_i^* D_j^*\right] K_{h,ij} \\ &\equiv I_{1n}^{(1)*} + I_{2n}^{(1)*} + I_{3n}^{(1)*}. \end{aligned}$$

The theorem is proved if we can show that $I_{in}^{(1)*}|\mathcal{F}_n = o_p(1)$ for $i = 1$ and 3 and $I_{2n}^{(1)*}/v_n^{(1)*}|\mathcal{F}_n \to N(0,1)$ in probability. The first part can be proved as in the proof of Theorem 6, and for the second part, see the discussion below.

For the second test, denote $m_i^* = \widehat{y}_i$ and define $\widehat{m}_i^*$ and $\widehat{u}_i^*$ by

$$\widehat{m}_i^* = \frac{1}{n-1} \sum\nolimits_{j \ne i} m_j^* L_{b,ij} \Big/ \widehat{f}_i,$$

and

$$\widehat{u}_i^* = \frac{1}{n-1} \sum\nolimits_{j \ne i} u_j^* L_{b,ij} \Big/ \widehat{f}_i.$$

Then using $\widehat{e}_i^* = y_i^* - \widehat{y}_i^* = m_i^* + u_i^* - (\widehat{m}_i^* + \widehat{u}_i^*)$, we get

$$\begin{aligned} I_n^{(2)*} &= \frac{nh^{d/2}}{n(n-1)} \sum_i \sum_{j \ne i} 1_i^\Gamma 1_j^\Gamma \left\{(m_i^* - \widehat{m}_i^*)\left(m_j^* - \widehat{m}_j^*\right) + u_i^* u_j^* + \widehat{u}_i^* \widehat{u}_j^*\right. \\ &\quad \left. +2u_i^*\left(m_j^* - \widehat{m}_j^*\right) - 2\widehat{u}_i^*\left(m_j^* - \widehat{m}_j^*\right) - 2u_i^*\widehat{u}_j^*\right\} K_{h,ij} \\ &\equiv I_{1n}^{(2)*} + I_{2n}^{(2)*} + I_{3n}^{(2)*} + 2I_{4n}^{(2)*} - 2I_{5n}^{(2)*} - 2I_{6n}^{(2)*}. \end{aligned}$$

The theorem is proved if we can show that $I_{in}^{(2)*}|\mathcal{F}_n = o_p(1)$ for $i = 1, 3, 4, 5, 6$ and $I_{2n}^{(2)*}/v_n^{(2)*}|\mathcal{F}_n \to N(0,1)$ in probability. The first part is similar to that of Theorem 7 under $H_0^{(2)}$. However, note that $m^*(\cdot)|\mathcal{F}_n$ as defined above satisfies $H_0^{(2)}$ even if $m(\cdot)$ is from $H_1^{(2)}$; see Gu et al. (2007) for a similar analysis in testing omitted variables. But there is some differences in showing the second part.

First, because $u_i^*|\mathcal{F}_n$ are mean zero and mutually independent and have variance $\widehat{e}_i^2$,

$$\frac{nh^{d/2}}{n(n-1)}\sum_i\sum_{j\neq i}1_i^\Gamma 1_j^\Gamma u_i^* u_j^* K_{h,ij} = \frac{2nh^{1/2}}{n(n-1)}\sum_i\sum_{j>i}1_i^\Gamma 1_j^\Gamma u_i^* u_j^* K_{h,ij} \equiv \sum_i\sum_{j>i}U_{n,ij}^*$$

is a second order degenerate U-statistic with conditional variance

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

Because $U_{n,ij}^*$ depends on i and j, we use the central limit theorem of de Jong (1987) for generalized quadratic forms rather than Hall (1984) to find the asymptotic distribution of $I_{2n}^{(2)*}$. From his Proposition 3.2, we know $I_{2n}^{(2)*}/v_n^{(2)}|\mathcal{F}_n \to N(0,1)$ in probability as long as

$$\begin{aligned}
G_I^* &= \sum_i\sum_{j>i}\mathbb{E}^*\left[U_{n,ij}^{*4}\right] = o_p\left(v_n^{(2)4}\right),\\
G_{II}^* &= \sum_i\sum_{j>i}\sum_{l>j>i}\mathbb{E}^*\left[U_{n,ij}^{*2}U_{n,il}^{*2} + U_{n,ji}^{*2}U_{n,jl}^{*2} + U_{n,li}^{*2}U_{n,lj}^{*2}\right] = o_p\left(v_n^{(2)4}\right),\\
G_{IV}^* &= \sum_i\sum_{j>i}\sum_{k>j>i}\sum_{l>k>j>i}\mathbb{E}^*\left[U_{n,ij}^*U_{n,ik}^*U_{n,lj}^*U_{n,lk}^* + U_{n,ij}^*U_{n,il}^*U_{n,kj}^*U_{n,kl}^* + U_{n,ik}^*U_{n,il}^*U_{n,jk}^*U_{n,jl}^*\right] = o_p\left(v_n^{(2)4}\right).
\end{aligned}$$

It is straightforward to show that

$$G_I^* = O_p\left(\left(n^2h^d\right)^{-1}\right), G_{II}^* = O_p(n^{-1}), G_{IV}^* = O_p(h^d),$$

see, e.g., the proof of Theorem 2 of Hsiao et al. (2007), so the result follows by $v_n^{(2)4} = O_p(1)$. Next, it is easy to check that $\mathbb{E}^*\left[v_n^{(2)*2}\right] = v_n^{(2)2} + o_p(1)$, and $Var^*\left(v_n^{(2)*2}\right) = o_p(1)$. Thus $I_{2n}^{(2)*}/v_n^{(2)*}|\mathcal{F}_n \to N(0,1)$ in probability. The analysis for $I_{2n}^{(1)*}$ is similar. ■

Appendix B: Propositions

Proposition 1 $\widehat{\gamma} - \gamma_0 = O_p(h)$.

Proof. We apply Lemma 4 of Porter and Yu (2011) to prove this result. Define $Q_n(\gamma)$ as the probability limit of $\widehat{Q}_n(\gamma)$. Lemma 1 shows that

$$\sup_{\gamma\in\Gamma}\left|\widehat{Q}_n(\gamma) - Q_n(\gamma)\right| \xrightarrow{p} 0,$$

where

$$Q_n(\gamma) = \int\left[\begin{array}{c}\int_{-1}^0\int K^x(u_x,x)k_-(u_q)m(x,\gamma+u_qh)f(x+u_xh,\gamma+u_qh)du_xdu_q\\ -\int_0^1\int K^x(u_x,x)k_+(u_q)m(x,\gamma+u_qh)f(x+u_xh,\gamma+u_qh)du_xdu_q\end{array}\right]^2 f(x)dx.$$

Let $\mathcal{N}_n = [\gamma_0 - h, \gamma_0 + h]$ and $\gamma_n = \arg\max_{\gamma\in\Gamma} Q_n(\gamma)$, then it remains to show that $\sup_{\gamma\in\Gamma\backslash\mathcal{N}_n} Q_n(\gamma) < Q_n(\gamma_n) - C$ for some positive constant C. It is easy to show that $\sup_{\gamma\in\Gamma\backslash\mathcal{N}_n} Q_n(\gamma) = O(h^2)$. On the contrary, for $\gamma \in \mathcal{N}_n$,

$Q_n(\gamma)$ behaves quite differently. Specifically, let $\gamma = \gamma_0 + ah$, $a \in (0,1)$, then

$$Q_n(\gamma) = \int \left[\begin{array}{l} \int_{-1}^{0} \int K^x(u_x, x) k_-(u_q) g(x, \gamma + u_q h) f(x + u_x h, \gamma + u_q h) du_x du_q \\ + \int_{-1}^{-a} \int K^x(u_x, x) k_-(u_q) \left(1 \ x \ \gamma + u_q h\right)' \delta_0 f(x + u_x h, \gamma + u_q h) du_x du_q \\ - \int_0^1 \int K^x(u_x, x) k_+(u_q) g(x, \gamma + u_q h) f(x + u_x h, \gamma + u_q h) du_x du_q \end{array} \right]^2 f(x) dx.$$

The difference of the first and the third terms in brackets is $O(h^2)$, so the second term will dominate. From Assumption I, $(1 \ x \ \gamma_0)' \delta_0 \neq 0$ for some $x \in \mathcal{X}$, so $\int \left[\int K^x(u_x, x) \left(1 \ x \ \gamma_0\right)' \delta_0 f(x, \gamma_0) du_x \right]^2 f(x) dx > C$ for some positive constant C. Because $k_-(0) > 0$ and $k_-(\cdot) \geq 0$, $\int_{-1}^{-a} k_-(u_q) du_q < 1$ and is a decreasing function of a. As a result, $Q_n(\gamma)$ is a decreasing function of a for $a \in (0,1)$ up to $O(h^2)$. Similarly, it is an increasing function of a for $a \in (-1, 0)$. So $Q_n(\gamma)$ is maximized at some $\gamma_n \in \mathcal{N}_n$ such that $Q_n(\gamma_n) > \sup_{\gamma \in \Gamma \backslash \mathcal{N}_n} |Q_n(\gamma)| + C/2$ for n large enough. The required result follows. ■

Proposition 2 $\widehat{\gamma} - \gamma_0 = O_p(n^{-1})$.

Proof. We use the standard shelling method (see, e.g., Theorem 3.2.5 of Van der Vaart and Wellner (1996)) to prove this result.

For each n, the parameter space can be partitioned into the "shells" $S_{l,n} = \left\{ \pi : 2^{l-1} < n \left| \gamma - \gamma_0 \right| \leq 2^l \right\}$ with l ranging over the integers. If $n \left| \widehat{\gamma} - \gamma_0 \right|$ is larger than 2^L for a given integer L, then $\widehat{\gamma}$ is in one of the shells $S_{l,n}$ with $l \geq L$. In that case the supremum of the map $\gamma \mapsto \widehat{Q}_n(\gamma) - \widehat{Q}_n(\gamma_0)$ over this shell is nonnegative by the property of $\widehat{\gamma}$. Note that

$$\begin{aligned} & P\left(n \left|\widehat{\gamma} - \gamma_0\right| > 2^L\right) \\ \leq \ & P\left(\sup_{2^L < n|\gamma - \gamma_0| \leq nh} \left(\frac{1}{n} \sum_{i=1}^n \widehat{\Delta}_i^2(\gamma) - \frac{1}{n} \sum_{i=1}^n \widehat{\Delta}_i^2(\gamma_0) \right) \geq 0 \right) + P\left(\left|\widehat{\gamma} - \gamma_0\right| \geq h\right) \\ \leq \ & \sum_{l=L}^{\log_2(nh)} P\left(\sup_{S_{l,n}} \frac{1}{n} \sum_{i=1}^n \widehat{\Delta}_i^2(\gamma) \geq \frac{1}{n} \sum_{i=1}^n \widehat{\Delta}_i^2(\gamma_0) \right) + P\left(\left|\widehat{\pi} - \pi_0\right| \geq h\right) \\ \leq \ & \sum_{l=L}^{\log_2(nh)} P\left(\sup_{S_{l,n}} \frac{1}{n} \sum_{i=1}^n \widehat{\Delta}_i^2(\gamma) 1(\Delta(x_i) > 0) \geq \frac{1}{n} \sum_{i=1}^n \widehat{\Delta}_i^2(\gamma_0) 1(\Delta(x_i) > 0) \right) \\ & + \sum_{l=L}^{\log_2(nh)} P\left(\sup_{S_{l,n}} \frac{1}{n} \sum_{i=1}^n \widehat{\Delta}_i^2(\gamma) 1(\Delta(x_i) < 0) \geq \frac{1}{n} \sum_{i=1}^n \widehat{\Delta}_i^2(\gamma_0) 1(\Delta(x_i) < 0) \right) \\ & + P\left(\left|\widehat{\pi} - \pi_0\right| \geq h\right) \\ \equiv \ & T1 + T2 + T3, \end{aligned}$$

where $\Delta(x_i) \equiv (1, x_i', \gamma_0) \delta_0$. $T3$ converges to zero by the last proposition, so we concentrate on the first two terms. $T2$ can be analyzed similar to $T1$, so we only consider $T1$ in the following discussion.

$$\begin{aligned} T1 \ \leq \ & \sum_{l=L}^{\log_2(nh)} P\left(\sup_{S_{l,n}} \frac{1}{n} \sum_{i=1}^n \left(\widehat{\Delta}_i(\gamma) - \widehat{\Delta}_i(\gamma_0) \right) 1(\Delta(x_i) > 0) > 0 \right) \\ & + \sum_{l=L}^{\log_2(nh)} P\left(\sup_{S_{l,n}} \frac{1}{n} \sum_{i=1}^n \left(\widehat{\Delta}_i(\gamma) + \widehat{\Delta}_i(\gamma_0) \right) 1(\Delta(x_i) > 0) < 0 \right). \end{aligned}$$

We concentrate on the first term since the second term is easier to analyze given that $\Delta(x_i) > 0$. To simplify notations, we neglect $1(\Delta(x_i) > 0)$ in the following discussion.

Note that

$$
\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\Delta}_i(\gamma)-\widehat{\Delta}_i(\gamma_0)\right) \\
= & \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\left(y_jK_{h,ij}^{\gamma-}-y_jK_{h,ij}^{\gamma+}\right)-\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\left(y_jK_{h,ij}^{\gamma_0^-}-y_jK_{h,ij}^{\gamma_0^+}\right) \\
= & \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\left[\left(m_jK_{h,ij}^{\gamma-}-m_jK_{h,ij}^{\gamma+}\right)-\left(m_jK_{h,ij}^{\gamma_0^-}-m_jK_{h,ij}^{\gamma_0^+}\right)\right] \\
& +\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\left(e_jK_{h,ij}^{\gamma-}-e_jK_{h,ij}^{\gamma+}\right)-\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\left(e_jK_{h,ij}^{\gamma_0^-}-e_jK_{h,ij}^{\gamma_0^+}\right) \\
\equiv & D1+D2,
\end{aligned}
$$

where $m_j = g_j + (1\ x_j'\ q_j)\delta_0 1(q_j \le \gamma_0)$ with $g_j = g(x_j, q_j)$. Suppose $\gamma_0 < \gamma < \gamma_0 + h$. Then

$$
\begin{aligned}
D1 = & \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}g_j\left(K_{h,ij}^{\gamma_0^+}-K_{h,ij}^{\gamma+}\right)+\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}g_j\left(K_{h,ij}^{\gamma-}-K_{h,ij}^{\gamma_0^-}\right) \\
& +\frac{1}{n\left(n-1\right)}\sum_{i=1}^{n}\sum_{j=1}^{n}(1\ x_j'\ q_j)\delta_0\left(K_{h,ij}^{\gamma-}-K_{h,ij}^{\gamma_0^-}\right)1(q_j\le\gamma_0) \\
\le & -C\frac{|\gamma-\gamma_0|}{h},
\end{aligned}
$$

for some $C > 0$ with probability approaching 1 by calculating the mean and variance of $D1$ in its U-projection, where the first two terms contribute only $O_p(|\gamma-\gamma_0|)$, and the third term contributes to $-C\frac{|\gamma-\gamma_0|}{h}$ because for each i, $K_{h,ij}^{\gamma-}$ covers j terms less than $K_{h,ij}^{\gamma_0^-}$ given that $\gamma > \gamma_0$ and $k_\pm(0) > 0$. In consequence, for $\eta \le h$,

$$
P\left(\sup_{|\gamma-\gamma_0|<\eta}\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\Delta}_i(\gamma)-\widehat{\Delta}_i(\gamma_0)\right)>0\right)\le P\left(\sup_{|\gamma-\gamma_0|<\eta}D2>C\frac{|\gamma_0-\gamma|}{h}\right).
$$

Notice that

$$
\begin{aligned}
D2 = & \frac{1}{n\left(n-1\right)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}e_j\left(K_{h,ij}^{\gamma-}-K_{h,ij}^{\gamma_0^-}\right)1(q_j\le\gamma_0)-\frac{1}{n\left(n-1\right)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}e_j\left(K_{h,ij}^{\gamma_0^+}-K_{h,ij}^{\gamma+}\right)1(q_j>\gamma) \\
& +\frac{1}{n\left(n-1\right)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}e_j\left(K_{h,ij}^{\gamma-}+K_{h,ij}^{\gamma_0^+}\right)1\left(\gamma_0<q_j\le\gamma\right).
\end{aligned}
$$

Conditional on x_i, the three summations $\sum_{j=1,j\neq i}^{n}\cdot$ on the right side are independent. Recall that $Var(Y) = Var(\mathbb{E}[Y|X]) + \mathbb{E}[Var(Y|X)]$ for random variables X and Y, so that

$$
\begin{aligned}
& Var(D2) \\
= & \frac{1}{n(n-1)}\mathbb{E}\left[\begin{array}{c} \mathbb{E}_i\left[e_j^2\left(K_{h,ij}^{\gamma-}-K_{h,ij}^{\gamma_0^-}\right)^2 1(q_j\leq\gamma_0)\right]+\mathbb{E}_i\left[e_j^2\left(K_{h,ij}^{\gamma_0^+}-K_{h,ij}^{\gamma+}\right)^2 1(q_j>\gamma)\right] \\ +\mathbb{E}_i\left[e_j^2\left(K_{h,ij}^{\gamma-}+K_{h,ij}^{\gamma_0^+}\right)^2 1\left(\gamma_0<q_j\leq\gamma\right)\right] \end{array}\right] \\
\leq & \frac{C}{n(n-1)h^{2d}}\left[h^d\left(\frac{\gamma-\gamma_0}{h}\right)^2+h^{d-1}\left|\gamma-\gamma_0\right|\right]\leq\frac{C\left|\gamma-\gamma_0\right|}{n(n-1)h^{d+1}},
\end{aligned}
$$

uniformly for $|\gamma-\gamma_0|<\eta$, where $\mathbb{E}_i[\cdot]=\mathbb{E}[\cdot|x_i]$. In consequence,

$$
\begin{aligned}
& P\left(\sup_{|\gamma-\gamma_0|<\eta}\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\Delta}_i(\gamma)-\widehat{\Delta}_i(\gamma_0)\right)>0\right) \\
\leq & C\,\mathbb{E}\left[\left(\sup_{|\gamma-\gamma_0|<\eta}D2\right)^2\right]\Big/\left(\frac{|\gamma-\gamma_0|}{h}\right)^2 \\
\leq & \frac{C\left|\gamma-\gamma_0\right|}{n(n-1)h^{d+1}}\Big/\frac{(\gamma-\gamma_0)^2}{h^2}\leq\frac{C}{n\left|\gamma-\gamma_0\right|}\frac{1}{nh^{d-1}},
\end{aligned}
$$

by Markov's inequality. So

$$
\begin{aligned}
& \sum_{l=L}^{\log_2(nh)}P\left(\sup_{S_{l,n}}\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\Delta}_i(\gamma)-\widehat{\Delta}_i(\gamma_0)\right)>0\right) \\
\leq & \sum_{l\geq L}\frac{C}{n\cdot 2^l/n}\frac{1}{nh^{d-1}}=\frac{C}{nh^{d-1}}\sum_{l\geq L}\frac{1}{2^l}\to 0
\end{aligned}
$$

as $L\to\infty$, and the proof is complete. ■

Proposition 3 *For v in any compact set of $\mathbb{R}$,*

$$
\begin{aligned}
& nh\left(\widehat{Q}_n\left(\gamma_0+\frac{v}{n}\right)-\widehat{Q}_n(\gamma_0)\right)\Big/2k_+(0) \\
= & -\sum_{i=1}^{n}\overline{z}_{1i}1\left(\gamma_0-\frac{v}{n}<q_i\leq\gamma_0\right)-\sum_{i=1}^{n}\overline{z}_{2i}1\left(\gamma_0<q_i\leq\gamma_0+\frac{v}{n}\right)+o_p(1).
\end{aligned}
$$

Proof. We use the same notation as the last proposition and denote $\gamma_0+\frac{v}{n}$ as γ_0^v. Then

$$
\begin{aligned}
nh\left(\widehat{Q}_n\left(\gamma_0^v\right)-\widehat{Q}_n(\gamma_0)\right) &= \sum_{i=1}^{n}h\widehat{\Delta}_i(\gamma_0^v)^2-\sum_{i=1}^{n}h\widehat{\Delta}_i(\gamma_0)^2 \\
&= \sum_{i=1}^{n}\left(\widehat{\Delta}_i(\gamma_0^v)+\widehat{\Delta}_i(\gamma_0)\right)h\left(\widehat{\Delta}_i(\gamma_0^v)-\widehat{\Delta}_i(\gamma_0)\right).
\end{aligned}
$$

Following Lemma B.1 of Newey (1994), we can show that $\widehat{\Delta}_i(\gamma_0^v)\xrightarrow{p}(1,x_i',\gamma_0)\,\delta_0 f(x_i,\gamma_0)\equiv\Delta_f(x_i)=O_p(1)$ uniformly in i and v, so $\widehat{\Delta}_i(\gamma_0^v)+\widehat{\Delta}_i(\gamma_0)\xrightarrow{p}2\Delta_f(x_i)$ uniformly in i and v. We concentrate on

$h\left(\widehat{\Delta}_i(\gamma_0^v)-\widehat{\Delta}_i(\gamma_0)\right)$. For simplicity, let $v>0$. Now,

$$\begin{aligned}
& h\left(\widehat{\Delta}_i(\gamma_0^v)-\widehat{\Delta}_i(\gamma_0)\right) \\
= & \left(\frac{h}{n-1}\sum_{j=1,j\neq i}^{n} y_j K_{h,ij}^{\gamma_0^v-}-\frac{h}{n-1}\sum_{j=1,j\neq i}^{n} y_j K_{h,ij}^{\gamma_0^v+}\right)-\left(\frac{h}{n-1}\sum_{j=1,j\neq i}^{n} y_j K_{h,ij}^{\gamma_0-}-\frac{h}{n-1}\sum_{j=1,j\neq i}^{n} y_j K_{h,ij}^{\gamma_0+}\right) \\
= & \left[\begin{array}{c} \frac{h}{n-1}\sum_{j=1,j\neq i}^{n}\left(g(x_j,q_j)+\left(1\ x_j'\ q_j\right)\delta_0+e_j\right)1(q_j\leq\gamma_0)K_{h,ij}^{\gamma_0^v-}-\frac{h}{n-1}\sum_{j=1,j\neq i}^{n}\left(g(x_j,q_j)+e_j\right)K_{h,ij}^{\gamma_0^v+} \\ +\frac{h}{n-1}\sum_{j=1,j\neq i}^{n}\left(g(x_j,q_j)+e_j\right)1(\gamma_0<q_j\leq\gamma_0^v)K_{h,ij}^{\gamma_0^v-} \end{array}\right] \\
& -\left[\frac{h}{n-1}\sum_{j=1,j\neq i}^{n}\left(g(x_j,q_j)+\left(1\ x_j'\ q_j\right)\delta_0+e_j\right)K_{h,ij}^{\gamma_0-}-\frac{h}{n-1}\sum_{j=1,j\neq i}^{n}\left(g(x_j,q_j)+e_j\right)K_{h,ij}^{\gamma_0+}\right] \\
= & T_{1i}+T_{2i}+T_{3i}+T_{4i}+T_{5i}+T_{6i},
\end{aligned}$$

where

$$\begin{aligned}
T_{1i} &= -\frac{h}{n-1}\sum_{j=1,j\neq i}^{n} g(x_j,q_j)\left(K_{h,ij}^{\gamma_0^v+}-K_{h,ij}^{\gamma_0+}\right), \\
T_{2i} &= \frac{h}{n-1}\sum_{j=1,j\neq i}^{n}\left[g(x_j,q_j)+\left(1\ x_j'\ q_j\right)\delta_0\right]\left(K_{h,ij}^{\gamma_0^v-}-K_{h,ij}^{\gamma_0-}\right), \\
T_{3i} &= -\frac{h}{n-1}\sum_{j=1,j\neq i}^{n} e_j 1(q_j>\gamma_0^v)\left(K_{h,ij}^{\gamma_0^v+}-K_{h,ij}^{\gamma_0+}\right), \\
T_{4i} &= \frac{h}{n-1}\sum_{j=1,j\neq i}^{n} e_j 1(q_j\leq\gamma_0)\left(K_{h,ij}^{\gamma_0^v-}-K_{h,ij}^{\gamma_0-}\right), \\
T_{5i} &= \frac{h}{n-1}\sum_{j=1,j\neq i}^{n} e_j 1(\gamma_0<q_j\leq\gamma_0^v)K_{h,ij}^{\gamma_0+},(*) \\
T_{6i} &= -\frac{h}{n-1}\sum_{j=1,j\neq i}^{n}\left[\left(1\ x_j'\ q_j\right)\delta_0-e_j\right]1(\gamma_0<q_j\leq\gamma_0^v)K_{h,ij}^{\gamma_0^v-}.(*)
\end{aligned}$$

Our target is to show that

$$\sum_{i=1}^{n}\left(T_{1i}+T_{2i}+T_{3i}+T_{4i}\right)=o_p(1),$$

and

$$\begin{aligned}
\sum_{i=1}^{n}\left(T_{5i}+T_{6i}\right)\Delta_f(x_i) &= k_+(0)\sum_{i=1}^{n}\left[-\left(1,x_i',\gamma_0\right)\delta_0+2e_i\right]f(x_i)\Delta_f(x_i)1\left(\gamma_0<q_i\leq\gamma_0^v\right)+o_p(1) \\
&= -k_+(0)\sum_{i=1}^{n}\overline{z}_{2i}1\left(\gamma_0<q_i\leq\gamma_0^v\right)+o_p(1).
\end{aligned}$$

The first result is shown in Lemma 2, and the second is shown in Lemma 3. ■

Proposition 4 *On any compact set of v, $nh^d\left(\widehat{\Delta}_o\left(\gamma_0+\frac{v}{a_n}\right)-\widehat{\Delta}_o\left(\gamma_0\right)\right)\Rightarrow D_o(v)$.*

Proof. The proof proceeds by establishing convergence of the finite dimensional distributions of $R(v) \equiv nh^d\left(\widehat{\Delta}_o\left(\gamma_0^v\right) - \widehat{\Delta}_o\left(\gamma_0\right)\right)$ to those of $D_o(v)$ and then showing that $R(v)$ is tight, where $\gamma_0^v = \gamma_0 + \frac{v}{a_n}$.

From the last proposition, $R(v)$ can be written as the sum of six terms:

$$R(v) = \sum_{l=1}^{6} T_l^{+} 1(v > 0) + \sum_{l=1}^{6} T_l^{-} 1(v < 0),$$

where T_l^+ is the same as T_{li} except that $\frac{h}{n-1}$ in T_{li} is changed to h^d, x_i is changed to x_o, $\sum_{j=1,j\neq i}^{n}$ changes to $\sum_{j=1}^{n}$, and $K_{h,ij}^{\gamma\pm}$ changes to $K_{h,j}^{\gamma\pm}$, and

$$\begin{aligned}
T_1^- &= h^d \sum_{j=1}^{n} g(x_j, q_j)\left(K_{h,j}^{\gamma_0+} - K_{h,j}^{\gamma_0^v+}\right), \\
T_2^- &= h^d \sum_{j=1}^{n} \left[g(x_j, q_j) + \left(1\ x_j'\ q_j\right)\delta_0\right]\left(K_{h,j}^{\gamma_0^v-} - K_{h,j}^{\gamma_0-}\right), \\
T_3^- &= -h^d \sum_{j=1}^{n} e_j 1(q_j > \gamma_0)\left(K_{h,j}^{\gamma_0^v+} - K_{h,j}^{\gamma_0+}\right), \\
T_4^- &= h^d \sum_{j=1}^{n} e_j 1(q_j \le \gamma_0^v)\left(K_{h,j}^{\gamma_0^v-} - K_{h,j}^{\gamma_0-}\right), \\
T_5^- &= -h^d \sum_{j=1}^{n} e_j 1(\gamma_0^v < q_j \le \gamma_0) K_{h,j}^{\gamma_0-}, (*) \\
T_6^- &= -h^d \sum_{j=1}^{n} \left[\left(1\ x_j'\ q_j\right)\delta_0 + e_j\right] 1(\gamma_0^v < q_j \le \gamma_0) K_{h,j}^{\gamma_0^v+}. (*)
\end{aligned}$$

Lemma 4 shows that $\sum_{l=1}^{4} T_l^+ + \sum_{l=1}^{4} T_l^- = o_p(1)$ uniformly in v, and Lemma 5 shows that for a fixed v,

$$T_5^+ + T_6^+ + T_5^- + T_6^- \xrightarrow{d} D_o(v).$$

We next show the tightness of $T_5^+ + T_6^+ + T_5^- + T_6^-$. Take T_5^+ to illustrate the argument. Suppose v_1 and v_2, $0 < v_1 < v_2 < \infty$, are stopping times. Then for any $\epsilon > 0$,

$$\begin{aligned}
& P\left(\sup_{|v_2 - v_1| < \eta} \left|T_5^+(v_2) - T_5^+(v_1)\right| > \epsilon\right) \\
\le\ & P\left(\sum_{j=1}^{n} K\left(\frac{x_j - x_o}{h}\right) k_+\left(\frac{q_j - \gamma_0}{h}\right) |e_j| \sup_{|v_2 - v_1| < \eta} 1(\gamma_0^{v_1} < q_j \le \gamma_0^{v_2}) > \epsilon\right) \\
\le\ & \sum_{j=1}^{n} \mathbb{E}\left[K\left(\frac{x_j - x_o}{h}\right) k_+\left(\frac{q_j - \gamma_0}{h}\right) |e_j| \sup_{|v_2 - v_1| < \eta} 1(\gamma_0^{v_1} < q_j \le \gamma_0^{v_2})\right] \Big/ \epsilon \\
\le\ & C\eta/\epsilon,
\end{aligned}$$

where the second inequality is from Markov's inequality, and C in the last inequality can take

$$\sup_{(x,q)\in N}\mathbb{E}\left[|e|\,|x,q\right]f(x,q)\sup_{u_x,u_q}K(u_x)k_+(u_q)$$

with N being a neighborhood of $(x_o',\gamma_0)'$. The required result now follows. ■

Proposition 5 $\begin{pmatrix}\sqrt{nh}\left(\widehat{\Delta}-\widehat{\Delta}^0\right)\\ \sqrt{nh}h\left(\widehat{\delta}_{xq}-\widehat{\delta}_{xq}^0\right)\end{pmatrix}\xrightarrow{p}0.$

Proof. We need only to show

$$\frac{1}{n}\sum_{i=1}^n k_h\left(\frac{q_i-\widehat{\gamma}}{h}\right)\begin{pmatrix}\sqrt{nh}\left(\widehat{a}_-(x_i)-\widehat{a}_+(x_i)\right)\\ \sqrt{nh}h\left(\widehat{b}_-(x_i)-\widehat{b}_+(x_i)\right)\end{pmatrix}-\frac{1}{n}\sum_{i=1}^n k_h\left(\frac{q_i-\gamma_0}{h}\right)\begin{pmatrix}\sqrt{nh}\left(\widehat{a}_-^0(x_i)-\widehat{a}_+^0(x_i)\right)\\ \sqrt{nh}h\left(\widehat{b}_-^0(x_i)-\widehat{b}_+^0(x_i)\right)\end{pmatrix}\xrightarrow{p}0,$$

and

$$\sqrt{nh}\left(\frac{1}{n}\sum_{i=1}^n k_h\left(\frac{q_i-\widehat{\gamma}}{h}\right)-\frac{1}{n}\sum_{i=1}^n k_h\left(\frac{q_i-\gamma_0}{h}\right)\right)\xrightarrow{p}0. \tag{21}$$

It is easy to see that the first result is implied by

$$\begin{aligned}&\sqrt{nh}\left[\left(\widehat{a}_-(x_i)-\widehat{a}_+(x_i)\right)-\left(\widehat{a}_-^0(x_i)-\widehat{a}_+^0(x_i)\right)\right]\xrightarrow{p}0\text{ uniformly in }x_i,\\ &\sqrt{nh}h\left[\left(\widehat{b}_-(x_i)-\widehat{b}_+(x_i)\right)-\left(\widehat{b}_-^0(x_i)-\widehat{b}_+^0(x_i)\right)\right]\xrightarrow{p}0\text{ uniformly in }x_i.\end{aligned}$$

Since $\widehat{\gamma}-\gamma_0=O_p(n^{-1})$, $\widehat{\gamma}$ falls into $\left[\gamma_0-\frac{C}{n},\gamma_0+\frac{C}{n}\right]$ for some positive C with any large probability when n is large enough. So we can just prove these results by replacing $\widehat{\gamma}$ by $\gamma_0+\frac{C}{n}\equiv\gamma_0^C$. The corresponding $\widehat{a}_\pm(x_i)$ and $\widehat{b}_\pm(x_i)$ are denoted as $\widehat{a}_\pm^C(x_i)$ and $\widehat{b}_\pm^C(x_i)$. Since the results for $\widehat{a}_-(x_i)$ and $\widehat{b}_-(x_i)$ are similarly proved, we need only prove that

$$\begin{aligned}&\sqrt{nh}\left[\widehat{a}_+^C(x_i)-\widehat{a}_+^0(x_i)\right]\xrightarrow{p}0\text{ uniformly in }x_i,\\ &\sqrt{nh}h\left[\widehat{b}_+^C(x_i)-\widehat{b}_+^0(x_i)\right]\xrightarrow{p}0\text{ uniformly in }x_i.\end{aligned} \tag{22}$$

Without loss of generality, suppose $C>0$. Lemma 6 shows (21), and Lemma 7 shows (22). ■

Proposition 6 *$I_{1n}^{(1)}$ is $o_p(1)$ under $H_0^{(1)}$, and is $O_p\left(nh^{d/2}\right)$ under $H_1^{(1)}$.*

Proof. Note that

$$\begin{aligned}I_{1n}&=\frac{nh^{d/2}}{n(n-1)}\sum_i\sum_{j\neq i}D_iD_jK_{h,ij}\\ &=\frac{nh^{d/2}}{n(n-1)}\sum_i\sum_{j\neq i}\left[m_i-\mathbf{x}_i'\widehat{\beta}-\mathbf{x}_i'\widehat{\delta}1\left(q_i\leq\widehat{\gamma}\right)\right]\left[m_j-\mathbf{x}_j'\widehat{\beta}-\mathbf{x}_j'\widehat{\delta}1\left(q_j\leq\widehat{\gamma}\right)\right]K_{h,ij}.\end{aligned}$$

Under $H_0^{(1)}$, $m_i=\mathbf{x}_i'\beta_0+\mathbf{x}_i'\delta_01\left(q_i\leq\gamma_0\right)$, so that

$$\begin{aligned}&m_i-\mathbf{x}_i'\widehat{\beta}-\mathbf{x}_i'\widehat{\delta}1\left(q_i\leq\widehat{\gamma}\right)\\ =\;&\mathbf{x}_i'\left(\beta_0-\widehat{\beta}\right)+\mathbf{x}_i'\left(\delta_0-\widehat{\delta}\right)1\left(q_i\leq\widehat{\gamma}\wedge\gamma_0\right)\\ &+\mathbf{x}_i'\delta_01\left(\widehat{\gamma}<q_i\leq\gamma_0\right)-\mathbf{x}_i'\widehat{\delta}1\left(\gamma_0<q_i\leq\widehat{\gamma}\right).\end{aligned}$$

As a result, I_{1n} has ten terms with a typical term of the form

$$T1=\left(\widehat{\beta}-\beta_0\right)'\left[\frac{nh^{d/2}}{n\left(n-1\right)}\sum_i\sum_{j\neq i}K_{h,ij}\mathbf{x}_i\mathbf{x}_j'\right]\left(\widehat{\beta}-\beta_0\right)$$

or

$$T2=\delta_0'\left[\frac{nh^{d/2}}{n\left(n-1\right)}\sum_i\sum_{j\neq i}K_{h,ij}\mathbf{x}_i\mathbf{x}_j'1\left(\widehat{\gamma}<q_i\leq\gamma_0\right)1\left(\widehat{\gamma}<q_j\leq\gamma_0\right)\right]\delta_0.$$

Given that $\widehat{\beta}-\beta_0=O_p(n^{-1/2})$, $\widehat{\delta}-\delta_0=O_p(n^{-1/2})$, and $\widehat{\gamma}-\gamma_0=O_p(n^{-1})$, it is easy to show that $T1=O_p(h^{d/2})$ and $T2=O_p(h^{d/2})$ since $\frac{1}{n(n-1)}\sum\limits_i\sum\limits_{j\neq i}K_{h,ij}\mathbf{x}_i\mathbf{x}_j'=O_p(1)$ and $\frac{1}{n(n-1)}\sum\limits_i\sum\limits_{j\neq i}K_{h,ij}\mathbf{x}_i\mathbf{x}_j'1\left(\widehat{\gamma}<q_i\leq\gamma_0\right)1\left(\widehat{\gamma}<q_j\leq\gamma_0\right)=O_p\left(n^{-1}\right)$.

We now analyze I_{1n} under $H_1^{(1)}$. There are three cases. Let

$$\left(\beta_o',\delta_o',\gamma_o\right)'=\arg\inf_{\beta,\delta,\gamma}\mathbb{E}\left[\left(y-\mathbf{x}'\beta+\mathbf{x}'\delta1\left(q\leq\gamma\right)\right)^2\right].$$

If $\delta_o=0$, then $\overline{m}(x,q)=\mathbf{x}'\beta_o$ and the model degenerates to the case analyzed in Zheng (1996). If $\delta_{xo}=0$ and $\delta_{\alpha o}+\gamma_o\delta_{qo}=0$, then $\overline{m}(x,q)$ takes the continuous threshold regression form of Chan and Tsay (1998). It follows that $\widehat{\beta}-\beta_o=O_p(n^{-1/2})$, $\widehat{\delta}-\delta_o=O_p(n^{-1/2})$, and $\widehat{\gamma}-\gamma_o=O_p(n^{-1/2})$. If $\delta_o\neq0$, then $\widehat{\beta}-\beta_o=O_p(n^{-1/2})$, $\widehat{\delta}-\delta_o=O_p(n^{-1/2})$, and $\widehat{\gamma}-\gamma_o=O_p(n^{-1})$. See Yu (2013a) for these results. We concentrate on the last case. Now,

$$I_{1n}=\frac{nh^{d/2}}{n\left(n-1\right)}\sum_i\sum_{j\neq i}\left(m_i-\overline{m}_i\right)\left(m_j-\overline{m}_j\right)K_{h,ij}(1+o_p(1)),$$

where $\overline{m}_i=\mathbf{x}_i'\beta_o-\mathbf{x}_i'\delta_o1\left(q_i\leq\gamma_o\right)$, so we need only calculate $\mathbb{E}\left[\left(m_i-\overline{m}_i\right)\left(m_j-\overline{m}_j\right)K_{h,ij}\right]$, which is equal to

$$\begin{aligned}&\int\left(m_i-\overline{m}_i\right)\left(m_j-\overline{m}_j\right)K_{h,ij}f_if_jdx_idq_idx_jdq_j\\\approx&\int\left(m_i-\overline{m}_i\right)^2K^x(u_x,x_i)k(u_q)f_i^2dx_idq_idu_xdu_q\\=&\int\left(m_i-\overline{m}_i\right)^2f_i^2dx_idq_i,\end{aligned}$$

The result follows. ■

Proposition 7 *$I_{1n}^{(2)}$ is $o_{P_m}(1)$ uniformly in m under $H_0^{(2)}$, and is $O_p\left(nh^{d/2}b\right)$ under $H_1^{(2)}$.*

Proof. Given that $\widehat{f}_i^{-1}=f_i^{-1}+o_p(1)$ and f_i is bounded uniformly over $(x_i,q_i)\in\mathcal{X}\times\Gamma$,

$$\begin{aligned}I_{1n}&=\frac{nh^{d/2}}{n\left(n-1\right)}\sum_i\sum_{j\neq i}1_i^\Gamma1_j^\Gamma\left(m_i-\widehat{m}_i\right)\widehat{f}_i\left(m_j-\widehat{m}_j\right)\widehat{f}_jK_{h,ij}\left(\widehat{f}_i^{-1}\widehat{f}_j^{-1}\right)\qquad(23)\\&\approx\frac{nh^{d/2}}{n\left(n-1\right)^3}\sum_i\sum_{j\neq i}\sum_{l\neq i}\sum_{k\neq j}1_i^\Gamma\left(m_i-m_l\right)L_{b,il}1_j^\Gamma\left(m_j-m_k\right)L_{b,jk}K_{h,ij}f_i^{-1}f_j^{-1}\\&=O_p\left(\frac{nh^{d/2}}{n\left(n-1\right)^3}\sum_i\sum_{j\neq i}\sum_{l\neq i}\sum_{k\neq j}1_i^\Gamma\left(m_i-m_l\right)L_{b,il}1_j^\Gamma\left(m_j-m_k\right)L_{b,jk}K_{h,ij}\right).\end{aligned}$$

Mimicking the proof of Proposition A.1 of Fan and Li (1996), we can show that under $H_0^{(2)}$, $I_{1n} = O_p\left(nh^{d/2}b^{2\eta}\right) = o_p(1)$. The only new result we need to employ is that $\left|\mathbb{E}_1\left[\left(m_2 - m_1\right) L_{b,21}\right]\right| = O_p\left(b^{\eta}\right)$, which is accomplished in Lemma 8.

We now analyze I_{1n} under $H_1^{(2)}$. It can be shown that the case where i, j, l, k are all different from each other dominates in the formula of the second equality of (23), so

$$I_{1n} \approx O_p(nh^{d/2}\mathbb{E}\left[1_1^{\Gamma}\left(m_1 - m_2\right) L_{b,12} 1_3^{\Gamma}\left(m_3 - m_4\right) L_{b,34} K_{h,13} f_1^{-1} f_3^{-1}\right]).$$

Because $h/b \to 0$, we can treat $(x_1, q_1) = (x_3, q_3)$. Specifically,

$$\begin{aligned}
& \mathbb{E}\left[1_1^{\Gamma}\left(m_1 - m_2\right) L_{b,12} 1_3^{\Gamma}\left(m_3 - m_4\right) L_{b,34} K_{h,13} f_1^{-1} f_3^{-1}\right] \\
= \ & \mathbb{E}\left[\begin{array}{c} 1_1^{\Gamma}\left(m_1 - m_2\right) L_{b,12} f_1^{-1} \int 1\left(q_1 + u_q h \in \Gamma\right)\left(m((x_1, q_1) + uh) - m_4\right) \\ \frac{1}{b^d} L^x\left(\frac{x_4 - x_1 - u_x h}{b}, x_1 + u_x h\right) l\left(\frac{q_4 - q_1 - u_q h}{b}\right) K^x\left(u_x, x_1\right) k\left(u_q\right) du \end{array}\right] \\
\approx \ & \mathbb{E}\left[1_1^{\Gamma}\left(m_1 - m_2\right) L_{b,12}\left(m_1 - m_4\right) L_{b,14} f_1^{-1}\right] \\
= \ & \mathbb{E}\left\{1_1^{\Gamma} f_1^{-1}\left\{\mathbb{E}_1\left[\left(m_1 - m_2\right) L_{b,12}\right]\right\}^2\right\} \\
= \ & \int_{\underline{\gamma}}^{\overline{\gamma}} \int\left[\int\left(m(x_1, q_1) - m(x_2\, q_2)\right) \frac{1}{b^d} L^x\left(\frac{x_2 - x_1}{b}, x_1\right) l\left(\frac{q_2 - q_1}{b}\right) f(x_2, q_2) dx_2 dq_2\right]^2 dx_1 dq_1 \\
\approx \ & O\left(b^{2\eta}\right) + \int_{\gamma_0 - b}^{\gamma_0 + b} \int\left[\int\left(m(x_1, q_1) - m(x_1 + u_x b, q_1 + u_q b)\right) L^x\left(u_x, x_1\right) l\left(u_q\right) f(x_1 + u_x b, q_1 + u_q b) du\right]^2 dx_1 dq_1 \\
\approx \ & O\left(b^{2\eta}\right) + \int_{\gamma_0 - b}^{\gamma_0 + b} \int\left[\begin{array}{c} \int_{\frac{\gamma_0 - q_1}{b}}^{1}\left(m(x_1, q_1) - m(x_1, q_1 + u_q b)\right) l\left(u_q\right) du \\ + \int_{-1}^{\frac{\gamma_0 - q_1}{b}}\left(m(x_1, q_1) - m(x_1, q_1 + u_q b)\right) l\left(u_q\right) du_q \end{array}\right]^2 f(x_1, \gamma_0)^2 dx_1 dq_1
\end{aligned}$$

where $u = (u_x, u_q)$. Under $H_1^{(2)}$,

$$\begin{aligned}
& \int_{\gamma_0 - b}^{\gamma_0 + b} \int\left[\begin{array}{c} \int_{\frac{\gamma_0 - q_1}{b}}^{1}\left(m(x_1, q_1) - m(x_1, q_1 + u_q b)\right) l\left(u_q\right) du \\ + \int_{-1}^{\frac{\gamma_0 - q_1}{b}}\left(m(x_1, q_1) - m(x_1, q_1 + u_q b)\right) l\left(u_q\right) du_q \end{array}\right]^2 f(x_1, \gamma_0)^2 dx_1 dq_1 \\
\approx \ & \int_{\gamma_0}^{\gamma_0 + b} \int\left[-\int_{\frac{\gamma_0 - q_1}{b}}^{1} m_+'(x_1) u_q b l(u_q) du_q + \int_{-1}^{\frac{\gamma_0 - q_1}{b}}\left(m_+(x_1) - m_-(x_1) + C u_q b\right) l(u_q) du_q\right]^2 f(x_1, \gamma_0)^2 dx_1 dq_1 \\
& + \int_{\gamma_0 - b}^{\gamma_0} \int\left[\int_{\frac{\gamma_0 - q_1}{b}}^{1}\left(m_-(x_1) - m_+(x_1) + C u_q b\right) l(u_q) du_q - \int_{-1}^{\frac{\gamma_0 - q_1}{b}} m_-'(x_1) u_q b l(u_q) du_q\right]^2 f(x_1, \gamma_0)^2 dx_1 dq_1 \\
\approx \ & b \int_0^1 \int\left[\int_{-1}^{-v}\left(m_+(x_1) - m_-(x_1)\right) l(u_q) du_q\right]^2 dx_1 dv \\
& + b \int_0^1 \int\left[\int_v^1\left(m_+(x_1) - m_-(x_1)\right) l(u_q) du_q\right]^2 f(x_1, \gamma_0)^2 dx_1 dv \\
= \ & 2b \int\left(m_+(x_1) - m_-(x_1)\right)^2 f(x_1, \gamma_0)^2 dx \int_0^1\left(\int_v^1 l(u_q) du_q\right)^2 dv,
\end{aligned}$$

where $m_{\pm}'(x) = \lim_{\gamma \to \gamma_0 \pm} \partial m(x, \gamma)/\partial\gamma$, and $m_{\pm}(x) = \lim_{\gamma \to \gamma_0 \pm} m(x, \gamma)$. The result follows. ■

Proposition 8 *$I_{2n}^{(1)} \xrightarrow{d} N\left(0, \Sigma^{(1)}\right)$ and $I_{2n}^{(2)} \xrightarrow{d} N\left(0, \Sigma^{(2)}\right)$.*

Proof. We only prove the second result since the first is proved in a similar way.

$$\begin{aligned} I_{2n}^{(2)} &= \frac{nh^{d/2}}{n(n-1)} \sum_i \sum_{j \neq i} 1_i^{\Gamma} 1_j^{\Gamma} u_i u_j K_{h,ij} \\ &\equiv \frac{nh^{d/2}}{n(n-1)} \sum_i \sum_{j \neq i} H_n(z_i, z_j) \equiv nh^{d/2} U_n, \end{aligned}$$

where U_n is a second order degenerate U-statistic with kernel function H_n. We can apply theorem 1 of Hall (1984) to find its asymptotic distribution. Two conditions need to be checked: (i) $\mathbb{E}[H_n^2(z_1, z_2)] < \infty$; (ii)

$$\frac{\mathbb{E}\left[G_n^2(z_1, z_2)\right] + n^{-1}\mathbb{E}[H_n^4(z_1, z_2)]}{\mathbb{E}^2[H_n^2(z_1, z_2)]} \to 0 \text{ as } n \to \infty,$$

where $G_n(z_1, z_2) = \mathbb{E}[H_n(z_3, z_1)H_n(z_3, z_2)|z_1, z_2]$. Because these checks follow similarly to lemma 3.3a of Zheng (1996) they are omitted here to save space. In conclusion

$$nU_n/\sqrt{2\mathbb{E}[H_n^2(z_1, z_2)]} \xrightarrow{d} N(0, 1).$$

It is easy to check that

$$\begin{aligned} \mathbb{E}[H_n^2(z_1, z_2)] &= \mathbb{E}\left[1_1^{\Gamma} 1_2^{\Gamma} K_{h,12}^2 \mathbb{E}[u_1^2|x_1, q_1]\mathbb{E}[u_2^2|x_2, q_2]\right] \\ &= \int_{\underline{\gamma}}^{\overline{\gamma}} \int \int_{\underline{\gamma}}^{\overline{\gamma}} \int \frac{1}{h^{2d}} K^x \left(\frac{x_2 - x_1}{h}, x_1\right)^2 k^2 \left(\frac{q_2 - q_1}{h}\right) \sigma^2(x_1, q_1)\sigma^2(x_2, q_2) f(x_1, q_1) f(x_2, q_2) dx_2 dq_2 dx_1 dq_1 \\ &= \int_{\underline{\gamma}}^{\overline{\gamma}} \int \int_{\frac{\underline{\gamma}-q}{h}}^{\frac{\overline{\gamma}-q}{h}} \int \frac{1}{h^d} K^x (u_x, x)^2 k^2(u_q)\sigma^2(x, q)\sigma^2(x + u_x h, q + u_q h) f(x, q) f(x + u_x h, q + u_q h) du dx dq \\ &\approx \frac{1}{h^d} \int_{\underline{\gamma}}^{\overline{\gamma}} \int \left[\int K^x (u_x, x)^2 k^2(u_q) du\right] \sigma^4(x, q) f^2(x, q) dx dq + o\left(\frac{1}{h^d}\right) \\ &\approx \frac{1}{h^d} \left[\int k^{2d}(u) du\right] \int_{\underline{\gamma}}^{\overline{\gamma}} \int \sigma^4(x, q) f^2(x, q) dx dq = \frac{1}{h^d} \frac{\Sigma^{(2)}}{2}, \end{aligned}$$

so the result follows. ■

Proposition 9 *$v_n^{(1)2} = \Sigma^{(1)} + o_p(1)$ under $H_0^{(1)}$ and the local alternative and $v_n^{(1)2} = O_p(1)$ under $H_1^{(1)}$.*

Proof. It can be shown that

$$\begin{aligned} &\frac{h^d}{n(n-1)} \sum_i \sum_{j \neq i} K_{h,ij}^2 \widehat{e}_i^2 \widehat{e}_j^2 \\ &= \frac{h^d}{n(n-1)} \sum_i \sum_{j \neq i} K_{h,ij}^2 (u_i + m_i - \overline{m}_i)^2 (u_j + m_j - \overline{m}_j)^2 + o_p(1) \\ &= h^d \mathbb{E}\left[K_{h,ij}^2 (u_i + m_i - \overline{m}_i)^2 (u_j + m_j - \overline{m}_j)^2\right] + o_p(1) \\ &= h^d \mathbb{E}\left[K_{h,ij}^2 \mathbb{E}_i\left[(u_i + m_i - \overline{m}_i)^2\right] \mathbb{E}_j\left[(u_j + m_j - \overline{m}_j)^2\right]\right] + o_p(1) \\ &= \int \int \left[\int K^x (u_x, x_i)^2 k^2(u_q) du\right] \left(\sigma_i^2 + (m_i - \overline{m}_i)^2\right)^2 f_i^2 dx_i dq_i + o_p(1) \\ &= \int k^{2d}(u) du \mathbb{E}\left[f(x, q)\left(\sigma^2(x, q) + (m - \overline{m})^2\right)^2\right] + o_p(1), \end{aligned}$$

where $\sigma_i^2 = \sigma^2(x_i, q_i)$. Under $H_0^{(1)}$, $m - \overline{m} = 0$. Under the local alternative, $\mathbb{E}\left[f(x,q)(m-\overline{m})^4\right] = O\left(n^{-2}h^{-d}\right) = o(1)$, and under $H_1^{(1)}$, $\mathbb{E}\left[f(x,q)(m-\overline{m})^4\right] = O(1)$. ■

Proposition 10 $v_n^{(2)2} = \Sigma^{(2)} + o_p(1)$ *under both* $H_0^{(2)}$ *and* $H_1^{(2)}$.

Proof. By similar steps as the last proposition, we have

$$\begin{aligned}
&\frac{h^d}{n(n-1)} \sum_i \sum_{j \neq i} 1_i^\Gamma 1_j^\Gamma K_{h,ij}^2 \widehat{e}_i^2 \widehat{e}_j^2 \\
= & \int k^{2d}(u) du \mathbb{E}\left[1_i^\Gamma f(x_i, q_i)\left(\sigma^2(x_i, q_i) + (m_i - \overline{m}_i)^2\right)^2\right] + o_p(1),
\end{aligned}$$

where $\overline{m}_i$ is redefined as $\mathbb{E}_i[\widehat{m}_i]$. Note that $\mathbb{E}\left[1_i^\Gamma f(x_i, q_i)(m_i - \overline{m}_i)^4\right]$ is at most $O(b)$ since $m_i - \overline{m}_i$ contributes only for $q \in [\gamma - b, \gamma + b]$. ■

Appendix C: Lemmas

Lemma 1 $\sup_{\gamma \in \Gamma}\left|\widehat{Q}_n(\gamma) - Q_n(\gamma)\right| \xrightarrow{p} 0$.

Proof. Noting that $\Gamma \times \mathcal{X}$ is compact we have from Lemma B.1 of Newey (1994) that

$$\sup_{\gamma \in \Gamma, x_i \in \mathcal{X}}\left|\widehat{\Delta}_i(\gamma) - \mathbb{E}_i[\widehat{\Delta}_i(\gamma)]\right| = O_p\left(\sqrt{\ln n / nh^d}\right).$$

Given that $\sup_{\gamma \in \Gamma, x_i \in \mathcal{X}} \mathbb{E}_i\left[\widehat{\Delta}_i(\gamma)\right] = O_p(1)$,

$$\begin{aligned}
\widehat{Q}_n(\gamma) &= \frac{1}{n}\sum_{i=1}^n \widehat{\Delta}_i^2(\gamma) = \frac{1}{n}\sum_{i=1}^n \left[\mathbb{E}_i\left[\widehat{\Delta}_i(\gamma)\right] + \left(\widehat{\Delta}_i(\gamma) - \mathbb{E}_i[\widehat{\Delta}_i(\gamma)]\right)\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_i^2\left[\widehat{\Delta}_i(\gamma)\right] + O_p\left(\sqrt{\ln n / nh^d}\right),
\end{aligned}$$

uniformly in γ. By a Glivenko-Cantelli theorem,

$$\sup_{\gamma \in \Gamma}\left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}_i^2\left[\widehat{\Delta}_i(\gamma)\right] - \mathbb{E}\left[\mathbb{E}_i^2\left[\widehat{\Delta}_i(\gamma)\right]\right]\right| \xrightarrow{p} 0.$$

Note that $\mathbb{E}\left[\mathbb{E}_i^2\left[\widehat{\Delta}_i(\gamma)\right]\right] = Q_n(\gamma)$, the result of interest follows. ■

Lemma 2 $\sum_{i=1}^n \sum_{l=1}^4 T_{li} = o_p(1)$ *uniformly in* v.

Proof. We take T_{4i} to illustrate and have

$$\begin{aligned}\sum_{i=1}^{n} T_{4i} &= \frac{h}{n-1}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} e_j 1(q_j \leq \gamma_0)\left(K_{h,ij}^{\gamma_0^v -} - K_{h,ij}^{\gamma_0 -}\right) \\ &= \frac{h}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} e_j 1(q_j \leq \gamma_0) n\left(K_{h,ij}^{\gamma_0^v -} - K_{h,ij}^{\gamma_0 -}\right) \\ &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} e_j 1(q_j \leq \gamma_0)\frac{1}{h}K_{h,ij}^{x} nh\left[k_-\left(\frac{q_j-\gamma_0^v}{h}\right) - k_-\left(\frac{q_j-\gamma_0}{h}\right)\right] \\ &= O\left(\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} e_j \frac{1(\gamma_0 - h \leq q_j \leq \gamma_0)}{h} K_{h,ij}^{x}\right) \equiv O\left(\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} P_n(X_i, X_j)\right)\end{aligned}$$

uniformly in v, where the second to last equality is from the Lipschitz continuity of $k_-(\cdot)$. By the U-statistic projection, see, e.g., Lemma 8.4 of Newey and McFadden (1994),

$$\begin{aligned}&\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} P_n(X_i, X_j) \\ &= \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[P_n(X_i, X_j)|X_j\right] + O_p\left(\frac{1}{n}\mathbb{E}\left[P_n(X_i, X_j)^2\right]^{1/2}\right).\end{aligned}$$

In our case, $\mathbb{E}\left[P_n(X_i, X_j)|X_j\right] = e_j \frac{1(\gamma_0 - h \leq q_j \leq \gamma_0)}{h}\int K_{h,ij}^{x} f(x_i)dx_i = O(e_j 1(\gamma_0 - h \leq q_j \leq \gamma_0)/h)$, and $\mathbb{E}\left[P_n(X_i, X_j)^2\right] \approx \frac{1}{h^d}\int \sigma^2(x_i, \gamma_0) f(x_i)dx_i = O\left(\frac{1}{h^d}\right)$, so

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[P_n(X_i, X_j)|X_j\right] &= O_p\left(\frac{1}{nh}\sum_{j=1} e_j 1(\gamma_0 - h \leq q_j \leq \gamma_0)\right) = o_p(1), \\ \frac{1}{n}\mathbb{E}\left[P_n(X_i, X_j)^2\right]^{1/2} &= O\left(\frac{1}{nh^{d/2}}\right) = o(1).\end{aligned}$$

■

Lemma 3 $\sum_{i=1}^{n}(T_{5i} + T_{6i})\Delta_f(x_i) = -k_+(0)\sum_{i=1}^{n}\left[(1, x_i', \gamma_0)\delta_0 - 2e_i\right] 1(\gamma_0 < q_i \leq \gamma_0^v) f(x_i)\Delta_f(x_i) + o_p(1).$

Proof. $\sum_{i=1}^{n} T_{5i}\Delta_f(x_i)$ is a U-statistic and we write

$$\begin{aligned}\sum_{i=1}^{n} T_{5i}\Delta_f(x_i) &= \frac{h}{n-1}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} e_j 1(\gamma_0 < q_j \leq \gamma_0^v) K_{h,ij}^{\gamma_0 +}\Delta_f(x_i) \equiv \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} P_n(X_i, X_j) \\ &= \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[P_n(X_i, X_j)|X_j\right] + O_p\left(\frac{1}{n}\mathbb{E}\left[P_n(X_i, X_j)^2\right]^{1/2}\right),\end{aligned}$$

where $P_n(X_i, X_j) = nhe_j 1(\gamma_0 < q_j \leq \gamma_0^v) K_{h,ij}^{\gamma_0+} \Delta_f(x_i)$ with $X_i = (x_i', q_i, e_i)'$, and the last equality is from Lemma 8.4 of Newey and McFadden (1994). Then

$$\begin{aligned}
\mathbb{E}\left[P_n(X_i, X_j) | X_j\right] &\approx ne_j 1(\gamma_0 < q_j \leq \gamma_0^v) k_+(0) \frac{1}{h^{d-1}} \int K_h^x(x_i - x_j, x_i) f(x_i) \Delta_f(x_i) dx_i \\
&\approx ne_j 1(\gamma_0 < q_j \leq \gamma_0^v) k_+(0) \int K^x(u_x, x_j + u_x h) f(x_j + u_x h) \Delta_f(x_j + u_x h) du_x \\
&\approx ne_j 1(\gamma_0 < q_j \leq \gamma_0^v) k_+(0) f(x_j) \Delta_f(x_j),
\end{aligned}$$

and

$$\mathbb{E}\left[P_n(X_i, X_j)^2\right] = O\left(n^2 h^2 \frac{1}{n} \frac{1}{h^{2d}} h^{d-1}\right) = O\left(\frac{n}{h^{d-1}}\right),$$

so that

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\left[P_n(X_i, X_j) | X_j\right] &= \sum_{i=1}^{n} e_i 1(\gamma_0 < q_i \leq \gamma_0^v) k_+(0) f(x_i) \Delta_f(x_i), \\
\frac{1}{n}\mathbb{E}\left[P_n(X_i, X_j)^2\right]^{1/2} &= \frac{1}{n}\sqrt{\frac{n}{h^{d-1}}} = \sqrt{\frac{1}{nh^{d-1}}} = o(1).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\sum_{i=1}^{n} T_{6i} \Delta_f(x_i) &= -\frac{h}{n-1} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \left[\mathbf{x}_j' \delta_0 - e_j\right] 1(\gamma_0 < q_j \leq \gamma_0^v) K_{h,ij}^{\gamma_0^v -} \Delta_f(x_i) \\
&= -\sum_{i=1}^{n} \left[(1, x_i', \gamma_0)\delta_0 - e_i\right] 1(\gamma_0 < q_i \leq \gamma_0^v) k_-(0) f(x_i) \Delta_f(x_i) + o_p(1).
\end{aligned}$$

The result follows by noting that $k_-(0) = k_+(0)$. ∎

Lemma 4 $\sum_{l=1}^{4} T_l^+ + \sum_{l=1}^{4} T_l^- = o_p(1)$ *uniformly in* v.

Proof. Take T_4^+ as an example.

$$\begin{aligned}
T_4^+ &= h^d \sum_{j=1}^{n} e_j 1(q_j \leq \gamma_0) \left(K_{h,j}^{\gamma_0^v -} - K_{h,j}^{\gamma_0 -}\right) \\
&= \frac{1}{nh^{d-1}} \sum_{j=1}^{n} e_j 1(q_j \leq \gamma_0) K^x\left(\frac{x_j - x_o}{h}, x_o\right) nh^d \left[k_-\left(\frac{q_j - \gamma_0^v}{h}\right) - k_-\left(\frac{q_j - \gamma_0}{h}\right)\right] \\
&= O\left(\frac{1}{nh^{d-1}} \sum_{j=1}^{n} e_j 1(\gamma_0 - h \leq q_j \leq \gamma_0) K^x\left(\frac{x_j - x_o}{h}, x_o\right)\right)
\end{aligned}$$

uniformly in v, where the last equality is from the Lipschitz continuity of $k_-(\cdot)$. Since

$$\mathbb{E}[T_4^{+2}] = O\left(\frac{1}{nh^{d-2}}\right) = o(1),$$

$T_4^+ = o_p(1)$. ∎

Lemma 5 $T_5^+ + T_6^+ + T_5^- + T_6^- \xrightarrow{d} D_o(v)$.

Proof. Take $T_5^+ + T_5^-$ as an example. We use the characteristic function to find its weak limit. Define $T_5^- = \sum_{j=1}^n T_{5j}^- \equiv \sum_{j=1}^n t_{5j}^- 1(\gamma_0^{v-} < q_j \le \gamma_0)$ and $T_5^- = \sum_{j=1}^n T_{5j}^+ = \sum_{j=1}^n t_{5j}^+ 1(\gamma_0 < q_j \le \gamma_0^{v+})$, where $t_{5j}^- = -h^d e_j K_{h,j}^{\gamma_0-}$, $t_{5j}^+ = h^d e_j K_{h,j}^{\gamma_0+}$, $v_- < 0$ and $v_+ > 0$. Note that

$$\begin{aligned}
\exp\left\{\sqrt{-1}s^- T_{5j}^-\right\} &= 1 + 1(\gamma_0^{v_-} < q_j \le \gamma_0)\left[\exp\left\{\sqrt{-1}s^- t_{5j}^-\right\} - 1\right], \\
\exp\left\{\sqrt{-1}s^+ T_{5j}^+\right\} &= 1 + 1(\gamma_0 < q_j \le \gamma_0^{v_+})\left[\exp\left\{\sqrt{-1}s^+ t_{5j}^+\right\} - 1\right].
\end{aligned}$$

Hence

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{\sqrt{-1}\left(s^- T_{5j}^- + s^+ T_{5j}^+\right)\right\}\right] = \mathbb{E}\left[\exp\left\{\sqrt{-1}s^- T_{5j}^-\right\}\right]\mathbb{E}\left[\exp\left\{\sqrt{-1}s^+ T_{5j}^+\right\}\right] \\
&\approx 1 + \mathbb{E}\left[1(\gamma_0^{v_-} < q_j \le \gamma_0)\mathbb{E}\left[\exp\left\{\sqrt{-1}s^- t_{5j}^-\right\} - 1\middle| q_j\right]\right] \\
&\quad + \mathbb{E}\left[1(\gamma_0 < q_j \le \gamma_0^{v_+})\mathbb{E}\left[\exp\left\{\sqrt{-1}s^+ t_{5j}^-\right\} - 1\middle| q_j\right]\right] \\
&\approx 1 + h^{d-1}\int_{\gamma_0^{v_-}}^{\gamma_0}\left[\int\int\left[\exp\left\{-\sqrt{-1}s^- e_j K(u_x) k_-\left(\tfrac{q_j-\gamma_0}{h}\right)\right\} - 1\right] f(e_j, x_o|q_j) de_j du_x\right] f(q_j) dq_j \\
&\quad + h^{d-1}\int_{\gamma_0}^{\gamma_0^{v_+}}\left[\int\int\left[\exp\left\{\sqrt{-1}s^+ e_j K(u_x) k_+\left(\tfrac{q_j-\gamma_0}{h}\right)\right\} - 1\right] f(e_j, x_o|q_j) de_j du_x\right] f(q_j) dq_j \\
&\approx 1 + \tfrac{v_-}{n} f_q(\gamma_0)\int\int\left[\exp\left\{-\sqrt{-1}s^- e_j K(u_x) k_-(0)\right\} - 1\right] f(e_j, x_o|q_j = \gamma_0-) de_j du_x \\
&\quad + \tfrac{v_+}{n} f_q(\gamma_0)\int\int\left[\exp\left\{\sqrt{-1}s^+ e_j K(u_x) k_+(0)\right\} - 1\right] f(e_j, x_o|q_j = \gamma_0+) de_j du_x \\
&= 1 + \tfrac{v_-}{n} f_q(\gamma_0) 2^{d-1}\int\int\left[\exp\left\{-\sqrt{-1}s^- e_j K(u_x) k_-(0)\right\} - 1\right] \tfrac{1(K(u_x)>0)}{\mathrm{Vol}(K(u_x)>0)} f(e_j, x_o|q_j = \gamma_0-) de_j du_x \\
&\quad + \tfrac{v_+}{n} f_q(\gamma_0) 2^{d-1}\int\int\left[\exp\left\{\sqrt{-1}s^+ e_j K(u_x) k_+(0)\right\} - 1\right] \tfrac{1(K(u_x)>0)}{\mathrm{Vol}(K(u_x)>0)} f(e_j, x_o|q_j = \gamma_0+) de_j du_x \\
&= 1 + \tfrac{v_-}{n} 2^{d-1} f_q(\gamma_0) f_{x|q}(x_o|\gamma_0)\mathbb{E}\left[\exp\left\{-\sqrt{-1}s^- e_j K(U_j^-) k_-(0)\right\} - 1\middle| x_j = x_o, q_j = \gamma_0-\right] \\
&\quad + \tfrac{v_+}{n} 2^{d-1} f_q(\gamma_0) f_{x|q}(x_o|\gamma_0)\mathbb{E}\left[\exp\left\{\sqrt{-1}s^+ e_j K(U_j^+) k_+(0)\right\} - 1\middle| x_j = x_o, q_j = \gamma_0+\right] \\
&= 1 + \tfrac{v_-}{n} 2^{d-1} f(x_o, \gamma_0)\mathbb{E}\left[\exp\left\{-\sqrt{-1}s^- e_j K(U_j^-) k_-(0)\right\} - 1\middle| x_j = x_o, q_j = \gamma_0-\right] \\
&\quad + \tfrac{v_+}{n} 2^{d-1} f(x_o, \gamma_0)\mathbb{E}\left[\exp\left\{\sqrt{-1}s^+ e_j K(U_j^+) k_+(0)\right\} - 1\middle| x_j = x_o, q_j = \gamma_0+\right],
\end{aligned}$$

where $\mathrm{Vol}(K(u_x) > 0) = 2^{d-1}$ is the volume of the area of u_x such that $K(u_x) > 0$, and U_j^- and U_j^+ are independent of $(e_j, x_j', q_j)'$ and follow a uniform distribution on the support of $K(\cdot)$. It follows that

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{\sqrt{-1}\left(s^-\sum_{j=1}^n T_{5j}^- + s^+\sum_{j=1}^n T_{5j}^+\right)\right\}\right] \\
&= \prod_{j=1}^n \mathbb{E}\left[\exp\left\{\sqrt{-1}\left(s^- T_{5j}^- + s^+ T_{5j}^+\right)\right\}\right] \\
&\to \exp\left\{v_- 2^{d-1} f(x_o, \gamma_0)\mathbb{E}\left[\exp\left\{-\sqrt{-1}s^- e K(U^-) k_-(0)\right\} - 1\middle| x = x_o, q = \gamma_0-\right]\right. \\
&\qquad \left. + v_+ 2^{d-1} f(x_o, \gamma_0)\mathbb{E}\left[\exp\left\{\sqrt{-1}s^+ e K(U^+) k_+(0)\right\} - 1\middle| x = x_o, q = \gamma_0+\right]\right\}.
\end{aligned}$$

This is the characteristic function of a compound Poisson process $D_5(\cdot)$ evaluated at v_- and v_+, where

$$D_5(v) = \begin{cases} \sum_{i=1}^{N_1(|v|)} \widetilde{z}_{1i}, & \text{if } v \le 0; \\ \sum_{i=1}^{N_2(v)} \widetilde{z}_{2i}, & \text{if } v > 0, \end{cases}$$

is a cadlag process with $D_5(0) = 0$, $\widetilde{z}_{1i} = -e_i^- K(U_i^-) k_-(0)$, $\widetilde{z}_{2i} = e_i^+ K(U_i^+) k_-(0)$, and $\left\{e_i^-, e_i^+, U_i^-, U_i^+\right\}_{i\ge 1}$, $N_1(\cdot)$ and $N_2(\cdot)$ are defined in Corollary 1. Generalizing this argument, we get the result of interest. ■

Lemma 6 $\sqrt{nh}\left(\frac{1}{n}\sum_{i=1}^n k_h\left(\frac{q_i-\gamma_0^C}{h}\right) - \frac{1}{n}\sum_{i=1}^n k_h\left(\frac{q_i-\gamma_0}{h}\right)\right) \xrightarrow{p} 0.$

Proof.

$$
\begin{aligned}
& \sqrt{nh}\left(\frac{1}{n}\sum_{i=1}^{n}k_h\left(\frac{q_i-\gamma_0^C}{h}\right)-\frac{1}{n}\sum_{i=1}^{n}k_h\left(\frac{q_i-\gamma_0}{h}\right)\right) \\
= & \frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\left[k\left(\frac{q_i-\gamma_0^C}{h}\right)-k\left(\frac{q_i-\gamma_0}{h}\right)\right] \\
\leq & \frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\frac{C}{nh}1(\gamma_0-h\leq q_i\leq\gamma_0^C+h)=O_p\left(\frac{C}{\sqrt{nh}}\right)=o_p(1),
\end{aligned}
$$

where the inequality is from the Lipschitz continuity of $k(\cdot)$. ■

Lemma 7 *Uniformly in x_i,*

$$
\begin{aligned}
\sqrt{nh}\left[\widehat{a}_+^C(x_i)-\widehat{a}_+^0(x_i)\right] &\xrightarrow{p} 0, \\
\sqrt{nh}h\left[\widehat{b}_+^C(x_i)-\widehat{b}_+^0(x_i)\right] &\xrightarrow{p} 0.
\end{aligned}
$$

Proof. Take the first result as an example. We have

$$
\begin{aligned}
& \sqrt{nh}\left[\widehat{a}_+^C(x_i)-\widehat{a}_+^0(x_i)\right] \\
= & \sqrt{nh}\left(\mathbf{e}_1'\left(M_{iC}^+\right)^{-1}r_{iC}^+-\mathbf{e}_1'\left(M_{i0}^+\right)^{-1}r_{i0}^+\right) \\
= & \sqrt{nh}\left[\mathbf{e}_1'\left(M_{i0}^+\right)^{-1}\left(r_{iC}^+-r_{i0}^+\right)-\mathbf{e}_1'\left(M_{i0}^+\right)^{-1}\left(M_{iC}^+-M_{i0}^+\right)\left(M_{i0}^+\right)^{-1}r_{i0}^+\right] \\
& +\sqrt{nh}\mathbf{e}_1'\left(M_{i0}^+\right)^{-1}\left(M_{iC}^+-M_{i0}^+\right)\left(M_{i0}^+\right)^{-1}\left(M_{iC}^+-M_{i0}^+\right)\left(M_{iC}^+\right)^{-1}r_{i0}^+ \\
& -\sqrt{nh}\mathbf{e}_1'\left(M_{i0}^+\right)^{-1}\left(M_{iC}^+-M_{i0}^+\right)\left(M_{iC}^+\right)^{-1}\left(r_{iC}^+-r_{i0}^+\right),
\end{aligned}
$$

where M_{iC}^+ and r_{iC}^+ are similarly defined as M_i^+ and r_i^+ but with $\widehat{\gamma}$ replaced by γ_0^C, and the decomposition in the last equality is from Lemma 2 of Yu (2010). Since $\left(M_{i0}^+\right)^{-1}$, $\left(M_{i0}^+\right)^{-1}$ and r_{i0}^+ are $O_p(1)$, we need only to show that

$$
\begin{aligned}
\sqrt{nh}\left(M_{iC}^+-M_{i0}^+\right) &\xrightarrow{p} 0 \text{ uniformly in } i, \\
\sqrt{nh}\left(r_{iC}^+-r_{i0}^+\right) &\xrightarrow{p} 0 \text{ uniformly in } i.
\end{aligned}
$$

Take the second result as an example.

$$
\begin{aligned}
& \sqrt{nh}\left(r_{iC}^+-r_{i0}^+\right)' \\
= & \frac{1}{\sqrt{nh}}\sum_{j=1,j\neq i}^{n}(x_j'-x_i',q_j-\gamma_0^C)^{S_p}K_h^x\left(x_j-x_i,x_i\right)k_+\left(\frac{q_j-\gamma_0^C}{h}\right)y_j \\
& -\frac{1}{\sqrt{nh}}\sum_{j=1,j\neq i}^{n}(x_j'-x_i',q_j-\gamma_0^C)^{S_p}K_h^x\left(x_j-x_i,x_i\right)k_+\left(\frac{q_j-\gamma_0}{h}\right)y_j.
\end{aligned}
$$

Take the following term of $\sqrt{nh}\left(r_{iC}^{+}-r_{i0}^{+}\right)$ as an example since it is the hardest to analyze.

$$
\begin{aligned}
& \left|\frac{1}{\sqrt{nh}}\sum_{j=1,j\neq i}^{n}\left(q_j-\gamma_0^C\right)^p K_h^x\left(x_j-x_i,x_i\right)k_+\left(\frac{q_j-\gamma_0^C}{h}\right)y_j\right. \\
& \left.-\frac{1}{\sqrt{nh}}\sum_{j=1,j\neq i}^{n}\left(q_j-\gamma_0\right)^p K_h^x\left(x_j-x_i,x_i\right)k_+\left(\frac{q_j-\gamma_0}{h}\right)y_j\right| \\
\leq & \left|\frac{1}{\sqrt{nh}}\sum_{j=1,j\neq i}^{n}K_h^x\left(x_j-x_i,x_i\right)\frac{C}{n}k_+\left(\frac{q_j-\gamma_0^C}{h}\right)y_j\right| \\
& +\left|\frac{1}{\sqrt{nh}}\sum_{j=1,j\neq i}^{n}K_h^x\left(x_j-x_i,x_i\right)\left(q_j-\gamma_0\right)^p 1(\gamma_0-h\leq q_j\leq\gamma_0^C+h)\frac{C}{nh}y_j\right|.
\end{aligned}
$$

From Lemma B.1 of Newey (1994), both terms on the right side converge to their expectations uniformly in i, but it is easy to see that these expectations are $O\left(h/\sqrt{nh}\right)=o(1)$. The results of interest follow. ■

Lemma 8 $\left|\mathbb{E}_1\left[\left(m_2-m_1\right)L_{b,21}\right]\right|=O_p\left(b^{\eta}\right)$.

Proof.

$$
\begin{aligned}
& \left|\mathbb{E}\left[\left(m(x_2,q_2)-m(x_1,q_1)\right)L_{b,21}|x_1,q_1\right]\right| \\
= & \left|\int\left(m(x_2,q_2)-m(x_1,q_1)\right)f(x_2,q_2)\frac{1}{b^d}L^x\left(\frac{x_2-x_1}{b},x_1\right)l\left(\frac{q_2-q_1}{b}\right)dx_2dq_2\right| \\
= & \left|\begin{array}{c}\int\left(Q_m\left(\left(x_2,q_2\right),\left(x_1,q_1\right)\right)+R_m\left(\left(x_2,q_2\right),\left(x_1,q_1\right)\right)\right)\left(f(x_1,q_1)+Q_f\left(\left(x_2,q_2\right),\left(x_1,q_1\right)\right)+R_f\left(\left(x_2,q_2\right),\left(x_1,q_1\right)\right)\right) \\ \frac{1}{b^d}L^x\left(\frac{x_2-x_1}{b},x_1\right)l\left(\frac{q_2-q_1}{b}\right)dx_2dq_2\end{array}\right|,
\end{aligned}
$$

where $Q_m\left(\left(x_2,q_2\right),\left(x_1,q_1\right)\right)$ is the $(s-1)$th-order Taylor expansion of $m(x_2,q_2)$ at $m(x_1,q_1)$, $R_m\left(\left(x_2,q_2\right),\left(x_1,q_1\right)\right)$ is the remainder term, $Q_f\left(\left(x_2,q_2\right),\left(x_1,q_1\right)\right)$ is $(\lambda-1)$th-order Taylor expansion of $f(x_2,q_2)$ at $f(x_1,q_1)$, and $R_f\left(\left(x_2,q_2\right),\left(x_1,q_1\right)\right)$ is the remainder term. From Assumption L,

$$
\int Q_m\left(\left(x_2,q_2\right),\left(x_1,q_1\right)\right)\left(f(x_1,q_1)+Q_f\left(\left(x_2,q_2\right),\left(x_1,q_1\right)\right)\right)\frac{1}{b^d}L^x\left(\frac{x_2-x_1}{b},x_1\right)l\left(\frac{q_2-q_1}{b}\right)dx_2dq=0,
$$

so $\left|\mathbb{E}\left[\left(m(x_2,q_2)-m(x_1,q_1)\right)L_{b,21}|x_1\right]\right|$ is bounded by

$$
\begin{aligned}
& \left|\int R_m\left(\left(x_2,q_2\right),\left(x_1,q_1\right)\right)f(x_1,q_1)\frac{1}{b^d}L^x\left(\frac{x_2-x_1}{b},x_1\right)l\left(\frac{q_2-q_1}{b}\right)dx_2dq_2\right| \\
& +\left|\int\left(m(x_2,q_2)-m(x_1,q_1)\right)R_f\left(\left(x_2,q_2\right),\left(x_1,q_1\right)\right)\frac{1}{b^d}L^x\left(\frac{x_2-x_1}{b},x_1\right)l\left(\frac{q_2-q_1}{b}\right)dx_2dq_2\right| \\
\leq & \ Cb^s+Cb^{\lambda+1}\leq Cb^{\eta},
\end{aligned}
$$

where $\eta=\min\left(\lambda+1,s\right)$. ■

Supplementary Materials

This supplement to the paper discusses the asymptotic distribution of the Wald-type and score-type test statistics under the null and local alternatives when instruments are available. We also provide implementation details for the use of Hansen's (1996) simulation method in the present context.

We first collect notation for future reference. Define the $n \times 1$ vectors Y and $\boldsymbol{\varepsilon}$ by stacking the variables y_i and ε_i, the $n \times (d+1)$ matrices X, $X_{\leq\gamma}$ and $X_{>\gamma}$ by stacking the vectors $\mathbf{x}_i'$, $\mathbf{x}_i' 1(q_i \leq \gamma)$ and $\mathbf{x}_i' 1(q_i > \gamma)$, and the $n \times d_z$ matrices Z, $Z_{\leq\gamma}$ and $Z_{>\gamma}$ are similarly defined. We use $\Rightarrow$ to signify weak convergence of a stochastic process on $\gamma \in \Gamma$. Define further that

$$\begin{aligned}
\Omega_1(\gamma) &= \mathbb{E}\left[z_i z_i' \varepsilon_i^2 1(q_i \leq \gamma)\right], \Omega_1(\gamma) = \mathbb{E}\left[z_i z_i' \varepsilon_i^2 1(q_i > \gamma)\right], \\
Q_1(\gamma) &= \mathbb{E}\left[z_i \mathbf{x}_i' 1(q_i \leq \gamma)\right], Q_2(\gamma) = \mathbb{E}\left[z_i \mathbf{x}_i' 1(q_i > \gamma)\right], \\
V_1(\gamma) &= \left[Q_1(\gamma)' \Omega_1(\gamma)^{-1} Q_1(\gamma)\right]^{-1}, V_2(\gamma) = \left[Q_2(\gamma)' \Omega_2(\gamma)^{-1} Q_2(\gamma)\right]^{-1}, \\
\Omega &= \mathbb{E}\left[z_i z_i' \varepsilon_i^2\right], Q = \mathbb{E}\left[z_i \mathbf{x}_i'\right], V = \left[Q' \Omega^{-1} Q\right]^{-1}.
\end{aligned}$$

$S_1(\gamma)$ is a mean-zero Gaussian process with covariance kernel $\mathbb{E}[S_1(\gamma_1) S_1(\gamma_2)'] = \Omega_1(\gamma_1 \wedge \gamma_2)$, $S = \lim_{\gamma \to \infty} S_1(\gamma)$, and $S_2(\gamma) = S - S_1(\gamma)$.

Given the threshold point γ, the 2SLS estimators for β_1 and β_2 are

$$\begin{aligned}
\widetilde{\beta}_1(\gamma) &= \left(\widehat{Q}_1(\gamma)' \left(\frac{1}{n} Z_{\leq\gamma}' Z_{\leq\gamma}\right)^{-1} \widehat{Q}_1(\gamma)\right)^{-1} \left(\widehat{Q}_1(\gamma)' \left(\frac{1}{n} Z_{\leq\gamma}' Z_{\leq\gamma}\right)^{-1} \frac{1}{n} Z_{\leq\gamma}' Y\right), \\
\widetilde{\beta}_2(\gamma) &= \left(\widehat{Q}_2(\gamma)' \left(\frac{1}{n} Z_{>\gamma}' Z_{>\gamma}\right)^{-1} \widehat{Q}_2(\gamma)\right)^{-1} \left(\widehat{Q}_2(\gamma)' \left(\frac{1}{n} Z_{>\gamma}' Z_{>\gamma}\right)^{-1} \frac{1}{n} Z_{>\gamma}' Y\right),
\end{aligned}$$

where $\widehat{Q}_1(\gamma) = n^{-1} \sum_{i=1}^n z_i \mathbf{x}_i' 1(q_i \leq \gamma)$ and $\widehat{Q}_2(\gamma) = n^{-1} \sum_{i=1}^n z_i \mathbf{x}_i' 1(q_i > \gamma)$. The residual from this equation is

$$\widetilde{\varepsilon}_i(\gamma) = y_i - \mathbf{x}_i' \widetilde{\beta}_1(\gamma) 1(q_i \leq \gamma) - \mathbf{x}_i' \widetilde{\beta}_2(\gamma) 1(q_i > \gamma).$$

The weight matrices

$$\widetilde{\Omega}_1(\gamma) = \frac{1}{n} \sum_{i=1}^n z_i z_i' \widetilde{\varepsilon}_i^2(\gamma) 1(q_i \leq \gamma), \ \widetilde{\Omega}_2(\gamma) = \frac{1}{n} \sum_{i=1}^n z_i z_i' \widetilde{\varepsilon}_i^2(\gamma) 1(q_i > \gamma).$$

The GMM estimators for β_1 and β_2 are

$$\begin{aligned}
\widehat{\beta}_1(\gamma) &= \left(\widehat{Q}_1(\gamma)' \widetilde{\Omega}_1^{-1}(\gamma) \widehat{Q}_1(\gamma)\right)^{-1} \left(\widehat{Q}_1(\gamma)' \widetilde{\Omega}_1^{-1}(\gamma) \frac{1}{n} Z_{\leq\gamma}' Y\right), \\
\widehat{\beta}_2(\gamma) &= \left(\widehat{Q}_2(\gamma)' \widetilde{\Omega}_2^{-1}(\gamma) \widehat{Q}_2(\gamma)\right)^{-1} \left(\widehat{Q}_2(\gamma)' \widetilde{\Omega}_2^{-1}(\gamma) \frac{1}{n} Z_{>\gamma}' Y\right).
\end{aligned}$$

The estimated covariance matrices for the GMM estimators are

$$\widehat{V}_1(\gamma) = \left(\widehat{Q}_1(\gamma)' \widetilde{\Omega}_1^{-1}(\gamma) \widehat{Q}_1(\gamma)\right)^{-1}, \ \widehat{V}_2(\gamma) = \left(\widehat{Q}_2(\gamma)' \widetilde{\Omega}_2^{-1}(\gamma) \widehat{Q}_2(\gamma)\right)^{-1}.$$

When H_0 holds, $\delta = 0$, and then the 2SLS estimator for β is

$$\widetilde{\beta} = \left(\widehat{Q}' \left(\frac{1}{n} Z'Z \right)^{-1} \widehat{Q} \right)^{-1} \left(\widehat{Q}' \left(\frac{1}{n} Z'Z \right)^{-1} \frac{1}{n} Z'Y \right),$$

where $\widehat{Q} = n^{-1} \sum_{i=1}^{n} z_i \mathbf{x}_i'$. Note here that the underlying assumption in this specification testing context is $\mathbb{E}[\varepsilon|z] = 0$, so that the 2SLS estimator can be applied. Correspondingly, the weight matrix is

$$\widetilde{\Omega} = \frac{1}{n} \sum_{i=1}^{n} z_i z_i' \widetilde{\varepsilon}_i^2$$

with $\widetilde{\varepsilon}_i = y_i - \mathbf{x}_i' \widetilde{\beta}$, the GMM estimators for β is

$$\widehat{\beta} = \left(\widehat{Q}' \widetilde{\Omega}^{-1} \widehat{Q} \right)^{-1} \left(\widehat{Q}' \widetilde{\Omega}^{-1} \frac{1}{n} Z'Y \right),$$

and the residual is

$$\widehat{\varepsilon}_i = y_i - \mathbf{x}_i' \widehat{\beta}.$$

The estimated covariance matrix for the GMM estimator is

$$\widehat{V} = \left(\widehat{Q}' \widetilde{\Omega}^{-1} \widehat{Q} \right)^{-1}.$$

1. Wald-type Tests

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

The test statistic is a functional of $W_n(\cdot)$. Two test statistics are most popular. The first is the Kolmogorov-Smirnov sup-type statistic

$$K_n^{\omega} = \sup_{\gamma \in \Gamma} \|W_n(\gamma)\|,$$

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

$$C_n^{\omega} = \int_{\Gamma} \|W_n(\gamma)\| \, w(\gamma) d\gamma,$$

where $w(\gamma)$ in C_n^{ω} is a known positive weight function with $\int_{\Gamma} w(\gamma) d\gamma = 1$. For example, $w(\tau) = 1/|\Gamma|$ with $|\Gamma|$ being the length of Γ; if we have some information on the locations where threshold effects are most likely to occur, we can impose larger weights on the neighborhoods of such locations. The choice of the norm $\|\cdot\|$ is also an issue. The Euclidean norm $\|\cdot\|_2$ is obviously natural, e.g., Caner and Hansen (2004) use (the square of) this norm. Yu (2013b) suggest using the ℓ_1 norm in testing quantile threshold effects, and Bai (1996) suggests using the ℓ_∞ norm in structural change tests.

The following theorem states the asymptotic distribution of a general continuous functional $g(\cdot)$ of $W_n(\cdot)$ under the local alternative $\delta_n = n^{-1/2} c$. The corresponding test statistic is denoted as g_n^{ω}.

Theorem 9 *If $\delta_n = n^{-1/2} c$, $\mathbb{E}\left[\|x\|^4\right] < \infty$, $\mathbb{E}[q^4] < \infty$, $\mathbb{E}[\varepsilon^4]$ and $\mathbb{E}[\|z\|^4] < \infty$, then*

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

where

$$
\begin{aligned}
W^c(\gamma) &= (V_1(\gamma)+V_2(\gamma))^{-1/2}\left[V_1(\gamma)\,Q_1(\gamma)'\,\Omega_1(\gamma)^{-1}\,Q_1(\gamma\wedge\gamma_0)-V_2(\gamma)\,Q_2(\gamma)'\,\Omega_2(\gamma)^{-1}\,Q_2(\gamma\vee\gamma_0)\right]c\\
&\quad+(V_1(\gamma)+V_2(\gamma))^{-1/2}\left[V_1(\gamma)\,Q_1(\gamma)'\,\Omega_1(\gamma)^{-1}\,S_1(\gamma)-V_2(\gamma)\,Q_2(\gamma)'\,\Omega_2(\gamma)^{-1}\,S_2(\gamma)\right].
\end{aligned}
$$

Proof. Under the local alternative $\delta_n = n^{-1/2}c$, $Y = X_{\le\gamma_0}(\beta+\delta_n) + X_{>\gamma_0}\beta + \boldsymbol{\varepsilon} = X\beta + X_{\le\gamma_0}\delta_n + \boldsymbol{\varepsilon}$, so that

$$
\begin{aligned}
\widetilde{\beta}_1(\gamma) &= \left(X'_{\le\gamma}Z_{\le\gamma}\left(Z'_{\le\gamma}Z_{\le\gamma}\right)^{-1}Z'_{\le\gamma}X_{\le\gamma}\right)^{-1}\left(X'_{\le\gamma}Z_{\le\gamma}\left(Z'_{\le\gamma}Z_{\le\gamma}\right)^{-1}Z'_{\le\gamma}Y\right)\\
&= \beta + O_p(1)\frac{1}{n}\sum\nolimits_{i=1}^{n} z_i\left[x_i'\delta_n 1\left(q_i\le\gamma_0\wedge\gamma\right)+\varepsilon_i 1\left(q_i\le\gamma\right)\right]\\
&= \beta + o_p(1) \text{ uniformly in } \gamma\in\Gamma.
\end{aligned}
$$

Similarly, $\widetilde{\beta}_2(\gamma)$ is uniformly consistent to β. As a result,

$$
\begin{aligned}
\widetilde{\varepsilon}_i(\gamma) &= y_i - \mathbf{x}_i'\widetilde{\beta}_1(\gamma)\,1(q_i\le\gamma) - \mathbf{x}_i'\widetilde{\beta}_2(\gamma)\,1(q_i>\gamma)\\
&= \mathbf{x}_i'\beta + \mathbf{x}_i'\delta_n 1(q_i\le\gamma_0) + \varepsilon_i - \mathbf{x}_i'\left(\beta + o_p(1)\right)\\
&= \varepsilon_i + o_p\left(\|\mathbf{x}_i\|\right) \text{ uniformly in } \gamma\in\Gamma,
\end{aligned}
$$

so that

$$
\begin{aligned}
\widetilde{\Omega}_1(\gamma) &= \frac{1}{n}\sum_{i=1}^{n} z_i z_i'\widetilde{\varepsilon}_i^2(\gamma)\,1(q_i\le\gamma)\\
&= \frac{1}{n}\sum_{i=1}^{n} z_i z_i'\left(\varepsilon_i + o_p\left(\|\mathbf{x}_i\|\right)\right)^2 1(q_i\le\gamma) \xrightarrow{p} \Omega_1(\gamma)
\end{aligned}
$$

uniformly in $\gamma\in\Gamma$ by a standard argument. Similarly, $\widetilde{\Omega}_2(\gamma) \xrightarrow{p} \Omega_2(\gamma)$ uniformly in $\gamma\in\Gamma$. Now,

$$
\sqrt{n}\left(\widehat{\beta}_1(\gamma)-\beta\right) = \left[\widehat{Q}_1(\gamma)'\,\widetilde{\Omega}_1(\gamma)^{-1}\,\widehat{Q}_1(\gamma)\right]^{-1}\left[\widehat{Q}_1(\gamma)'\,\widetilde{\Omega}_1(\gamma)^{-1}\frac{1}{\sqrt{n}}Z'_{\le\gamma}\left(X_{\le\gamma_0}\delta_n+\boldsymbol{\varepsilon}\right)\right],
$$

where $\widehat{Q}_1(\gamma) \xrightarrow{p} Q_1(\gamma)$, $\frac{1}{\sqrt{n}}Z'_{\le\gamma}X_{\le\gamma_0}\delta_n \xrightarrow{p} Q_1(\gamma\wedge\gamma_0)\,c$ uniformly in $\gamma\in\Gamma$, and $\frac{1}{\sqrt{n}}Z'_{\le\gamma}\boldsymbol{\varepsilon} \Rightarrow S_1(\gamma)$. Hence

$$
\sqrt{n}\left(\widehat{\beta}_1(\gamma)-\beta\right) \Rightarrow V_1(\gamma)\,Q_1(\gamma)'\,\Omega_1(\gamma)^{-1}\left[Q_1(\gamma\wedge\gamma_0)\,c + S_1(\gamma)\right].
$$

Similarly,

$$
\sqrt{n}\left(\widehat{\beta}_2(\gamma)-\beta\right) \Rightarrow V_2(\gamma)\,Q_2(\gamma)\,\Omega_2(\gamma)^{-1}\left[Q_2(\gamma\vee\gamma_0)\,c + S_2(\gamma)\right].
$$

From the arguments above and by the continuous mapping theorem, $\widehat{V}_1(\gamma) \xrightarrow{p} V_1(\gamma)$ and $\widehat{V}_2(\gamma) \xrightarrow{p} V_2(\gamma)$ uniformly in $\gamma\in\Gamma$. Finally, $W_n(\gamma) \Rightarrow W^c(\gamma)$ as specified in the theorem, where the second part of $W^c(\gamma)$ is the process in Theorem 4 of Caner and Hansen (2004). ■

2. Score-type Tests

The score-type tests are based on

$$T_n(\gamma) = \left[n^{-1} \sum_{i=1}^{n} \left(z_i 1(q_i \le \gamma) - \widehat{Q}_1(\gamma) \widehat{V} \widehat{Q}' \widetilde{\Omega}^{-1} z_i \right) \left(z_i 1(q_i \le \gamma) - \widehat{Q}_1(\gamma) \widehat{V} \widehat{Q}' \widetilde{\Omega}^{-1} z_i \right)' \widehat{\varepsilon}_i^2 \right]^{-1/2}$$
$$\cdot n^{-1/2} \sum_{i=1}^{n} \left[z_i 1(q_i \le \gamma) - \widehat{Q}_1(\gamma) \widehat{V} \widehat{Q}' \widetilde{\Omega}^{-1} z_i \right] \widehat{\varepsilon}_i, \gamma \in \Gamma.$$

Note here that although $\widehat{Q}_1(\gamma) \widehat{V} \widehat{Q}' \widetilde{\Omega}^{-1} n^{-1/2} \sum_{i=1}^{n} z_i \widehat{\varepsilon}_i = o_p(1)$, $z_i 1(q_i \le \gamma)$ is recentered by $\widehat{Q}_1(\gamma) \widehat{V} \widehat{Q}' \widetilde{\Omega}^{-1} z_i$. This is because the effect of $\widehat{\beta}$ will not disappear asymptotically so the asymptotic distribution of $n^{-1/2} \sum_{i=1}^{n} z_i 1(q_i \le \gamma) \widehat{\varepsilon}_i$ differs from $n^{-1/2} \sum_{i=1}^{n} z_i 1(q_i \le \gamma) \varepsilon_i$ under H_0. Recentering is to offset the effect of $\widehat{\beta}$. Since only $\widehat{\beta}$ is used in the construction of $T_n(\cdot)$, this type of tests is constructed under H_0 and only one GMM estimator needs to be constructed. This significantly lightens the computation burden. Given $T_n(\cdot)$, we can similarly construct the Kolmogorov-Smirnov sup-type statistic K_n^s and the Cramér–von Mises average-type statistic C_n^s.

The following theorem states the asymptotic distribution of a general continuous functional $g(\cdot)$ of $T_n(\cdot)$ under the local alternative $\delta_n = n^{-1/2} c$. The corresponding test statistic is denoted as g_n^s.

Theorem 10 *If* $\delta_n = n^{-1/2} c$, $\mathbb{E}\left[\|x\|^4 \right] < \infty$, $\mathbb{E}[q^4] < \infty$, $\mathbb{E}[\varepsilon^4]$ *and* $\mathbb{E}[\|z\|^4] < \infty$, *then*

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

where

$$T^c(\gamma) = H(\gamma, \gamma)^{-1/2} \left\{ S(\gamma) + \left[Q_1(\gamma \wedge \gamma_0) - Q_1(\gamma) V Q' \Omega^{-1} Q_1(\gamma_0) \right] c \right\},$$

with $S(\gamma)$ *being a mean zero Gaussian process with covariance kernel*

$$H(\gamma_1, \gamma_2) = \mathbb{E}\left[\left(z_i 1(q_i \le \gamma_1) - Q_1(\gamma_1) V Q' \Omega^{-1} z_i \right) \left(z_i 1(q_i \le \gamma_2) - Q_1(\gamma_2) V Q' \Omega^{-1} z_i \right)' \varepsilon_i^2 \right].$$

Proof. As in the last theorem, we can show $\widehat{\beta} \xrightarrow{p} \beta$, $\widetilde{\Omega} \xrightarrow{p} \Omega$, and $\widehat{V} \xrightarrow{p} V$ under the local alternative.

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

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

$n^{-1/2}\sum_{i=1}^{n} z_i 1(q_i \leq \gamma)\varepsilon_i \Rightarrow S_1(\gamma)$. Next,

$$
\begin{aligned}
& n^{-1/2}\sum_{i=1}^{n} \widehat{Q}_1(\gamma)\widehat{V}\widehat{Q}'\widetilde{\Omega}^{-1}z_i\widehat{\varepsilon}_i \\
= \; & \widehat{Q}_1(\gamma)\left(\widehat{Q}'\widetilde{\Omega}^{-1}\widehat{Q}\right)^{-1}\widehat{Q}'\widetilde{\Omega}^{-1}n^{-1/2}\sum_{i=1}^{n} z_i\left(-\mathbf{x}_i'(\widehat{\beta}-\beta)+\mathbf{x}_i'\delta_n 1(q_i\leq\gamma_0)+\varepsilon_i\right) \\
= \; & -\widehat{Q}_1(\gamma)\sqrt{n}\left(\widehat{\beta}-\beta\right)+\widehat{Q}_1(\gamma)\widehat{V}\widehat{Q}'\widetilde{\Omega}^{-1}\left(n^{-1}\sum_{i=1}^{n} z_i\mathbf{x}_i'1(q_i\leq\gamma_0)\right)c \\
& +\widehat{Q}_1(\gamma)\widehat{V}\widehat{Q}'\widetilde{\Omega}^{-1}\left(n^{-1/2}\sum_{i=1}^{n} z_i\varepsilon_i\right),
\end{aligned}
$$

where the second term in the last equality converges in probability to $Q_1(\gamma)VQ'\Omega^{-1}Q_1(\gamma_0)c$ uniformly in $\gamma\in\Gamma$, and $n^{-1/2}\sum_{i=1}^{n} z_i\varepsilon_i \xrightarrow{d} N(0,\Omega)$. In summary,

$$
\begin{aligned}
& n^{-1/2}\sum_{i=1}^{n}\left[z_i 1(q_i\leq\gamma)-\widehat{Q}_1(\gamma)\widehat{V}\widehat{Q}'\widetilde{\Omega}^{-1}z_i\right]\widehat{\varepsilon}_i \\
= \; & n^{-1/2}\sum_{i=1}^{n}\left[z_i 1(q_i\leq\gamma)-\widehat{Q}_1(\gamma)\widehat{V}\widehat{Q}'\widetilde{\Omega}^{-1}z_i\right]\varepsilon_i \\
& +\left[Q_1(\gamma\wedge\gamma_0)-Q_1(\gamma)VQ'\Omega^{-1}Q_1(\gamma_0)\right]c+o_p(1) \\
\Rightarrow \; & S(\gamma)+\left[Q_1(\gamma\wedge\gamma_0)-Q_1(\gamma)VQ'\Omega^{-1}Q_1(\gamma_0)\right]c,
\end{aligned}
$$

and it is not hard to show $n^{-1}\sum_{i=1}^{n}\left(z_i 1(q_i\leq\gamma)-\widehat{Q}_1(\gamma)\widehat{V}\widehat{Q}'\widetilde{\Omega}^{-1}z_i\right)\left(z_i 1(q_i\leq\gamma)-\widehat{Q}_1(\gamma)\widehat{V}\widehat{Q}'\widetilde{\Omega}^{-1}z_i\right)'\widehat{\varepsilon}_i^2 \xrightarrow{p} H(\gamma,\gamma)$ uniformly in $\gamma\in\Gamma$, so the results of the theorem follow. ■

To understand $S(\gamma)$, consider a simple case where $\mathbf{x}=(1,x')'$, q follows a uniform distribution on $[0,1]$ and is independent of $(z',x',e)'$. In this case,

$$
H(\gamma_1,\gamma_2)=(\gamma_1\wedge\gamma_2)\,\Omega-\gamma_1\gamma_2 QVQ'.
$$

If $d_z=d+1$, i.e., the model is just-identified, then

$$
\begin{aligned}
H(\gamma_1,\gamma_2) &= \mathbb{E}\left[\left(z_i 1(q_i\leq\gamma_1)-Q_1(\gamma_1)Q^{-1}z_i\right)\left(z_i 1(q_i\leq\gamma_2)-Q_1(\gamma_2)Q^{-1}z_i\right)'\varepsilon_i^2\right] \\
&= \Omega_1(\gamma_1\wedge\gamma_2)-Q_1(\gamma_1)Q^{-1}\Omega_1(\gamma_2)-\Omega_1(\gamma_1)Q'^{-1}Q_1(\gamma_2)'+Q_1(\gamma_1)Q^{-1}\Omega Q'^{-1}Q_1(\gamma_2)',
\end{aligned}
$$

and we can let

$$
\begin{aligned}
T_n(\gamma) = & \left[n^{-1}\sum_{i=1}^{n}\left(z_i 1(q_i\leq\gamma)-\widehat{Q}_1(\gamma)\widehat{Q}^{-1}z_i\right)\left(z_i 1(q_i\leq\gamma)-\widehat{Q}_1(\gamma)\widehat{Q}^{-1}z_i\right)'\widehat{\varepsilon}_i^2\right]^{-1/2} \\
& \cdot n^{-1/2}\sum_{i=1}^{n}\left[z_i 1(q_i\leq\gamma)-\widehat{Q}_1(\gamma)\widehat{Q}^{-1}z_i\right]\widehat{\varepsilon}_i,\gamma\in\Gamma.
\end{aligned}
\tag{24}
$$

Combining these two cases,[39] $H(\gamma_1,\gamma_2)$ reduces to $(\gamma_1\wedge\gamma_2-\gamma_1\gamma_2)\,\Omega$. In other words, $\Omega^{-1/2}S(\gamma)$ is a standard d-dimensional Brownian Bridge. Now, the local power is generated by $\left[Q_1(\gamma\wedge\gamma_0)-Q_1(\gamma)VQ'\Omega^{-1}Q_1(\gamma_0)\right]c=(\gamma\wedge\gamma_0-\gamma\gamma_0)\,Qc$. Of course, the construction of $T_n(\gamma)$ can be greatly simplified in this simple case, e.g., let

$$
T_n(\gamma)=\widetilde{\Omega}^{-1/2}\cdot n^{-1/2}\sum_{i=1}^{n}\left[z_i 1(q_i\leq\gamma)-\gamma z_i\right]\widehat{\varepsilon}_i
$$

[39] Now, $d_z=d$.

which converges to the standard d-dimensional Brownian Bridge. In linear regression, we need only replace z_i in all formula of (24) by $\mathbf{x}_i$.

3. Simulating the Critical Values

The asymptotic distributions in the above two theorems are nonpivotal, but the simulation method in Hansen (1996) can be extended to the present case. More specifically, let $\{\xi_i^*\}_{i=1}^n$ be i.i.d. $N(0,1)$ random variables, and set

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

and

$$T_n^*(\gamma) = \left[n^{-1}\sum_{i=1}^n \left(z_i 1(q_i \le \gamma) - \widehat{Q}_1(\gamma)\widehat{V}\widehat{Q}'\widetilde{\Omega}^{-1}z_i\right)\left(z_i 1(q_i \le \gamma) - \widehat{Q}_1(\gamma)\widehat{V}\widehat{Q}'\widetilde{\Omega}^{-1}z_i\right)'\widehat{\varepsilon}_i^2\right]^{-1/2} \cdot n^{-1/2}\sum_{i=1}^n \left[z_i 1(q_i \le \gamma) - \widehat{Q}_1(\gamma)\widehat{V}\widehat{Q}'\widetilde{\Omega}^{-1}z_i\right]\widehat{\varepsilon}_i\xi_i^*, \gamma \in \Gamma, \tag{25}$$

where $\widehat{\beta}_1^*(\gamma)$ and $\widehat{\beta}_2^*(\gamma)$ are similarly defined as $\widehat{\beta}_1(\gamma)$ and $\widehat{\beta}_2(\gamma)$ with the only difference being that y_i is replaced by $\widetilde{\varepsilon}_i(\gamma)\xi_i^*$; more specifically,

$$\begin{aligned}
\widehat{\beta}_1^*(\gamma) &= \left(\widehat{Q}_1(\gamma)'\widetilde{\Omega}_1^{-1}(\gamma)\widehat{Q}_1(\gamma)\right)^{-1}\left(\widehat{Q}_1(\gamma)'\widetilde{\Omega}_1^{-1}(\gamma)\frac{1}{n}\sum\nolimits_{i=1}^n z_i 1(q_i \le \gamma)\widetilde{\varepsilon}_i(\gamma)\xi_i^*\right), \\
\widehat{\beta}_2^*(\gamma) &= \left(\widehat{Q}_2(\gamma)'\widetilde{\Omega}_2^{-1}(\gamma)\widehat{Q}_2(\gamma)\right)^{-1}\left(\widehat{Q}_2(\gamma)'\widetilde{\Omega}_2^{-1}(\gamma)\frac{1}{n}\sum\nolimits_{i=1}^n z_i 1(q_i > \gamma)\widetilde{\varepsilon}_i(\gamma)\xi_i^*\right).
\end{aligned}$$

Our test rejects H_0 if g_n^ω (g_n^s) is greater than the $(1-\alpha)$th conditional quantile of $g(W_n^*(\gamma))$ $(g(T_n^*(\gamma)))$. Equivalently, the p-value transformation can be employed. Take the score test as an example. Define $p_n^* = 1 - F_n^*(g_n^s)$, and $p_n = 1 - F_0(g_n^s)$, where F_n^* is the conditional distribution of $g(T_n^*(\gamma))$ given the original data, and F_0 is the asymptotic distribution of $g(T_n(\gamma))$ under the null. Our test rejects H_0 if $p_n^* \le \alpha$. By stochastic equicontinuity of the $T_n(\gamma)$ process, we can replace Γ by finite grids with the distance between adjacent grid points going to zero as $n \to \infty$. A natural choice of the grids for Γ is the q_i's in Γ. Also, the conditional distribution can be approximated by standard simulation techniques. More specifically, the following procedure is used.

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

Step 2: set $T_n^{j*}(\gamma_l)$ as in (25), where $\{\gamma_l\}_{l=1}^L$ is a grid approximation of Γ. Note here that the same $\{\xi_{ij}^*\}_{i=1}^n$ are used for all γ_l, $l = 1, \cdots, L$.

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

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

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

It can be shown that $p_n^* = p_n + o_p(1)$ under both the null and local alternative. Hence $p_n^* \xrightarrow{d} p_c = 1 - F_0(g_c^s)$ under the local alternative, and $p_n^* \xrightarrow{d} U$, the uniform distribution on $[0,1]$, under the null. The proof is similar to that of Yu (2013b), so it is omitted here.