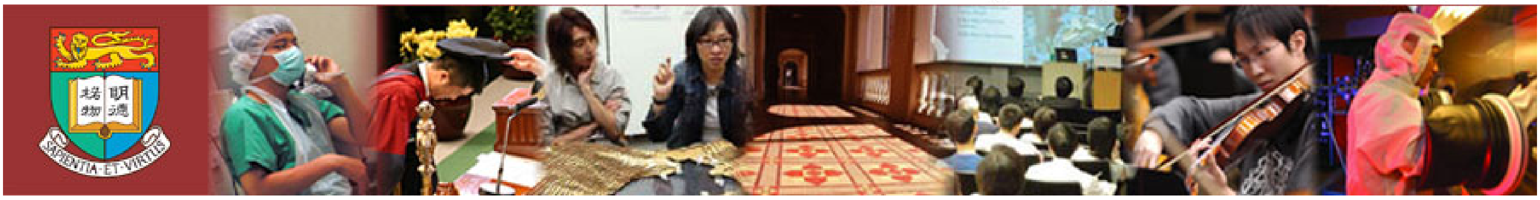

# Threshold Regression with Endogeneity[*]

Ping Yu[†]
University of Hong Kong

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

First Draft: October 2012
This Version: July 2015

**Abstract**

This paper studies estimation 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. 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, U-statistic, Threshold treatment model, 401(k) plan.

JEL-CLASSIFICATION: C21, C24, 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, Howell Tong, Andrew Weiss, Ka-fu Wong, Steven Pai Xu and seminar participants at ANU, HKU, NUS, Rutgers, UNSW, UoA and 2013NASM for helpful comments. Phillips acknowledges support from the NSF under Grant No. SES 1258258.

[†] School of Economics and Finance, The University of Hong Kong, Pokfulam Road, Hong Kong; corresponding author email: pingyu@hku.hk.

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

# 1 Introduction

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 $\gamma$, the coefficients $\beta_1$ and $\beta_2$ are the respective threshold parameters, and $\varepsilon$ 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 $\theta \equiv \left(\beta', \delta', \gamma\right)'$.

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 (2012) 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 $\theta$ 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 endogenous 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 on this topic. For instance, Caner and Hansen (2004) use the asymptotic framework of Hansen (2000), where $\delta$ 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 $\gamma$. 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 (2013a) shows that three typical 2SLS estimators of $\gamma$ are generally inconsistent. This finding motivates us to search for general consistent estimators of $\gamma$. One of the main contributions of the present paper is to show that when only $\gamma$ and $\delta$ are of interest, as in the typical case,[1] these parameters are both identified even without instruments. This result has meaningful significance 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 $\delta$ 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

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

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

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

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 $\gamma$ 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 $\gamma$ 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 $\gamma$, so that only the local information around $\gamma$ 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, 2015) 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 $\gamma$, we propose two estimators of $\delta$ that are suggested from the partial linear model structure of (2) that applies for known $\gamma$.[2] An important difference between (2) and the usual partial linear structure is that both parametric and nonparametric components of $m(x,q) = g(x,q) + \mathbf{x}'\delta 1\left(q \leq \gamma\right)$ 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 $\mathbf{x}'1\left(q \leq \gamma\right)$ is fully determined by $(x,q)$ given $\gamma$, 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), $Z = (x',q)$, $X = \mathbf{x}'1\left(q \leq \gamma\right)$, $\beta = \delta$ and $\theta\left(Z\right) = g(x,q)$.

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

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

for all $\eta(x,q)$, suggesting that the (partial linear) component $\mathbf{x}'\delta 1\left(q\leq\gamma\right)$ cannot be separated from $g(x,q)$ in the composite function $m\left(x,q\right)$. 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 $\delta$ in the discontinuity structure of the component $\mathbf{x}'\delta 1\left(q\leq\gamma\right)$ that arises from the jump in $m(x,q)$ at $q=\gamma$. It is this jump discontinuity that assures identification of the linear coefficients $\delta$.

Although the coefficient vector $\delta$ 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 $\mathbb{E}\left[\mathbf{x}'\delta 1\left(q\leq\gamma\right)|x,q\right]$ 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 $\delta$ 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 $\delta$ can be estimated at a $\sqrt{n}$ rate. In this case, for the linear endogenous threshold model (1), $\beta$ can also be estimated at a $\sqrt{n}$ rate. So the role of instruments in the model (1) is to provide identification for $\beta$ and to improve the convergence rate of estimates of $\delta$. As for the threshold parameter $\gamma$ in (1), our results show that $\gamma$ 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 $\gamma$ 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 $\beta$, $\delta$ and $\gamma$: only for $\beta$ do instruments have the conventional role of delivering identification, whereas for $\delta$ and $\gamma$ the presence of instruments serves to improve convergence rates or efficiency.

A brief simulation study is included to test the adequacy of the asymptotic theory of the estimation 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. 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 $\gamma$ and $\delta$ and derive their limit distributions. Section 3 investigates the role of instruments. Section 4 covers some extensions and simplifications of our analysis. Section 5 reports the results of some finite sample simulations. Section 6 presents an empirical application to explore the effects of 401(k) retirement programs on savings. Section 7 concludes. Proofs with supporting propositions and lemmas are given in Appendices A, B and C, respectively. A Supplement to the paper contains additional material, derivations, and proofs of subsidiary results.

A word on notation. The letter $C$ denotes a generic positive constant whose value may change in each occurrence. The parameters $\beta$ and $\delta$ 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 $\ell$ is used to indicate the two regimes in (1), and is not written out explicitly as '$\ell=1,2'$ except in Section 6 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 specified; 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 (IDKE)

This section introduces a new methodology for consistently estimating $\gamma$ and $\delta$ when instruments are absent. The method involves a nonparametric kernel estimator that we call the integrated difference kernel estimator (IDKE) A related estimator of $\gamma$ that is already in the literature is the difference kernel estimator (DKE) of Qiu et al. (1991) where $q$ is the only covariate. When there are additional covariates as in our setup, Delgado and Hidalgo (2000) suggested that the DKE continue to be used to estimate $\gamma$. In the supplementary materials, we explain some difficulties that arise in applying the DKE in the current case; we also explain the difficulties in applying other estimators such as the LSE and the partial linear estimator (PLE). In the following discussion, we concentrate on describing the construction of the IDKE, developing the limit theory for the IDKE and associated coefficient estimates, and providing an intuitive rationale for the identification and consistent estimation of $\gamma$ and $\delta$ without instruments.

## 2.1 Construction of the IDKE of $\gamma$

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

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

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

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

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

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

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

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

(iv) implies that $k_-(\cdot,r)$ is Lipschitz on $(-\infty,r]$ and $k_+(\cdot,r)$ is Lipschitz on $[-r,\infty)$. This assumption is important in deriving the asymptotic distribution of the IDKE of $\gamma$; see 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 $\infty$, and $\Gamma$ and $\Lambda$ are compact.

Since $\delta_0$ 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; a more explicit version of the non-zero restriction on $\delta_0$ is imposed in Assumption I of Section 2.2 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 tail of the distribution of $x$.

Define

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

and

$$k_h(u,t) = \begin{cases} \frac{1}{h} k\left(\frac{u}{h}\right), & \text{if } h \le t \le 1-h, \\ \frac{1}{h} k_+\left(\frac{u}{h}, \frac{t}{h}\right), & \text{if } 0 \le t \le h, \\ \frac{1}{h} k_-\left(\frac{u}{h}, \frac{1-t}{h}\right), & \text{if } 1-h \le t \le 1. \end{cases} . \tag{4}$$

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 $\gamma$.[3] Formally, we require

**Assumption K**: $k_h(u,t)$ takes the form of (4) 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 $\gamma$.

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

$$\begin{aligned} \widehat{\gamma} &= \arg\max_\gamma \frac{1}{n}\sum_{i=1}^n \left[ \frac{1}{n-1} \sum_{j=1, j\neq i}^n y_j K_{h,ij}^{\gamma-} - \frac{1}{n-1} \sum_{j=1, j\neq i}^n y_j K_{h,ij}^{\gamma+} \right]^2 \\ &\equiv \arg\max_\gamma \frac{1}{n}\sum_{i=1}^n \widehat{\Delta}_i^2\left(\gamma\right) \equiv \arg\max_\gamma \widehat{Q}_n\left(\gamma\right), \end{aligned} \tag{5}$$

where

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

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

with

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

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 (5) excludes $j = i$, which is a standard strategy in converting a V-statistic to a U-statistic. Also, the normalization factor $\sum_{j=1,j\neq i}^n K^{\gamma\pm}_{h,ij}$ does not appear in the construction of $\widehat{\gamma}$, thereby avoiding random denominator issues in conditional mean estimation and simplifying the derivation of the limit distribution of $\widehat{\gamma}$, a technique that dates back at least to Powell et al. (1989). This form of $\widehat{\gamma}$ has some practical advantages especially when $d$ is large. Since the conditional mean is estimated at the boundary point $q = \gamma$, the local linear smoother (LLS) or the local polynomial estimator (LPE) might be considered to ameliorate bias. However, when $d$ is large, there are not many data points in a $h$ neighborhood of $(x_i', \gamma)'$. 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. Furthermore, different from the regular parameter (such as the conditional mean) estimation, use of the LLS does not affect the first-order asymptotic distribution of $\widehat{\gamma}$.

The objective function in (5) may be viewed as a nonparametric extension of the objective function of the parametric LSE of $\gamma$. With some preliminary algebra, it can be shown that the parametric LSE of $\gamma$ satisfies

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

where $\widehat{\delta}$ is the LSE of $\delta$ based on the splitting of $\gamma$, 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 $\mathbf{x}_i'1(q_i > \gamma)$, respectively. The objective function of $\widehat{\gamma}^P_{LSE}$ uses the weighted average form of $\mathbf{X}\widehat{\delta}$ which is the conditional mean differences at all $\mathbf{x}_i$'s.[4] The weights in (5) are essentially given by $f(x_i, \gamma)$ (the probability limit of $n^{-1}\sum_{j=1,j\neq i}^n K^{\gamma\pm}_{h,ij}$), 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 $\gamma$.

## 2.2 Limit Theory for the IDKE

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

[4] 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 $\mathbf{X} = (x_1, \cdots, x_n)'$.

assumption is made for simplicity, just as in Robinson (1988), and is not critical to the methodology or the limit theory. The second part of Assumption F implies that $\gamma_0$ is not on the boundary of $\mathcal{Q}$. 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 $\gamma$, there will be a jump in the limit only if $\gamma=\gamma_0$ which provides identifying information. As a result, the threshold point can be identified and consistently estimated by maximizing $\widehat{Q}_n(\gamma)$. Given that $\mathbb{E}[y|x,q=\gamma_0+]-\mathbb{E}[y|x,q=\gamma_0-]=(1,x',\gamma_0)\,\delta_0$, we need the following assumption to identify $\gamma_0$.

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

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

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

Here, $\overline{z}_{1i}$ represents the effect on $\widehat{Q}_n(\gamma)$ when the threshold point is displaced on the left of $\gamma_0$, 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*

$$D\left(v\right)=\left\{\begin{array}{l}\sum\limits_{i=1}^{N_1(|v|)} z_{1i},\ \textit{if } v\leq 0,\\ \sum\limits_{i=1}^{N_2(v)} z_{2i},\ \textit{if } v>0,\end{array}\right.$$

*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 $N_\ell\left(\cdot\right)$ 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; see the supplementary materials for a brief summary. 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 $\mathbb{E}\left[y|x,q\right]$; see (5). 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 (5) 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 $\varepsilon_1$ and $\varepsilon_2$ 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

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

where $\widehat{\Delta}_o\left(\gamma\right)=\frac{1}{n}\sum\limits_{j=1}^{n}y_jK_{h,j}^{\gamma-}-\frac{1}{n}\sum\limits_{j=1}^{n}y_jK_{h,j}^{\gamma+}$ with

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

and where $x_o$ is some fixed point in the interior of $\mathcal{X}$. As explained in the supplementary materials, selection of $x_o$ is difficult from both theory and practical perspectives. As distinct from the DKE, the IDKE procedure integrates the jump information over all $x_i$'s, thereby removing the problem of choosing $x_o$. Further, use of all the data ensures that the IDKE has greater identifying capability than the DKE. For ease of expression in the following corollary, define $K\left(u_x\right)=\prod_{l=1}^{d-1}k(u_{x_l})$.

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

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

*where*

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

*is a cadlag process with* $D(0)=0$, $z_{1i}=\left[2\left(1,x_o',\gamma_0\right)\delta_0 e_i^- + \delta_0'\left(1,x_o',\gamma_0\right)'\left(1,x_o',\gamma_0\right)\delta_0\right]K(U_i^-)$ *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* $K(\cdot)$, $z_{2i}=\left[-2\left(1,x_o',\gamma_0\right)\delta_0 e_i^+ + \delta_0'\left(1,x_o',\gamma_0\right)'\left(1,x_o',\gamma_0\right)\delta_0\right]K(U_i^+)$ *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* $U_i^-$, $\left\{e_i^-,e_i^+,U_i^-,U_i^+\right\}_{i\geq 1}$, $N_1(\cdot)$ *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 $\gamma$. Nevertheless, the convergence rate is still faster than that of Delgado and Hidalgo (2000). In their setup in terms of the DKE, $k_+(0)=k_-(0)=0$,[5] so that data in the neighborhood of $\gamma_0$ are not used in estimating $\gamma_0$. 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.

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 $d=1$.[6]

## 2.3 Estimation of $\delta$

Given $\widehat{\gamma}$, we can estimate $\delta$ as if $\gamma_0$ were known. Due to the superconsistency of $\widehat{\gamma}$, the asymptotic distribution of our estimator $\widehat{\delta}$ is unaffected by the estimation of $\gamma$ and is the same as when $\gamma_0$ is known. We provide two estimators of $\delta$, both of which are based on the observation that

$$m_-(x)-m_+(x)\equiv\mathbb{E}[y|x,q=\gamma_0-]-\mathbb{E}[y|x,q=\gamma_0+]=\delta_{\alpha 0}+x'\delta_{x0}+\gamma_0\delta_{q0}. \tag{6}$$

[5] 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.

[6] In (2), if we neglect the data on $x$, the relationship between $y$ and $q$ is $y=E[g(x,q)|q]+E[\mathbf{x}'\delta|q]1(q\leq\gamma)+v$ with $v=e+g(x,q)-E[g(x,q)|q]+(\mathbf{x}'\delta-E[\mathbf{x}'\delta|q])\,1(q\leq\gamma)$ satisfying $E[v|q]=0$. From Porter and Yu (2011), in the limit distribution of the DKE, $z_{1i}=2E[\mathbf{x}'\delta|q=\gamma_0]v_i^-+(E[\mathbf{x}'\delta|q=\gamma_0])^2$ and $z_{2i}=-2E[\mathbf{x}'\delta|q=\gamma_0]v_i^++(E[\mathbf{x}'\delta|q=\gamma_0])^2$ with $v_i^{\pm}$ similarly defined as $e_i^{\pm}$, so $E[z_{\ell i}]=(E[\mathbf{x}'\delta|q=\gamma_0])^2$. On the other hand, if we neglect $f(x_i,\gamma_0)f(x_i)$ in the limit distribution of the IDKE, $E[z_{\ell i}]=E[(\mathbf{x}'\delta)^2\,|q=\gamma_0]\geq(E[\mathbf{x}'\delta|q=\gamma_0])^2$, 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.

The first estimator of $\delta$ is the IDKE. From (6), $\delta_{x0}$ and $\delta_{q0}$ 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 (6),

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

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

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

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 $\xi^{S_p}$ 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 (5).

If $\gamma_0$ 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

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

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 $l$th 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,*

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

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

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

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

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

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 $\delta$ 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,*

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

*where*

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

*and*

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

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

The convergence rate of the DKE of $\Delta_0$ 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 $\widehat{\Delta}$,

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

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 $B_\Delta$ 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 $\Sigma_\Delta$ 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 $\delta$ is based on another implication of (6), namely that $\delta_0$ 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 $\delta$. 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 $\delta_q$ is zero. To avoid this problem, we may regress $\widehat{a}_-(x_i) - \widehat{a}_+(x_i)$ only on $(1, x_i')'$. Specifically, define

$$(\overline{\delta}_\alpha, \widetilde{\delta}_x')' = \arg\min_{\underline{\delta}} \frac{1}{n}\sum_{i=1}^{n} k\left(\frac{q_i-\widehat{\gamma}}{h}\right)\left[\widehat{a}_-(x_i) - \widehat{a}_+(x_i) - (1, x_i')\,\underline{\delta}\right]^2. \tag{7}$$

Note that $\overline{\delta}_\alpha$ estimates $\delta_{\alpha 0} + \gamma_0\delta_{q0}$, so we can estimate $\delta_{\alpha 0}$ by

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

where $\widehat{\delta}_q$ is the IDKE of $\delta_{q0}$. 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 $\Psi$ as

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

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

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

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

*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* $\gamma_0 = 0$,

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

*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*

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

*where*

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

*with $\Sigma_q$ 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 $\delta$ (which are necessary for statistical inference) is a standard econometric exercise, it is omitted here.

## 2.4 Intuition for the Identifiability of $\gamma$ and $\delta$

Although our analysis shows that $\gamma$ and $\delta$ 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

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

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 1. 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, 2015) 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

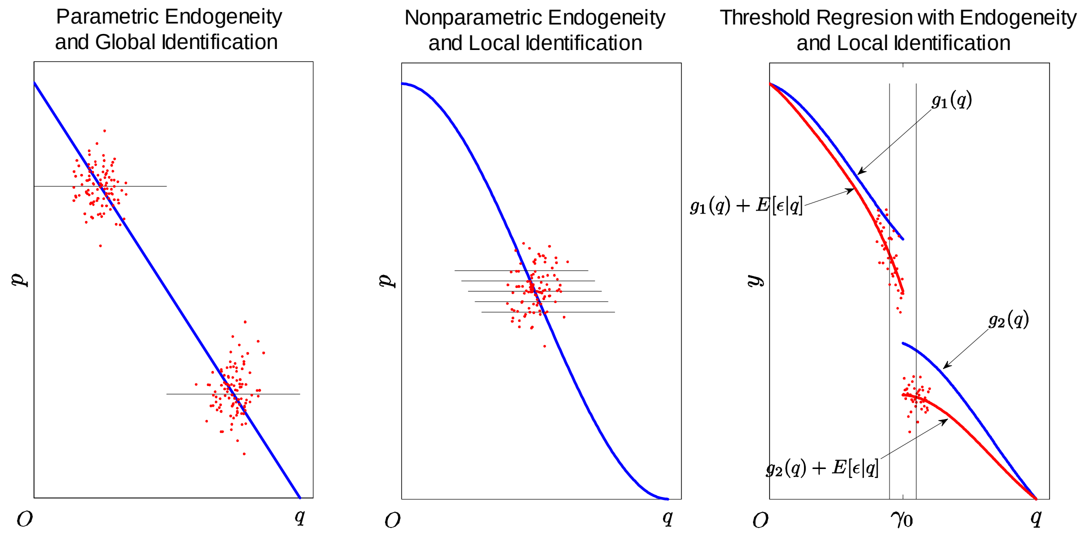

Figure 1: Graphical Intuition for the Identification of the Demand Curve under Endogeneity in Parametric, Nonparametric and Threshold Regression Models

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, 2015), 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 1 gives some graphical intuition exhibiting this identification scheme.

In threshold regression with endogeneity, the system contains a local shifter that helps to identify $\gamma_0$ 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 $\gamma_0$ even in nonparametric threshold regression with endogeneity. To be explicit, suppose $y_i = g(q_i) + \varepsilon_i = g_1(q_i)1(q_i \le \gamma) + g_2(q_i)1(q_i > \gamma) + \varepsilon_i$, where $g_1$ and $g_2$ are smooth functions with $g_1(\gamma_0) \ne g_2(\gamma_0)$, and $\mathbb{E}[\varepsilon|q] \ne 0$. 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

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

In other words, we may use the indicator $1(q > \gamma)$ to shift $y$ from the left neighborhood of $\gamma$ to the right neighborhood, and check which shifter provides the largest variation in $\mathbb{E}[y]$. Carefully checking this objective function, we see that it is the numerator of the Wald estimator using only local-to-$\gamma$ data.[7] 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,

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

the RDD literature concentrates on identifying the jump size, while we are interested in the jump location.[8] 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 4.1 for further discussion 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 (6) can be identified for any $x_i$. Following the discussion in Section 2.3, $\widehat{\delta}$ or $\widetilde{\delta}$ can be used to identify $\delta_0$. The right panel of Figure 1 illustrates this intuition concerning the identification schemes for $\gamma_0$ and $\delta_0$.

# 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 $\beta$, $\delta$ and $\gamma$ 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 $\beta$ cannot be identified without instrumentation since the effect of $\mathbf{x}'\beta$ and $\mathbb{E}[\varepsilon|x, q]$ are intermixed, just as the parameter $\beta$ 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 $\delta$ and $\gamma$ can be identified, with $\delta$ being estimable at a nonparametric rate whereas $\gamma$ is estimable at the same rate as the parametric case. In this section, we first study the optimal rate of convergence for estimates of $\delta$ and then give the optimal estimation rate for $\gamma$ from the existing literature.

To obtain the optimal rate of convergence for $\delta$, 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 $\theta$ be a functional defined on $\mathcal{P}$. Given an estimator $\widehat{\theta}$ of $\theta$ 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

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

(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.[9] 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 $\theta$. The goal is to find an achievable lower bound for the minimax risk defined by

$$\inf_{\widehat{\theta}}R\left(\widehat{\theta},\mathcal{P}\right)=\inf_{\widehat{\theta}}\sup_{P\in\mathcal{P}}\mathbb{E}_P\left[L\left(\widehat{\theta},\theta(P)\right)\right]. \tag{8}$$

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 $\gamma_0$ can be estimated at rate $n$, its estimation does not affect the optimal rate of convergence of $\delta$. We therefore assume that $\gamma_0$ is known in deriving the optimal rate of convergence of $\delta$.[10] Now $P\in\mathcal{P}$ is characterized by $\delta$ and $g(x,q)$ as follows

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

where $\mu$ 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 $\theta$ can be any element of $\delta$, e.g., $\delta_\alpha\left(P_{g,\delta}\right)=\delta_\alpha$. The following theorem provides upper bounds for the rates of convergence.

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

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

*and*

$$\begin{aligned}\varliminf_{n\to\infty}\inf_{\widehat{\delta}_\alpha}\sup_{P\in\mathcal{P}(s,B)}P\left(\left|n^{\frac{s-1}{2s+1}}\left(\widehat{\delta}_\alpha-\delta_\alpha(P)\right)\right|>\frac{\epsilon}{2}\right)&\ge C\text{ if }\gamma_0\ne0,\\ \varliminf_{n\to\infty}\inf_{\widehat{\delta}_\alpha}\sup_{P\in\mathcal{P}(s,B)}P\left(\left|n^{\frac{s}{2s+1}}\left(\widehat{\delta}_\alpha-\delta_\alpha(P)\right)\right|>\frac{\epsilon}{2}\right)&\ge C\text{ if }\gamma_0=0\end{aligned}$$

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

This theorem has interesting consequences. First, the main result is that we can estimate $\delta$ at most at a nonparametric rate. Second, estimation of $\delta$ does not suffer the curse of dimensionality. Specifically, an upper bound to the rate of convergence for $\delta_x$ is the same as for one-dimensional conditional mean estimation, and the upper bound for $\delta_q$ is the same as for one-dimensional slope estimation. As for $\delta_\alpha$, the upper bound depends on whether $\gamma_0=0$ or not: if $\gamma_0\ne0$, the upper bound is the same as in slope estimation; otherwise, it is the same as in level estimation. The upper bound for $\delta_q$ is not a surprise because $\delta_q$ is the slope difference in the neighborhood of $q=\gamma_0$. However, it may seems mysterious why $\delta_x$, as the *slope* difference in the

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

[10] The problem with unknown $\gamma_0$ is harder than the problem with known $\gamma_0$, so the upper bounds in Theorem 4 below are also the upper bounds for the problem with unknown $\gamma_0$. Given that these upper bounds are achievable even if $\gamma_0$ were unknown, these bounds are also the optimal rates of convergence with unknown $\gamma_0$.

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 $\gamma_0$ 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 $\gamma_0$ 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 (2014b) for more discussion on this point.

With this intuition on the optimal rate for $\delta_x$, the upper bound for $\delta_\alpha$ 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 $\delta_x$ has the optimal rate $\frac{s}{2s+1}$, so the optimal rate for $\delta_\alpha$ 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 $\delta_x$, which is $\frac{s}{2s+1}$. Otherwise, its optimal rate is determined by the optimal rate of $\delta_q$, 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 $\delta_\alpha$, $\delta_x$ and $\delta_q$, 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 $\delta_x$, 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 $\delta$ 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 $\delta_x$.[11] 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 $\gamma$, the optimality result is more subtle. In the parametric model, Yu (2012) shows that the Bayes estimator is efficient in the minimax 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 $\gamma_0$ in the semiparametric case; in other words, the nonparametric components of the model do not affect the efficiency of $\gamma_0$, so that $\gamma_0$ can be estimated as if these components were known. Specifically, the following procedure is used to adaptively estimate $\gamma_0$ in the present case.

**Algorithm G**:

**Step 1:** Compute the IDKE $\left(\widehat{\gamma},\widehat{\delta}'\right)'$, $\widehat{g}(x_i,q_i)=\frac{1}{(n-1)\widehat{f}_i}\sum\limits_{j=1,j\neq i}^{n}K_{h,ij}\left(y_j-\mathbf{x}_j'\widehat{\delta}1(q_j\leq\widehat{\gamma})\right)$ 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 $\widehat{f}_i=\frac{1}{n-1}\sum\limits_{j=1,j\neq i}^{n}K_{h,ij}$ with

[11] Another estimator that fully exploits the linear index structure of the model is the PLE of $\delta$ as discussed in the supplementary materials. We conjecture that this estimator also achieves the optimal rate of $\delta$. 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 $d=1$.

$K_{h,ij} = K^x_{h,ij} \cdot k_h(q_j - q_i)$ is the kernel estimator of $f_i \equiv f(x_i, q_i)$.

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

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

is the estimated likelihood function.

The asymptotic distribution of $\widehat{\gamma}_o$ is $\arg\min\limits_t \int_{\mathbb{R}} l\,(t-v)\,p^*(v)dv$, 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 $\overline{z}_{1i} \equiv \ln\frac{f_{e|x,q}\left(e_i + \mathbf{x}_i'\delta_0 \middle| x_i, q_i\right)}{f_{e|x,q}(e_i|x_i,q_i)}$ and $\overline{z}_{2i} \equiv \ln\frac{f_{e|x,q}\left(e_i - \mathbf{x}_i'\delta_0 \middle| x_i, q_i\right)}{f_{e|x,q}(e_i|x_i,q_i)}$. Note also that the nonparametric posterior interval (NPI) based on $\widehat{\mathcal{L}}_n(\gamma)$ is a valid confidence interval for $\gamma_0$;[12] for other inference methods for $\gamma$, see Liao et al. (2015).

## 3.2 Optimality Results With Instruments

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

$$\mathbb{E}\left[z\varepsilon 1\,(q \le \gamma_0)\right] = 0, \text{ and } \mathbb{E}\left[z\varepsilon 1\,(q > \gamma_0)\right] = 0, \tag{9}$$

where $z \in \mathbb{R}^{d_z}$ with $d_z \ge d+1$. Note that here we do not require $\mathbb{E}[\varepsilon|z,q] = 0$ as in Caner and Hansen (2004) to identify $\left(\beta', \delta'\right)'$.[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 $\gamma_0$ can be consistently estimated by the IDKE, we can treat it as known in constructing the GMM objective function and estimates. Specifically,

$$\left(\widehat{\beta}'_{GMM}, \widehat{\delta}'_{GMM}\right)' = \arg\min_{\beta,\delta} n\overline{m}_n(\beta,\delta)' W_n \overline{m}_n(\beta,\delta), \tag{10}$$

where

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

and $W_n$ is a consistent estimator of the inverse of

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

[12] See Section 4.1 of Yu (2014a) for a summary of valid inference methods in threshold regression without endogeneity.

[13] Since $\delta$ is already identified, we need only one of the two moment conditions in (9) to identify $\beta$.

For example, $W_n$ can be the inverse of the sample analog of $\Omega$, say,

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

where $\widetilde{\varepsilon}_i = y_i - \mathbf{x}_i'\widetilde{\beta} - \mathbf{x}_i'\widetilde{\delta}1(q_i \le \widehat{\gamma})$, and $\left(\widetilde{\beta}', \widetilde{\delta}'\right)'$ is the 2SLS estimator of $\left(\beta', \delta'\right)'$ which is defined as the minimizer of (10) with

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

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

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

is a consistent estimator of

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

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

**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*

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

*where the inverse of $\Omega$ 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 $\delta$, in contrast to the purely local identification information that is used when $z$ is absent. Meanwhile, $\beta$, 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 $\gamma$ other than the IDKE.

From Hansen (1982), $\left(G'\Omega^{-1}G\right)^{-1}$ is the optimal asymptotic variance under moment conditions (9) with $\gamma_0$ known. Actually, according to Yu (2008), the GMM estimator is semiparametrically efficient even when $\gamma_0$ is unknown and the estimate $\widehat{\gamma}$ is used, as long as the loss function imposed on $(\beta', \delta')'$ and $\gamma$ 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 $\Omega$, 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 $\gamma$, 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 $\varepsilon_i$ (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 $\gamma_0$ by $\arg\min\limits_t \int_\Gamma l_n(t-\gamma)\widehat{\mathcal{L}}_n(\gamma)\pi(\gamma)\,d\gamma$, 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 $\gamma$ improves its efficiency without affecting the convergence rate.

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

$$\begin{aligned} y &= \delta 1(q \leq \gamma) + \varepsilon, \\ \mathbb{E}[\varepsilon|q] &= g(q) \neq 0, \, \mathbb{E}[\varepsilon] = 0. \end{aligned} \tag{11}$$

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 $\pi$ 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 $\pi_0^2$, so the instrument $z$ indeed improves the efficiency of $\gamma$ estimation. Table 1 provides numerical results for this example based on the algorithms in Appendix D of Yu (2012). 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 $\pi_0$ increases, $z$ indeed provides more information about $\gamma$ 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 $\gamma$ 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 $\beta$ and $\delta$ and achieve a $\sqrt{n}$ convergence rate.

| | 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 $\gamma$ Estimation by $z$: $\rho_0 = 0.5$, $\delta_0 = 1$, and $\gamma_0 = 0$.

[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.

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

| | Without Instruments | With Instruments |
|---|---|---|
| $\beta$ | Unidentified | $\sqrt{n}$-consistency |
| $\delta$ | Nonparametric Consistency | $\sqrt{n}$-consistency |
| $\gamma$ | $n$-consistency | Efficiency Improvement |

Table 2: The Role of Instruments on the Estimation Properties of the Parameters

# 4 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.

## 4.1 $\mathbb{E}[\varepsilon|x, q]$ 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

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

If $\gamma_0 = 0$, then $\mathbb{E}\left[\varepsilon|q\right]$ is continuous, but has a cusp at $q = \gamma_0$. If $\gamma_0 \neq 0$, then $\mathbb{E}\left[\varepsilon|q\right]$ 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 |
| $\beta$ | Unidentified | $\sqrt{n}$-consistency | Unidentified | $\sqrt{n}$-consistency |
| $\delta_\alpha, \delta_q$ | Unidentified | $\sqrt{n}$-consistency | Unidentified | $\sqrt{n}$-consistency |
| $\delta_x$ | Nonparametric Consistency | $\sqrt{n}$-consistency | Unidentified | $\sqrt{n}$-consistency |
| $\gamma$ | $n$-consistency | Efficiency Improvement | $n$-consistency | Efficiency Improvement |

Table 3: The Role of Instrumentation for Estimation Properties of 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 $\delta_x$ in Section 2.3 are still applicable. On the other hand, $\delta_\alpha$ and $\delta_q$ 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 $\delta_{q0}$ cannot be individually identified since $\delta_{q0}$ cannot be identified due to the cusp at $\gamma_0$.[15] 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 $q=\gamma_0$,[16] no elements of $\delta$ can be identified, but $\gamma$ can still be identified and estimated at the rate of $n$ by the IDKE.

## 4.2 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 $\beta_\ell$ and $\delta$ are partitioned according to the partition of $\mathbf{x}=(1,x_1',x_2')'$ as $\left(\beta_{\ell\alpha},\beta_{\ell1}',\beta_{\ell2}'\right)$ 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.

First, the IDKE of $\gamma$ can be simplified. For each $\gamma\in\Gamma$, $\mathbb{E}[y|x_i,q=\gamma-]$ can be estimated as follows. The components $\beta_{11}$ 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

$$\min_{\beta_{11},a_1,\cdots,a_n}\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n K_{h,ij}^{\underline{x}_2}k_h^-(q_j-\gamma)\left[y_j-a_i-x_{1j}'\beta_{11}\right]^2, \tag{12}$$

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 (5). Note that $\beta_{11}$ is the same for all $\underline{x}_{2i}$ in the objective function (12). 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

$$\mathbb{E}\left[(x_1-\mathbb{E}[x_1|\underline{x}_2,q=\gamma-])(x_1-\mathbb{E}[x_1|\underline{x}_2,q=\gamma-])'\right|q=\gamma-\left.\right]>0,$$

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 $\gamma_0$ 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 $q\leq\widehat{\gamma}$ and $q>\widehat{\gamma}$ to estimate $\beta_{11}$ and $\beta_{21}$ using either the double residual regression method of Robinson (1988) or the pairwise difference estimator of Powell (1987, 2001).

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

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

The resulting estimators are expected to be $\sqrt{n}$-consistent when

$$\mathbb{E}\left[(x_1 - \mathbb{E}[x_1|x_2])(x_1 - \mathbb{E}[x_1|x_2])' 1(q \le \gamma_0)\right] > 0 \text{ and } \mathbb{E}\left[(x_1 - \mathbb{E}[x_1|x_2])(x_1 - \mathbb{E}[x_1|x_2])' 1(q > \gamma_0)\right] > 0.^{17}$$

Note that here we use all the data with $q \le \widehat{\gamma}$ to estimate $\beta_{11}$ but only the data in the left $h$ neighborhood of $q = \gamma$ to estimate $\beta_{11}$ in (12). 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 (12) may actually be estimating $\beta_{21}$. Given the estimates of $\beta_{11}$ and $\beta_{21}$, which we still denote as $\widehat{\beta}_{11}$ and $\widehat{\beta}_{21}$ to simplify notation, we can construct

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

which satisfies

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

So here $\delta_\alpha$ and $\delta_2$ can be estimated using the procedures in Section 2.3 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 6, where $x_2$ is binary, we show that even kernel smoothing is not required. If we further assume that $\varepsilon$ 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.[18]

# 5 Simulation Results

This section presents a simple simulation study designed to assess the adequacy of the limit theory. The simulation compares the efficiency of the IDKE and DKE of $\gamma$.

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 $\varepsilon|\left(x, q\right) \sim N(-q, 0.2^2)$. The threshold effect does not depend on $x$ and $x|q$ is uniformly distributed, 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 the formula $Cn^{-1/6}$ with proportionality constants $C = 0.3$, 0.5 and 0.7.[19] 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.

[17] 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 $E[\underline{mm}'] > 0$, where $\underline{m} = \begin{pmatrix} (x_1 1(q \le \gamma_0) - E[x_1 1(q \le \gamma_0)|x_2])\, 1(q \le \gamma_0) \\ (x_1 1(q > \gamma_0) - E[x_1 1(q > \gamma_0)|x_2])\, 1(q > \gamma_0) \end{pmatrix}$.

[18] 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 course, $\varepsilon$ cannot be independent of $(x', q)'$, but it is quite possible that $\varepsilon$ 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.

[19] $C = 0.3$ 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 $N = n \times (2Cn^{-1/6}) \times Cn^{-1/6} = 2C^2 n^{2/3}$ data points in a $h$ neighborhood of $(x_i, \gamma)$. When $c = 0.3$ and $n = 200$, $N \approx 6$. When $c = 0.7$ and $n = 800$, $N \approx 84$.

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)

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 2 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 $O(n)$[20]. 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 2.1. 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.

# 6 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

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

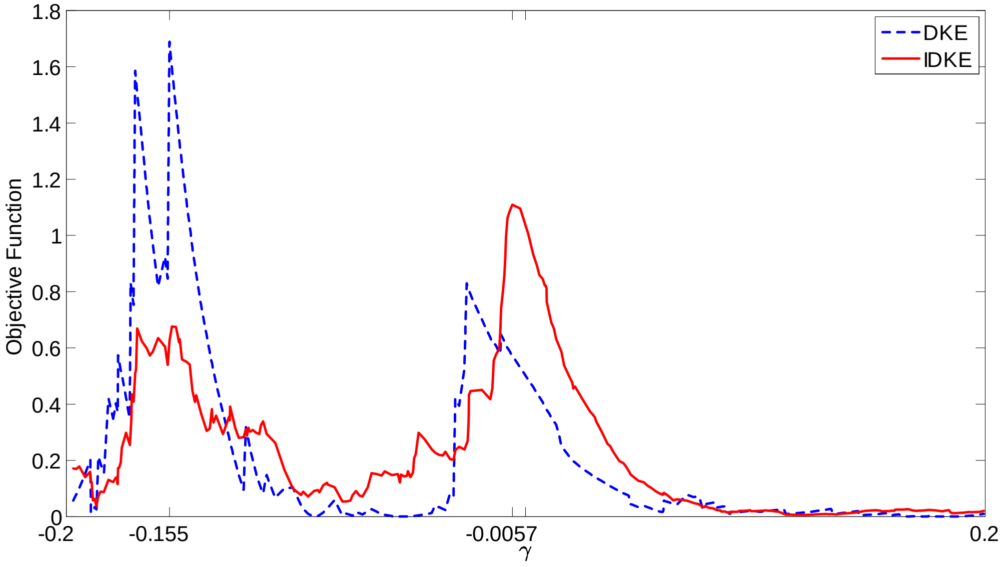

Figure 2: Objective Functions of the DKE and IDKE

employee 401(k) contributions. Therefore, only workers in firms that offer such programs are eligible, whereas IRAs are open to all.[21]

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).[22] 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.

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-

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

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

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 5 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

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

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).

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

Table 5: 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.

The findings that emerge from Table 5 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[23]

$$y = \begin{cases} D\alpha_1 + X'\beta_1 + \varepsilon, & inc \leq \gamma_1, \\ D\alpha_2 + X'\beta_2 + \varepsilon, & \gamma_1 < inc \leq \gamma_2, \\ D\alpha_3 + X'\beta_3 + \varepsilon, & inc > \gamma_2, \end{cases} \tag{13}$$

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

Model (13) is very special since the only endogenous variable $D$ is binary. As in Section 4.2, suppose $\varepsilon$ 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 $D$.[24] In other words, the relationship between $y$ and $(D, X')'$ satisfies

$$y = \begin{cases} D\widetilde{\alpha}_1 + \widetilde{\beta}_{10} + \underline{X}'\underline{\beta}_1 + e, & inc \leq \gamma_1, \\ D\widetilde{\alpha}_2 + \widetilde{\beta}_{20} + \underline{X}'\underline{\beta}_2 + e, & \gamma_1 < inc \leq \gamma_2, \\ D\widetilde{\alpha}_3 + \widetilde{\beta}_{30} + \underline{X}'\underline{\beta}_3 + e, & inc > \gamma_2, \end{cases} \tag{14}$$

where $\underline{X}$ ($\underline{\beta}_\ell$) is $X$ ($\beta_\ell$) excluding the constant (the intercept), $\widetilde{\alpha}_\ell$ and $\widetilde{\beta}_{\ell 0}$, $\ell = 1, 2, 3$, may differ from those in (13), but $\underline{\beta}_\ell$, $\ell = 1, 2, 3$, are the same as in (13). The new error term satisfies $\mathbb{E}[e|D, X] = 0$. Given this structure, the LSEs of $\gamma_1$ and $\gamma_2$ are consistent although the LSEs of $\alpha_\ell$, $\ell = 1, 2, 3$, are inconsistent. We use the sequential estimation procedure of Bai (1997) to consistently estimate $\gamma_1$ and $\gamma_2$. Given a consistent estimator of $\gamma_1$ and $\gamma_2$, $\alpha_\ell$ and $\beta_\ell$ can be consistently estimated by the 2SLS procedure developed here, and a consistent estimate of $\varepsilon$ 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.

Given the LSE of $\gamma_1$ and $\gamma_2$, we can use the modified Algorithm G in Section 3.2 to improve efficiency in estimation of $\gamma_1$ and $\gamma_2$. 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.[25] Then as argued in Section 4.2, we need only estimate $f(\varepsilon|D, z)$ to construct the nonparametric posterior interval (NPI) for $\gamma$. In other words, only three univariate density functions are estimated.[26] 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 (13) to construct the NPI for $\gamma_1$, and combine Regimes II and III to construct the NPI for $\gamma_2$, rather than constructing the NPI for $\gamma_1$ and $\gamma_2$ simultaneously. All implementation details and code are available upon request.

| | $\gamma_1$ | $\gamma_2$ | $n$ 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 | $[42.810, 42.876]$ | $[71.087, 71.358]$ | | | |

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

[24] 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; see Yu (2015) for more discussions on the consistency of the LSE for the threshold points in the presence of endogeneity.

[25] $E[\varepsilon|D, X] = E[\varepsilon|D]$ 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.

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

Table 6: Estimates of $\gamma_1$ and $\gamma_2$, the NPI and Numbers of Data Points in Each Regime

Another estimator of $\gamma_1$ and $\gamma_2$ 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, 2013a). 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 $\gamma_1$ and $\gamma_2$ mentioned above and the corresponding three regimes are summarized in Table 6. Some of the findings in Table 6 are summarized as follows. First, Regime I is the same for all estimators. Second, compared to the Caner-Hansen 2SLS estimator, the LSE of $\gamma_2$ 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 (2013a) 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 $\gamma_1$ and $\gamma_2$ are narrow (each interval covers only 12 data points), which indicates that regime splitting by the posterior mean (or median) is precise here.

Table 7 reports the OLS and 2SLS estimates of $\left(\alpha_\ell, \beta_\ell'\right)'$ 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 $\alpha_\ell$ 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 $\alpha_\ell$ 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 5, 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.[27],[28] Such limitations do not exist for 401(k) plans, although there is a maximum deductible level.[29] The analysis above shows

[27] 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.

[28] 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 7. 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 (unlike 401(k) eligibility) is trivially satisfied and is not a valid instrument.

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

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 developed in Liao et al. (2015) 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 (14) with no instruments available. We adapt both the (sup and average) Wald test and the score test to this environment; all four tests reject the null strongly with $p$-values equal to zero. The second test is based on (13) 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 $\gamma_1$ and $\gamma_2$ and 2SLS estimation of $\alpha_\ell$ 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 $\gamma$ and $\alpha$ parameters provided we use the model (14) for compliers. A formal extension of our analysis to this framework is of interest but 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 7: 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.

# 7 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 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, the estimation 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. Assumption H does not provide specific criteria for bandwidth selection besides the constraints on rates. Porter and Yu (2011) suggest using cross validation to select $h$ 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.

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 and Chan et al. (2014) for an analysis of the conditionally heteroscedastic AR model with thresholds. 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. The estimation techniques developed in this paper can also be used to other microeconometric models. For example, (i) the transformation model, $\Lambda(y) = \mathbf{x}'\beta + \mathbf{x}'\delta 1(q \leq \gamma) + \varepsilon$, where $\Lambda(\cdot)$ is an unknown strictly increasing transformation; (ii) the limited dependent variable model, $y = \Lambda(y^*)$, where $\Lambda(\cdot)$ is a known noninvertible function, and the latent variable $y^* = \mathbf{x}'\beta + \mathbf{x}'\delta 1(q \leq \gamma) + \varepsilon$; (iii) threshold panel data models, $y_{it} = \alpha_i + \mathbf{x}_{it}'\beta + \mathbf{x}_{it}'\delta 1(q_{it} \leq \gamma) + \varepsilon_{it}$, where $\mathbf{x}_{it}$ may include lagged $y_{it}$'s and $q_{it}$ may be a lagged $y_{it}$. As long as the *observed* dependent variable has a jump at $q = \gamma$, we can use the IDKE to identify the threshold point $\gamma$.

7. 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.

8. Due to space constraints, the present paper has concentrated on model identification and estimation. In a parallel and complementary work, Liao et al. (2015) develop methodology for specification testing, focussing on a test for endogeneity and a test for the presence of threshold effects, working as here both with and without instruments.

# 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.

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.

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., D. Li, S. Ling and H. Tong, 2014, On Conditionally Heteroscedastic AR Models with Thresholds, *Statistica Sinica*, 24, 625-652.

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.

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.

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

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

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

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

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.

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.

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

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.

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.

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

Liao, Q., P.C.B. Phillips and P. Yu, 2015, Inferences and Specification Testing in Threshold Regression with Endogeneity, mimeo, HKU.

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, forthcoming, *Journal of Econometrics*.

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.

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.

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.

Wang, Q. and P. Phillips, 2015, Nonparametric Cointegrating Regression with Endogeneity and Long Memory", *Econometric Theory*, 2015 (forthcoming).

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

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

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

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

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

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

Yu, P., 2014b, A Note on Average Derivative Estimation, mimeo, Department of Economics, HKU.

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

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

# 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. Propositions that are used in these derivations are given in Appendix B and additional lemmas that are needed are given in Appendix C. Also, to save space, the proofs for the asymptotic distributions of $\delta$ estimators in Section 2.3 and GMM estimators in Section 3.2 are relegated to supplementary materials.

**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 $n\left(\widehat{\gamma}-\gamma_0\right)$. From Proposition 3, for $v$ in any compact set of $\mathbb{R}$,

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

where $\overline{z}_{1i}$ and $\overline{z}_{2i}$ are defined in the main text. Now, we can obtain the asymptotic distribution of $n\left(\widehat{\gamma}-\gamma_0\right)$ by applying the same argument as in the proofs of Theorem 1 and 2 in Yu (2012). 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 $nh^{d-1}\left(\widetilde{\gamma}-\gamma_0\right) = O_p(1)$ are similar to Theorem 1, so are omitted here. We concentrate on deriving the weak limit of the localized process $nh^d\left(\widehat{\Delta}_o^2\left(\gamma\right)-\widehat{\Delta}_o^2\left(\gamma_0\right)\right)$ for $\gamma$ in an $\left(nh^{d-1}\right)^{-1}$ neighborhood of $\gamma_0$ .

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

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

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

$$\begin{aligned}
& \mathbb{E}\left[\widehat{\Delta}_o\left(\gamma\right)\right] = \int_{-1}^0\int K^x(u_x,x_o)k_-(u_q)g(x_o+u_xh,\gamma+u_qh)f(x_o+u_xh,\gamma+u_qh)du_xdu_q \\
& +\int_{-1}^{\frac{\gamma_0-\gamma}{h}}\int K^x(u_x,x_o)k_-(u_q)\left(1,(x_o+u_xh)',\gamma+u_qh\right)\delta_0 f(x_o+u_xh,\gamma+u_qh)du_xdu_q \\
& -\int_0^1\int K^x(u_x,x_o)k_+(u_q)g(x_o+u_xh,\gamma+u_qh)f(x_o+u_xh,\gamma+u_qh)du_xdu_q \\
= \; & \left(1,x_o',\gamma_0\right)\delta_0 f(x_o,\gamma_0)+O(h).
\end{aligned}$$

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

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

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

$$D_o\left(v\right)=\begin{cases}\sum\limits_{i=1}^{N_1(|v|)} z_{1i}, \text{ if } v\leq 0,\\ \sum\limits_{i=1}^{N_2(v)} z_{2i}, \text{ if } v>0,\end{cases}$$

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

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

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

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

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

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

and

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

Given the weak limit of $nh^d\left(\widehat{\Delta}_o^2\left(\gamma_0+\frac{v}{a_n}\right)-\widehat{\Delta}_o^2\left(\gamma_0\right)\right)$, 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 (2012). Since these checks are all similar, we omit the details here and only note that $\arg\max\limits_v \overline{D}(v)=\arg\min\limits_v D(v)$, given that $k_-\left(0\right)=k_+\left(0\right)>0$ and $f(x_o,\gamma_0)>0$. ■

**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

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

where the infimum is taken over the measurable function $\phi$ 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 $\mu$ 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

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

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

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

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

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

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

$$\inf_{\widehat{\delta}_\alpha} \sup_{\mathbb{P} \in \mathcal{P}(s,B)} \mathbb{P}\left(\left|\widehat{\delta}_\alpha - \delta_\alpha(\mathbb{P})\right| > \epsilon/2\right) \geq \frac{1}{2}\left(\frac{3}{2} - \exp\left(\sum_{i=1}^n \nu_i^2\right)\right) \tag{15}$$

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 $\delta_P$ are changed to $g_Q$ and $\delta_Q$, respectively. We now specify $g$ and $\delta$ 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 $\xi$ is a positive constant, $\eta = n^{-1/(2s+1)}$, $\varphi_q$ 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 $\xi$ is small enough, say $\xi^2/2 \leq \log(5/4)$, we have

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

It follows from (15) that

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

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

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

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

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

The above argument also shows that the optimal rate of convergence for $\delta_q$ is $n^{-\frac{s-1}{2s+1}}$. As for $\delta_x$, 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 $\varphi_q$ is an infinitely differentiable function in $q$ satisfying (i) $\varphi_q(v) = 0$ for $v \geq 0$, (ii) $\varphi_q(v) = 1$, for $v \leq -\zeta$, and (iii) $\varphi_q(v) \in (0,1)$ for $v \in (-\zeta, 0)$. Then

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

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

# Appendix B: Propositions

The following propositions are needed in the proof of Theorem 1 and Corollary 1 and hold under the conditions of that theorem.

**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

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

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, $\left(1, x', \gamma_0\right) \delta_0 \neq 0$ 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.

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

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

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

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

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

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 less $j$ terms 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

$$\begin{aligned} D2 \quad = \quad & \frac{1}{n\,(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} e_j\left(K_{h,ij}^{\gamma-}-K_{h,ij}^{\gamma_0^-}\right)1(q_j\le\gamma_0)+\frac{1}{n\,(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} e_j\left(K_{h,ij}^{\gamma_0^+}-K_{h,ij}^{\gamma+}\right)1(q_j>\gamma) \\ & +\frac{1}{n\,(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} e_j\left(K_{h,ij}^{\gamma-}+K_{h,ij}^{\gamma_0^+}\right)1\left(\gamma_0<q_j\le\gamma\right)\equiv D_{21}+D_{22}+D_{23}. \end{aligned}$$

By Lemma 8.4 of Newey and McFadden (1994), we can show

$$D_{21} \approx \frac{1}{n}\sum_{j=1}^{n}\frac{e_j}{h}\left[k_-\left(\frac{q_j-\gamma}{h}\right)-k_-\left(\frac{q_j-\gamma_0}{h}\right)\right]1(q_j \le \gamma_0),$$

so $Var\left(D_{21}\right) = O\left(\frac{1}{nh}\left(\frac{\gamma-\gamma_0}{h}\right)^2\right)$. Similarly, $Var(D_{22}) = O\left(\frac{1}{nh}\left(\frac{\gamma-\gamma_0}{h}\right)^2\right)$. As to $D_{23}$, we can show

$$D_{23} \approx \frac{1}{n}\sum_{j=1}^{n}\frac{e_j}{h}\left[k_-\left(\frac{q_j-\gamma}{h}\right)+k_+\left(\frac{q_j-\gamma_0}{h}\right)\right]1\left(\gamma_0 < q_j \le \gamma\right),$$

so $Var\left(D_{23}\right) = O\left(\frac{1}{nh}\frac{|\gamma-\gamma_0|}{h}\right)$. By the independence of U-projections of $D_{21}, D_{22}$ and $D_{23}$, we have

$$Var\left(D2\right) = O\left(\frac{1}{nh}\left(\frac{\gamma-\gamma_0}{h}\right)^2+\frac{1}{nh}\frac{|\gamma-\gamma_0|}{h}\right) = O\left(\frac{1}{nh}\frac{|\gamma-\gamma_0|}{h}\right).$$

In consequence,

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

by Markov's inequality. So

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

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

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

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

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

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

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 $x_o$, $\sum\limits_{j=1,j\neq i}^{n}$ changes to $\sum\limits_{j=1}^{n}$, and $K_{h,ij}^{\gamma\pm}$ changes to $K_{h,j}^{\gamma\pm}$, and

$$\begin{aligned}
T_1^{-} &= h^d\sum_{j=1}^{n}g(x_j,q_j)\left(K_{h,j}^{\gamma_0+}-K_{h,j}^{\gamma_0^v+}\right),\\
T_2^{-} &= h^d\sum_{j=1}^{n}\left[g(x_j,q_j)+\left(1,x_j',q_j\right)\delta_0\right]\left(K_{h,j}^{\gamma_0^v-}-K_{h,j}^{\gamma_0-}\right),\\
T_3^{-} &= -h^d\sum_{j=1}^{n}e_j1(q_j>\gamma_0)\left(K_{h,j}^{\gamma_0^v+}-K_{h,j}^{\gamma_0+}\right),\\
T_4^{-} &= h^d\sum_{j=1}^{n}e_j1(q_j\leq\gamma_0^v)\left(K_{h,j}^{\gamma_0^v-}-K_{h,j}^{\gamma_0-}\right),\\
T_5^{-} &= -h^d\sum_{j=1}^{n}e_j1(\gamma_0^v<q_j\leq\gamma_0)K_{h,j}^{\gamma_0-},(*)\\
T_6^{-} &= -h^d\sum_{j=1}^{n}\left[\left(1,x_j',q_j\right)\delta_0+e_j\right]1(\gamma_0^v<q_j\leq\gamma_0)K_{h,j}^{\gamma_0^v+}.(*)
\end{aligned}$$

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 $\epsilon>0$,

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

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. ■

# Appendix C: Lemmas

To save space, the proofs for all lemmas are relegated to the supplementary materials.

**Lemma 1** $\sup_{\gamma\in\Gamma}\left|\widehat{Q}_n(\gamma) - Q_n(\gamma)\right| \xrightarrow{p} 0.$

**Lemma 2** $\sum_{i=1}^{n}\sum_{l=1}^{4} T_{li} = o_p(1)$ *uniformly in* $v$.

**Lemma 3** $\sum_{i=1}^{n} \left(T_{5i} + T_{6i}\right) \Delta_f(x_i) = -k_+(0) \sum_{i=1}^{n} \left[\left(1, x_i', \gamma_0\right) \delta_0 - 2e_i\right] 1(\gamma_0 < q_i \leq \gamma_0^v) f(x_i) \Delta_f(x_i) + o_p(1).$

**Lemma 4** $\sum_{l=1}^{4} T_l^+ + \sum_{l=1}^{4} T_l^- = o_p(1)$ *uniformly in* $v$.

**Lemma 5** $T_5^+ + T_6^+ + T_5^- + T_6^- \xrightarrow{d} D_o(v).$