Threshold Regression with Nonparametric Sample Splitting

Yoonseok Lee* Yulong Wang†
Syracuse University Syracuse University

July 2022

Abstract

This paper develops a threshold regression model where an unknown relationship between two variables nonparametrically determines the threshold. We allow the observations to be cross-sectionally dependent so that the model can be applied to determine an unknown spatial border for sample splitting over a random field. We derive the uniform rate of convergence and the nonstandard limiting distribution of the nonparametric threshold estimator. We also obtain the root-n consistency and the asymptotic normality of the regression coefficient estimator. We illustrate empirical relevance of this new model by estimating the tipping point in social segregation problems as a function of demographic characteristics; and determining metropolitan area boundaries using nighttime light intensity collected from satellite imagery.

Keywords: threshold regression, sample splitting, nonparametric, random field, tipping point, metropolitan area boundary.
JEL Classifications: C14, C21, C24, R1

* *Address*: Department of Economics and Center for Policy Research, Syracuse University, 426 Eggers Hall, Syracuse, NY 13244. *E-mail*: `ylee41@syr.edu`

† *Address*: Department of Economics and Center for Policy Research, Syracuse University, 127 Eggers Hall, Syracuse, NY 13244. *E-mail*: `ywang402@syr.edu`

1 Introduction

Sample splitting and threshold regression models have spawned a vast literature in econometrics and statistics. Existing studies typically specify the sample splitting criteria in a parametric way as whether a single random variable or a linear combination of variables crosses some unknown threshold. See, for example, Hansen (2000), Caner and Hansen (2004), Seo and Linton (2007), Lee, Seo, and Shin (2011), Li and Ling (2012), Yu (2012), Hidalgo, Lee, and Seo (2019), Yu and Fan (2021), and Lee, Liao, Seo, and Shin (2021). In this paper, we study a novel extension to consider a *nonparametric* sample splitting model. Such an extension leads to new theoretical results and substantially generalizes the empirical applicability of threshold models.

Specifically, we consider a model given by

$$y_i = x_i^\top \beta_0 + x_i^\top \delta_0 \mathbf{1}\left[q_i \leq \gamma_0\left(s_i\right)\right] + u_i \tag{1}$$

for the ith entity, where $\mathbf{1}[\cdot]$ is the binary indicator. In this model, the marginal effect of x_i to y_i can be different across i as $(\beta_0 + \delta_0)$ or β_0 depending on whether $q_i \leq \gamma_0(s_i)$ or not. The threshold function $\gamma_0(\cdot)$ is unknown, and the main parameters of interest are β_0, δ_0, and $\gamma_0(\cdot)$. The novel feature of this model is that the sample splitting is determined by an unknown relationship between two variables q_i and s_i, and their relationship is characterized by the nonparametric threshold function $\gamma_0(\cdot)$. In contrast, the classical threshold regression models assume $\gamma_0(\cdot)$ to be a constant or a linear index. This new specification can handle interesting cases that have not been studied. For example, we can consider the threshold to be heterogeneous and specific to each observation i if we see $\gamma_0(s_i) = \gamma_{0i}$; or the threshold to be determined by the direction of a prediction error if we consider some moment condition $\gamma_0(s_i) = \mathbb{E}[q_i|s_i]$. Apparently, when $\gamma_0(s) = \gamma_0$ or $\gamma_0(s) = \gamma_0 s$ for some parameter γ_0 and $s \neq 0$, it reduces to the standard threshold regression model.

The new model is motivated by the following two applications: estimating potentially heterogeneous thresholds in public economics and determining spatial sample splitting in urban economics. The first one is about the tipping point problem by Schelling (1971), who analyzes the phenomenon that a neighborhood's white population substantially decreases once the minority share exceeds a certain threshold, called the tipping point. Card, Mas, and Rothstein (2008) empirically estimate the tipping point model by considering the constant threshold regression, $y_i = \beta_{10} + \delta_{10}\mathbf{1}[q_i \leq \gamma_0] + x_{2i}^\top \beta_{20} + u_i$, where y_i is the white population change in a decade and q_i is the initial minority share in the ith tract. The parameters δ_{10} and γ_0 denote the change size and the threshold, respectively. In Section VII of Card, Mas, and Rothstein (2008), however, they find that the tipping point γ_0 varies depending on the attitudes of white residents toward the minority. This finding motivates us to study the more general model (1) than the constant threshold model by specifying the tipping point γ_0 as a nonparametric function of local demographic characteristics. We estimate such a tipping point function in Section 6.1.

For the second application, we use the model (1) to determine metropolitan area boundaries, which is a fundamental problem in urban economics. Recently, many studies propose to use nighttime light intensity collected from satellite imagery to define the metropolitan area. They set an *ad hoc* level of light intensity as a threshold and categorize a pixel in the satellite imagery as a part of the metropolitan area if the light intensity of that pixel is higher than the threshold. See, for example, Rozenfeld, Rybski, Gabaix, and Makse (2011), Henderson, Storeygard, and Weil (2012), Dingel, Miscio, and Davis (2021), and Baragwanath, Goldblatt, Hanson, and Khandelwal (2021). In contrast, the model (1) can provide a data-driven guidance of choosing the intensity threshold from the econometric perspective, if we define y_i as the light intensity in the ith pixel and (q_i, s_i) as the location information of that pixel (more precisely, the coordinate of a point on a rotated map as described in Section 4). In Section 6.2, we estimate the metropolitan area of Dallas, Texas, especially its development from 1995 to 2010, and find substantially different results from the conventional approaches. To the best of our knowledge, this is the first study to nonparametrically determine the metropolitan area using a threshold model.

We develop a two-step estimation procedure of (1), where we estimate $\gamma_0(\cdot)$ by the local constant least squares. Under the shrinking threshold asymptotics as in Bai (1997), Bai and Perron (1998), and Hansen (2000), we show that the nonparametric estimator $\widehat{\gamma}(\cdot)$ is uniformly consistent and has a nonstandard limiting distribution. Based on this result, we develop a pointwise specification test of $\gamma_0(s)$ for any given s, which enables us to construct a confidence interval by inverting the test. Besides, the parametric part $(\widehat{\beta}^\top, \widehat{\delta}^\top)^\top$ is shown to satisfy the root-n asymptotic normality.

We highlight the novel features of the new estimator as follows. First, since the nonparametric function $\gamma_0(\cdot)$ is inside the indicator function, technical proofs of the asymptotic results are non-standard. In particular, we establish the uniform rate of convergence of $\widehat{\gamma}(\cdot)$, which involves substantially more complicated derivations than the standard (constant) threshold regression model. Second, we find that, unlike the standard kernel estimator, $\widehat{\gamma}(\cdot)$ is asymptotically unbiased even if the optimal bandwidth is used. Also, when the change size δ_0 shrinks very slowly, the optimal rate of convergence of $\widehat{\gamma}(\cdot)$ becomes close to the root-n rate. In the standard kernel regression, such a fast rate of convergence can be obtained when the unknown function is infinitely differentiable, while we only require the second-order differentiability of $\gamma_0(\cdot)$. Third, to limit the influence of the first-step estimation error to the second-step estimation, we propose to use the observations that are sufficiently away from the estimated threshold function $\widehat{\gamma}(\cdot)$ when obtaining the parametric estimators $(\widehat{\beta}^\top, \widehat{\delta}^\top)^\top$. The choice of this distance is obtained by the uniform convergence rate of $\widehat{\gamma}(\cdot)$. Fourth, we let the variables be cross-sectionally dependent by considering the strong-mixing random field as in Conley (1999) and Conley and Molinari (2007). This generalization allows us to study nonparametric sample splitting of spatial observations. For instance, if we let (q_i, s_i) correspond to the geographical location (i.e., latitude and longitude on the map), then the threshold identifies a unknown border yielding two-dimensional sample splitting. In more general contexts, the model can be applied to identify social or economic segregation over interacting agents. Finally, noting that $\mathbf{1}\left[q_i \leq \gamma_0\left(s_i\right)\right]$ can be considered as the special case of $\mathbf{1}\left[g_0\left(q_i, s_i\right) \leq 0\right]$ when g_0 is monotonically increasing in q_i, we discuss how to extend the proposed method to such a more general case that leads to a threshold contour model.

The rest of the paper is organized as follows. Section 2 sets up the model, establishes the identification, and defines the estimator. Section 3 derives the asymptotic properties of the estimators and develops a likelihood ratio test of the threshold function. Section 4 describes how to extend the main model to estimate a threshold contour. Section 5 studies small sample properties of the proposed statistics by Monte Carlo simulations. Section 6 applies the new method to estimate the tipping point function and to determine metropolitan areas. Section 7 concludes this paper with some remarks. The main proofs are in the Appendix, and all the omitted proofs are collected in the supplementary material.

We use the following notations. Let $\rightarrow_p$ denote convergence in probability, $\rightarrow_d$ convergence in distribution, and $\Rightarrow$ weak convergence of the underlying probability measure as $n \rightarrow \infty$. Let $\lfloor r \rfloor$ denote the biggest integer smaller than or equal to r, $\mathbf{1}[E]$ the indicator function of a generic event E, and $\|A\|$ the Euclidean norm of a vector or matrix A. For any set B, let $|B|$ as the cardinality of B.

2 Nonparametric Threshold

We assume spatial processes located on an evenly spaced lattice $\Lambda \subset \mathbb{R}^2$, as in Conley (1999), Conley and Molinari (2007), and Carbon, Francq, and Tran (2007).[1] We consider the threshold regression model given by (1), which is

$$y_i = x_i^\top \beta_0 + x_i^\top \delta_0 \mathbf{1}\left[q_i \leq \gamma_0\left(s_i\right)\right] + u_i,$$

where the observations $\{(y_i, x_i^\top, q_i, s_i)^\top \in \mathbb{R}^{1+\dim(x)+1+1}; i \in \Lambda_n\}$ are a triangular array of real random variables defined on some probability space with Λ_n being a fixed sequence of finite subsets of Λ. In this setup, the cardinality of Λ_n, $n = |\Lambda_n|$, is the sample size and $\sum_{i \in \Lambda_n}$ denotes the summation of all observations. For readability, we postpone the regularity conditions on Λ_n in Assumption A later. The threshold function $\gamma_0 : \mathbb{R} \rightarrow \mathbb{R}$ as well as the regression coefficients $\theta_0 = (\beta_0^\top, \delta_0^\top)^\top \in \mathbb{R}^{2\dim(x)}$ are unknown, and they are the parameters of interest. Since we consider a shrinking threshold effect, the parameter δ_0 is to depend on the sample size n and hence δ_0 and θ_0 should be written as δ_{n0} and θ_{n0}, respectively. However, we write δ_0 and θ_0 for simplicity. We let $\mathcal{Q} \subset \mathbb{R}$ and $\mathcal{S} \subset \mathbb{R}$ denote the supports of q_i and s_i, respectively, which can be unbounded. We also let the space of $\gamma_0(s)$ for any s be a compact set $\Gamma \subset \mathcal{Q}$.

We first establish the identification, which requires the following conditions.

[1] It can be extended to an unevenly spaced lattice as in Bolthausen (1982) and Jenish and Prucha (2009) with substantially more complicated notations (cf. footnote 9 in Conley (1999)).

Assumption ID

(i) $\mathbb{E}[u_i|x_i, q_i, s_i] = 0$ *almost surely.*

(ii) $\mathbb{E}\left[x_i x_i^\top\right] > \mathbb{E}\left[x_i x_i^\top \mathbf{1}[q_i \leq \gamma]\right] > 0$ *for any* $\gamma \in \Gamma$.

(iii) For any $s \in \mathcal{S}$, *there exists* $\varepsilon(s) > 0$ *such that* $\varepsilon(s) < \mathbb{P}(q_i \leq \gamma_0(s_i)|s_i = s) < 1 - \varepsilon(s)$ *and* $\delta_0^\top \mathbb{E}\left[x_i x_i^\top|q_i = q, s_i = s\right]\delta_0 > 0$ *for all* $(q, s) \in \mathcal{Q} \times \mathcal{S}$.

(iv) q_i *is continuously distributed with a conditional density* $f(q|s)$ *satisfying* $0 < C_1 < f(q|s) < C_2 < \infty$ *for all* $(q, s) \in \mathcal{Q} \times \mathcal{S}$ *and some constants* C_1 and C_2.

Assumption ID is mild. The condition (i) excludes endogeneity, and (ii) is the full rank condition to identify β_0 and δ_0. The conditions (ii) and (iii) require that the location of the threshold is not on the boundary of the support of q_i for any $s \in \mathcal{S}$, which is inevitable for identification and has been commonly assumed in the existing threshold literature (e.g., Hansen (2000)). If $\gamma_0(s)$ reaches the boundary of q_i for some $s \in \mathcal{S}$, then no observation exists on one side of the threshold $\gamma_0(s)$, and identification is failed at this s. The second condition in (iii) assumes the coefficient change exists (i.e., $\delta_0 \neq 0$). Note that it does not require $\mathbb{E}\left[x_i x_i^\top|q_i = q, s_i = s\right]$ to be of full rank, and hence q_i or s_i can be one of the elements of x_i (e.g., the threshold autoregressive model by Tong (1983)) or a linear combination of x_i. The condition (iv) requires the conditional density of q_i given any s_i is strictly positive and bounded in Γ.

Under Assumption ID, the following theorem establishes the identification of the semiparametric threshold regression model (1).

Theorem 1 *Under Assumption ID,* $(\beta_0^\top, \delta_0^\top, \gamma_0(s))^\top$ *is the unique minimizer of* $\mathbb{E}[(y_i - x_i^\top\beta - x_i^\top\delta\mathbf{1}[q_i \leq \gamma])^2|s_i = s]$ *over* $(\beta^\top, \delta^\top, \gamma)^\top \in \mathbb{R}^{2\dim(x)} \times \Gamma$ *for each given* $s \in \mathcal{S}$.

Given the identification, we estimate this semiparametric model in two steps. First, for given $s \in \mathcal{S}$, we fix $\gamma_0(s) = \gamma$ and obtain $\widehat{\beta}(\gamma; s)$ and $\widehat{\delta}(\gamma; s)$ by the local constant least squares conditional on γ:

$$(\widehat{\beta}(\gamma; s)^\top, \widehat{\delta}(\gamma; s)^\top)^\top = \arg\min_{\beta,\delta} Q_n(\beta, \delta, \gamma; s), \tag{2}$$

where

$$Q_n(\beta, \delta, \gamma; s) = \sum_{i \in \Lambda_n} \left(y_i - x_i^\top\beta - x_i^\top\delta\mathbf{1}[q_i \leq \gamma]\right)^2 K\left(\frac{s_i - s}{b_n}\right) \tag{3}$$

for some kernel function $K(\cdot)$ and a bandwidth parameter b_n. Then $\gamma_0(s)$ is estimated by

$$\widehat{\gamma}(s) = \arg\min_{\gamma \in \Gamma_n} Q_n(\gamma; s) \tag{4}$$

for given s, where $\Gamma_n = \Gamma \cap \{q_1, \ldots, q_n\}$ and $Q_n(\gamma; s)$ is the concentrated sum of squares defined as

$$Q_n(\gamma; s) = Q_n\left(\widehat{\beta}(\gamma; s), \widehat{\delta}(\gamma; s), \gamma; s\right). \tag{5}$$

Note that, given s, the nonparametric estimator $\widehat{\gamma}(s)$ can be seen as a local version of the standard (constant) threshold regression estimator. Therefore, the computation of (4) requires one-dimensional grid search of the threshold for only n times over Γ_n as in the standard threshold regression estimation. We need to obtain $\widehat{\gamma}(s_i)$ for all $i \in \Lambda_n$ for the second step estimation below.

In the second step, we estimate the parametric components β_0 and δ_0 by least squares. To minimize any potential influence from the first-step estimation error in $\widehat{\gamma}(\cdot)$, we estimate β_0 and $\delta_0^* = \beta_0 + \delta_0$ using the observations that are sufficiently away from the estimated threshold. This is implemented by considering

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i \in \Lambda_n} \left(y_i - x_i^\top \beta \right)^2 \mathbf{1}\left[q_i > \widehat{\gamma}(s_i) + \pi_n \right] \mathbf{1}[s_i \in \mathcal{S}_0], \tag{6}$$

$$\widehat{\delta}^* = \arg\min_{\delta^*} \sum_{i \in \Lambda_n} \left(y_i - x_i^\top \delta^* \right)^2 \mathbf{1}\left[q_i < \widehat{\gamma}(s_i) - \pi_n \right] \mathbf{1}[s_i \in \mathcal{S}_0] \tag{7}$$

for some constant $\pi_n > 0$ satisfying $\pi_n \to 0$ as $n \to \infty$, which is defined later. The change size δ can be estimated as $\widehat{\delta} = \widehat{\delta}^* - \widehat{\beta}$. Note that the support of s_i, $\mathcal{S}$, is not necessarily bounded. To avoid any potential technical complexity in the second-step estimator, however, we focus on the estimates $\widehat{\gamma}(s)$ over some compact subset of the support $\mathcal{S}_0 \subset \mathcal{S}$.

For the asymptotic behavior of the threshold estimator, the existing literature typically assumes martingale difference arrays (e.g., Hansen (2000) and Lee, Liao, Seo, and Shin (2021)) or random samples (e.g., Yu (2012) and Yu and Fan (2021)). In this paper, we allow for cross-sectional dependence by considering spatial α-mixing processes as in Bolthausen (1982) and Conley (1999). More precisely, for any indices (or locations) $i, j \in \Lambda$, we define the metric $\lambda(i,j) = \max_{1 \le \ell \le \dim(\Lambda)} |i_\ell - j_\ell|$ and the corresponding norm $\max_{1 \le \ell \le \dim(\Lambda)} |i_\ell|$, where i_ℓ denotes the ℓth component of i. The distance of any two subsets $\Lambda_1, \Lambda_2 \subset \Lambda$ is defined as $\lambda(\Lambda_1, \Lambda_2) = \inf\{\lambda(i,j) : i \in \Lambda_1, j \in \Lambda_2\}$. We let $\mathcal{F}_\Lambda$ be the σ-algebra generated by a random sequence $\left(x_i^\top, q_i, s_i, u_i\right)^\top$ for $i \in \Lambda$ and define the spatial α-mixing coefficient as

$$\alpha_{k,l}(m) = \sup\left\{ |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| : A \in \mathcal{F}_{\Lambda_1},\, B \in \mathcal{F}_{\Lambda_2},\, \lambda(\Lambda_1, \Lambda_2) \ge m \right\}, \tag{8}$$

where $|\Lambda_1| \le k$ and $|\Lambda_2| \le l$. Without loss of generality, we assume $\alpha_{k,l}(0) = 1$ and $\alpha_{k,l}(m)$ is monotonically decreasing in m for all k and l.

The following conditions are imposed for deriving the asymptotic properties of our semiparametric estimators. Let $f(q,s)$ be the joint density function of (q_i, s_i) and

$$\begin{aligned} D(q,s) &= \mathbb{E}[x_i x_i^\top | (q_i, s_i) = (q,s)], \\ V(q,s) &= \mathbb{E}[x_i x_i^\top u_i^2 | (q_i, s_i) = (q,s)]. \end{aligned}$$

Assumption A

(i) The lattice $\Lambda_n \subset \mathbb{R}^2$ is infinitely countable; for any $i, j \in \Lambda_n$, $\lambda(i,j) \geq \lambda_0 > 1$; and $\lim_{n\to\infty} |\partial\Lambda_n|/n = 0$, where $\partial\Lambda_n = \{i \in \Lambda_n : \exists j \notin \Lambda_n \text{ with } \lambda(i,j) = 1\}$.

(ii) $\delta_0 = c_0 n^{-\epsilon}$ for some $c_0 \neq 0$ and $\epsilon \in (0, 1/2)$; $\left(c_0^\top, \beta_0^\top\right)^\top$ belongs to some compact subset of $\mathbb{R}^{2\dim(x)}$.

(iii) $\left(x_i^\top, q_i, s_i, u_i\right)^\top$ is strictly stationary and α-mixing with the mixing coefficient $\alpha_{k,l}(m)$ defined in (8), which satisfies that for all k and *l, $\alpha_{k,l}(m) \leq C_1 \exp(-C_2 m)$ for some positive constants C_1 and C_2.*

(iv) $0 < \mathbb{E}\left[u_i^2 | x_i, q_i, s_i\right] < \infty$ almost surely.

(v) There exist some finite constants $\varphi > 0$ and *$C > 0$ such that $\mathbb{E}[||x_i x_i^\top||^{2(2+\varphi)}|(q_i, s_i) = (q, s)] < C$ and $\mathbb{E}[||x_i u_i||^{2(2+\varphi)}|(q_i, s_i) = (q, s)] < C$ uniformly in (q, s).*

(vi) $\gamma_0 : \mathcal{S} \mapsto \Gamma$ is a twice continuously differentiable function with bounded derivatives.

(vii) $D(q,s)$, $V(q,s)$, and $f(q,s)$ are uniformly bounded in (q,s), continuous in q, and twice continuously differentiable in s with bounded derivatives. For any $i, j \in \Lambda_n$, the joint density of $(q_i, q_j, s_i, s_j)^\intercal$ and $\mathbb{E}[||x_i x_j^\top u_i u_j|||q_i, q_j, s_i, s_j]$ are uniformly bounded almost surely and continuously differentiable in all components.

(viii) $c_0^\top D(\gamma_0(s), s)\, c_0 > 0$, $c_0^\top V(\gamma_0(s), s)\, c_0 > 0$, and $f(\gamma_0(s), s) > 0$ for all $s \in \mathcal{S}$.

(ix) As $n \to \infty$, $b_n \to 0$, $n^{1-2\epsilon} b_n / \log n \to \infty$, $\log n / n b_n^2 \to 0$, and $n^{1/(1+\varphi)} b_n \to \infty$ for $\varphi > 0$ given in (v).

(x) $K(\cdot)$ is a positive second-order kernel, which is Lipschitz, symmetric around zero, and non-increasing on $\mathbb{R}^+$. It also satisfies $\int K(v)\, dv = 1$, $\int K^\ell(v)\, dv < \infty$, $\int v^2 K^\ell(v)\, dv < \infty$ for $\ell \leq 2(2+\varphi)$ and $\varphi > 0$ given in (v).

Assumption A is mild and common in the existing literature. In particular, the condition (i) is the same as in Bolthausen (1982) and Jenish and Prucha (2009) to define the latent random field, which assumes all the elements in Λ_n are located at distances at least λ_0 from each other. The distance λ_0 can be any strictly positive value and we impose $\lambda_0 > 1$ without loss of generality. The condition (ii) adopts the widely used shrinking change size setup as in Bai (1997), Bai and Perron (1998), and Hansen (2000) to obtain a limiting distribution that is free of nuisance parameters. In contrast, a constant change size (when $\epsilon = 0$) leads to a complicated asymptotic distribution of the threshold estimator, which depends on nuisance parameters (e.g., Chan (1993)). The condition (iii) is required to establish the maximal inequality and uniform convergence in a spatially dependent random field. We impose a stronger condition than Jenish and Prucha (2009) to obtain the maximal inequality uniformly over γ and s. We could weaken this condition such that $\alpha_{k,l}(m)$ decays at a polynomial rate (e.g., $\alpha_{k,l}(m) \leq Cm^{-r}$ for some $r > 8$ and a constant C as in Carbon, Francq, and Tran (2007)) if we impose higher moment restrictions in the condition (v). However, this exponential decay rate simplifies the technical proofs. The conditions (iv) to (viii) are similar to Assumption 1 of Hansen (2000), where we impose additional moment restrictions to control for spatial dependence. The condition (ix) imposes restrictions on the bandwidth b_n, which depends on ϵ and φ. The condition (x) holds for many commonly used kernel functions including the Gaussian and the uniform kernels.

We assume $\gamma_0(\cdot)$ to be a function from $\mathcal{S}$ to Γ in Assumption A-(vi), which is not necessarily one-to-one. For this reason, sample splitting based on $\mathbf{1}[q_i \leq \gamma_0(s_i)]$ can be different from that based on $\mathbf{1}[s_i \geq \breve{\gamma}_0(q_i)]$ for some function $\breve{\gamma}_0(\cdot)$. Instead of restricting $\gamma_0(\cdot)$ to be one-to-one in this paper, we presume that one knows which variables should be respectively assigned as q_i and s_i from the context. Alternatively, we can consider a function $g_0(q, s)$, which is monotonically increasing in q for any s, and $\mathbf{1}[q_i \leq \gamma_0(s_i)]$ can be viewed as a special case of $\mathbf{1}[g_0(q_i, s_i) \leq 0]$. We discuss such extension to identify a threshold contour in Section 4.

3 Asymptotic Results

We first obtain the asymptotic properties of $\widehat{\gamma}(s)$. The following theorem derives the pointwise consistency and the pointwise rate of convergence of $\widehat{\gamma}(s)$ at the interior points of $\mathcal{S}$, say in $\mathcal{S}_0 \subset \mathcal{S}$.

Theorem 2 *For a given $s \in \mathcal{S}_0$, under Assumptions ID and A, $\widehat{\gamma}(s) \rightarrow_p \gamma_0(s)$ as $n \rightarrow \infty$. Furthermore,*

$$\widehat{\gamma}(s) - \gamma_0(s) = O_p\left(\frac{1}{n^{1-2\epsilon} b_n}\right) \tag{9}$$

provided that $n^{1-2\epsilon} b_n^2$ does not diverge.

The pointwise rate of convergence of $\widehat{\gamma}(s)$ depends on two parameters, ϵ and b_n. It is decreasing in ϵ like the parametric (constant) threshold case: a larger ϵ reduces the threshold effect $\delta_0 = c_0 n^{-\epsilon}$ and hence decreases the effective sampling information on the threshold. Since we estimate $\gamma_0(\cdot)$ using the kernel estimation method, the rate of convergence depends on the bandwidth b_n as well. As in the standard kernel estimator case, a smaller bandwidth decreases the effective local sample size, which reduces the precision of the estimator $\widehat{\gamma}(s)$. Therefore, in order to have a sufficiently fast rate of convergence, we need to choose b_n large enough when the threshold effect δ_0 is expected to be small (i.e., when ϵ is close to 1/2).

Unlike the standard kernel estimator, (9) does not manifests the typical bias-variance trade-off in the local constant estimator $\widehat{\gamma}(s)$. Hence, it seems like that we could improve the rate of convergence by choosing a larger bandwidth b_n. However, b_n cannot be chosen too large to result in $n^{1-2\epsilon} b_n^2 \rightarrow \infty$, under which $n^{1-2\epsilon} b_n(\widehat{\gamma}(s) - \gamma_0(s))$ is no longer $O_p(1)$ in Theorem 3 below. In fact, the reason that we cannot see the bias-variance trade-off in (9) is because we restrict that $n^{1-2\epsilon} b_n^2$ does not diverge. Since we do not have the closed-form expression of

$\widehat{\gamma}(s) = \arg\min_{\gamma\in\Gamma_n} Q_n(\gamma; s)$ in (4), we obtain the convergence rate of $\widehat{\gamma}(s)$ indirectly through the convergence rate of $|Q_n(\widehat{\gamma}(s); s) - Q_n(\gamma_0(s); s)|$. Hence we cannot readily obtain the explicit stochastic orders of the bias and the variance of $\widehat{\gamma}(s)$ as in the standard nonparametric analysis. However, we can find that the components determining the bias and the variance of $\widehat{\gamma}(s)$ are, respectively, $O(b_n)$ and $O((n^{1-2\epsilon}b_n)^{-2})$ in the proof of Theorems 2 and 3. In particular, the $O(b_n)$ bias corresponds to the boundary bias of the typical local constant estimator with a bounded support. This bias is also expected in our case because we estimate the threshold $\gamma_0(s)$ using the one-side observations (q_i, s_i) such that $q_i \leq \gamma_0(s_i)$. Assuming $n^{1-2\epsilon}b_n^2 = (n^{1-2\epsilon}b_n)b_n \to \varrho < \infty$ is equivalent to balancing stochastic orders of the squared bias and the variance, and hence resulting in the optimal bandwidth choice as in the standard local constant estimation analysis.

More precisely, under the restriction $n^{1-2\epsilon}b_n^2 \to \varrho < \infty$, we can find the largest and optimal bandwidth as $b_n^* = n^{-(1-2\epsilon)/2}c^*$ for some constant $0 < c^* < \infty$, which yields the fastest pointwise rate of convergence of $\widehat{\gamma}(s)$ as $n^{-(1-2\epsilon)/2}$. Note that, when the change size δ_0 shrinks very slowly with n (i.e., ϵ is close to 0), the rate of convergence of $\widehat{\gamma}(\cdot)$ becomes close to $n^{-1/2}$. This $\sqrt{n}$-rate can be obtained in the standard kernel regression if the unknown function is infinitely differentiable, while we only require the second-order differentiability of $\gamma_0(\cdot)$.

The next theorem derives the limiting distribution of $\widehat{\gamma}(s)$. We let $W(\cdot)$ be a two-sided Brownian motion defined as in Hansen (2000):

$$W(r) = W_1(-r)\mathbf{1}[r<0] + W_2(r)\mathbf{1}[r>0], \tag{10}$$

where $W_1(\cdot)$ and $W_2(\cdot)$ are independent standard Brownian motions on $[0,\infty)$.

Theorem 3 *Under Assumptions ID and A, for a given $s \in \mathcal{S}_0$, if $n^{1-2\epsilon}b_n^2 \to \varrho \in (0,\infty)$,*

$$n^{1-2\epsilon}b_n(\widehat{\gamma}(s) - \gamma_0(s)) \to_d \xi(s) \arg\max_{r\in\mathbb{R}} (W(r) + \mu(r,\varrho;s)) \tag{11}$$

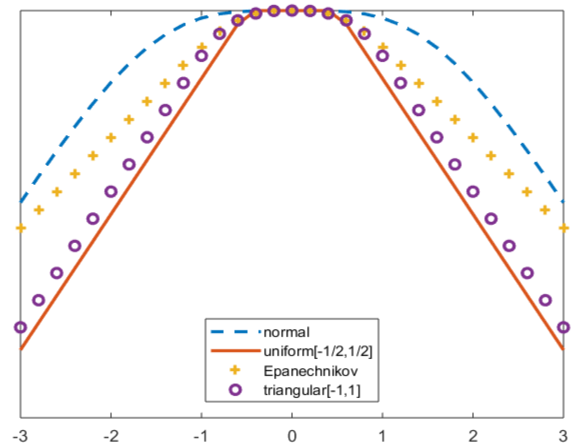
as $n\to\infty$, where

$$\begin{aligned}
\mu(r,\varrho;s) &= -|r|\,\psi_0(r,\varrho;s) + \frac{\varrho|\dot{\gamma}_0(s)|}{\xi(s)}\psi_1(r,\varrho;s),\\
\psi_j(r,\varrho;s) &= \int_0^{|r|\xi(s)/(\varrho|\dot{\gamma}_0(s)|)} t^j K(t)\,dt \quad \text{for } j=0,1,\\
\xi(s) &= \frac{\kappa_2 c_0^\top V(\gamma_0(s),s)c_0}{\left(c_0^\top D(\gamma_0(s),s)c_0\right)^2 f(\gamma_0(s),s)}
\end{aligned}$$

with $\kappa_2 = \int K(v)^2dv$ and $\dot{\gamma}_0(s)$ is the first derivative of γ_0 at s. Furthermore,

$$\mathbb{E}\left[\arg\max_{r\in\mathbb{R}}(W(r)+\mu(r,\varrho;s))\right] = 0.$$

Figure 1: Drift function $\mu(r, \varrho; s)$ for different kernels (color online)



The drift term $\mu(r, \varrho; s)$ in (11) depends on the constant $\varrho = \lim_{n\to\infty} n^{1-2\epsilon} b_n^2$ and the steepness of $\gamma_0(\cdot)$ at s, $|\dot{\gamma}_0(s)|$. It is important to note that having this drift term in the limiting expression does not mean that the limiting distribution of $\widehat{\gamma}(s)$ has a non-zero mean, even when we use the optimal bandwidth satisfying $n^{1-2\epsilon} b_n^2 \to \varrho \in (0, \infty)$. This is because the drift function $\mu(r, \varrho; s)$ is symmetric about zero and hence the limiting random variable $\arg\max_{r\in\mathbb{R}} (W(r) + \mu(r, \varrho; s))$ is mean zero.[2] Figure 1 depicts the drift function $\mu(r, \varrho; s)$ for various kernels when $\xi(s)/(\varrho|\dot{\gamma}_0(s)|) = 1$.

Since the limiting distribution in (11) depends on unknown components in the drift term, like ϱ and $\dot{\gamma}_0(s)$, it is hard to use this result for further inference. We instead suggest undersmoothing for practical use. More precisely, if we suppose $n^{1-2\epsilon} b_n^2 \to 0$ as $n \to \infty$, then the limiting distribution in (11) simplifies to[3]

$$n^{1-2\epsilon} b_n (\widehat{\gamma}(s) - \gamma_0(s)) \to_d \xi(s) \arg\max_{r\in\mathbb{R}} \left(W(r) - \frac{|r|}{2} \right)$$

as $n \to \infty$, which appears the same as in the parametric case in Hansen (2000) except for the scaling factor $n^{1-2\epsilon} b_n$. The distribution of $\arg\max_{r\in\mathbb{R}} (W(r) - |r|/2)$ is known (e.g., Bhattacharya and Brockwell (1976) and Bai (1997)), which is also described in Hansen (2000, p.581). The $\xi(s)$ term determines the scale of the distribution at given s in the way that it increases in the conditional variance $\mathbb{E}[u_i^2|x_i, q_i, s_i]$ and decreases in the size of the threshold constant c_0 and the density of (q_i, s_i) near the threshold.

[2] In general, we can show that the random variable $\arg\max_{r\in\mathbb{R}} (W(r) + \mu(r))$ always has mean zero if $\mu(r)$ is a non-random function that is symmetric about zero and monotonically decreasing fast enough. This result might be of independent research interest and is summarized in Lemma A.11 in the Appendix.

[3] We let $\psi_0(r, 0; s) = \int_0^\infty K(t)\, dt = 1/2$ and $\psi_1(r, 0; s) = \int_0^\infty tK(t)\, dt < \infty$.

For inference of $\gamma_0(s)$ given any $s \in \mathcal{S}_0$, we can consider a pointwise likelihood ratio test statistic for

$$H_0 : \gamma_0 (s) = \gamma_* (s) \quad \text{against} \quad H_1 : \gamma_0 (s) \neq \gamma_* (s) , \tag{12}$$

which is given as

$$LR_n(s) = \sum_{i \in \Lambda_n} \frac{Q_n (\gamma_* (s) , s) - Q_n (\widehat{\gamma} (s) , s)}{Q_n (\widehat{\gamma} (s) , s)} K \left(\frac{s_i - s}{b_n} \right). \tag{13}$$

The following corollary obtains the limiting null distribution of this test statistic. By inverting the likelihood ratio statistic, we can form a pointwise confidence interval for $\gamma_0 (s)$.

Corollary 1 *Suppose $n^{1-2\epsilon} b_n^2 \to 0$ as $n \to \infty$. Under the same condition in Theorem 3, for any fixed $s \in \mathcal{S}_0$, the test statistic in (13) satisfies*

$$LR_n(s) \to_d \xi_{LR} (s) \max_{r \in \mathbb{R}} (2W (r) - |r|) \tag{14}$$

as $n \to \infty$ under the null hypothesis (12), where

$$\xi_{LR} (s) = \frac{\kappa_2 c_0^\top V (\gamma_0 (s) , s) c_0}{\sigma^2(s) c_0^\top D (\gamma_0 (s) , s) c_0}$$

with $\sigma^2(s) = \mathbb{E} \left[u_i^2 | s_i = s \right]$ and $\kappa_2 = \int K(v)^2 dv$.

When $\mathbb{E}[u_i^2 | x_i, q_i, s_i] = \mathbb{E}[u_i^2 | s_i]$, which is the case of local conditional homoskedasticity, the scale parameter $\xi_{LR} (s)$ is simplified as κ_2, and hence the limiting null distribution of $LR_n(s)$ becomes free of nuisance parameters and the same for all $s \in \mathcal{S}_0$. Though this limiting distribution is still nonstandard, the critical values in this case can be simulated using the same method as Hansen (2000, p.582) with the scale adjusted by κ_2. More precisely, since the distribution function of $\zeta = \max_{r \in \mathbb{R}} (2W (r) - |r|)$ is given as $\mathbb{P}(\zeta \leq z) = (1 - \exp(-z/2))^2 \mathbf{1} [z \geq 0]$, the distribution function of $\zeta^* = \kappa_2 \zeta$ is $\mathbb{P}(\zeta^* \leq z) = (1 - \exp(-z/2\kappa_2))^2 \mathbf{1} [z \geq 0]$, where ζ^* is the limiting random variable of $LR_n(s)$ given in (14) under the local conditional homoskedasticity. For instance, the critical values are reported in Table 1 when the Gaussian kernel is used, where $\kappa_2 = (2\sqrt{\pi})^{-1}$ is about 0.2821 in this case.

In general, we can estimate $\xi_{LR} (s)$ by

$$\widehat{\xi}_{LR} (s) = \frac{\kappa_2 \widehat{\delta}^\top \widehat{V} (\widehat{\gamma} (s) , s) \widehat{\delta}}{\widehat{\sigma}^2(s) \widehat{\delta}^\top \widehat{D} (\widehat{\gamma} (s) , s) \widehat{\delta}},$$

Table 1: Simulated Critical Values of the LR Test (Gaussian Kernel)

$\mathbb{P}(\zeta^* > cv)$	0.800	0.850	0.900	0.925	0.950	0.975	0.990
cv	1.268	1.439	1.675	1.842	2.074	2.469	2.988

Note: ζ^* is the limiting distribution of $LR_n(s)$ under the local conditional homoskedasticity. The Gaussian kernel is used.

where $\widehat{\delta}$ is from (6) and (7), and $\widehat{\sigma}^2(s)$, $\widehat{D}\left(\widehat{\gamma}\left(s\right), s\right)$, and $\widehat{V}\left(\widehat{\gamma}\left(s\right), s\right)$ are the standard Nadaraya-Watson estimators at $s \in \mathcal{S}_0$. In particular, we let $\widehat{\sigma}^2(s) = \sum_{i \in \Lambda_n} \omega_{1i}(s)\widehat{u}_i^2$ with $\widehat{u}_i = y_i - x_i^\top \widehat{\beta} - x_i^\top \widehat{\delta} \mathbf{1}\left[q_i \leq \widehat{\gamma}\left(s_i\right)\right]$,

$$\widehat{D}\left(\widehat{\gamma}\left(s\right), s\right) = \sum_{i \in \Lambda_n} \omega_{2i}(s) x_i x_i^\top, \text{ and } \widehat{V}\left(\widehat{\gamma}\left(s\right), s\right) = \sum_{i \in \Lambda_n} \omega_{2i}(s) x_i x_i^\top \widehat{u}_i^2,$$

where

$$\omega_{1i}(s) = \frac{K\left((s_i - s)/b_n\right)}{\sum_{j \in \Lambda_n} K\left((s_j - s)/b_n\right)} \text{ and } \omega_{2i}(s) = \frac{\mathbb{K}\left((q_i - \widehat{\gamma}\left(s\right))/b_n', (s_i - s)/b_n''\right)}{\sum_{j \in \Lambda_n} \mathbb{K}\left((q_j - \widehat{\gamma}\left(s\right))/b_n', (s_j - s)/b_n''\right)}$$

for some bivariate kernel function $\mathbb{K}(\cdot,\cdot)$ and bandwidth parameters (b_n', b_n'').

Finally, we show the $\sqrt{n}$-consistency of $\widehat{\beta}$ and $\widehat{\delta}^*$ in (6) and (7). For this purpose, we first obtain the uniform rate of convergence of $\widehat{\gamma}\left(s\right)$.

Theorem 4 *Under Assumptions ID and A,*

$$\sup_{s \in \mathcal{S}_0} \left|\widehat{\gamma}\left(s\right) - \gamma_0\left(s\right)\right| = O_p\left(\frac{\log n}{n^{1-2\epsilon} b_n}\right)$$

provided that $n^{1-2\epsilon} b_n^2$ *does not diverge.*

Apparently, the uniform consistency of $\widehat{\gamma}\left(s\right)$ follows when $\log n/(n^{1-2\epsilon} b_n) \to 0$ as $n \to \infty$. Based on this uniform convergence, the following theorem derives the joint limiting distribution of $\widehat{\beta}$ and $\widehat{\delta}^*$. We let $\widehat{\theta}^* = (\widehat{\beta}^\top, \widehat{\delta}^{*\top})^\top$ and $\theta_0^* = (\beta_0^\top, \delta_0^{*\top})^\top$.

Theorem 5 *Suppose the conditions in Theorem 4 hold. If we let* $\pi_n > 0$ *such that* $\pi_n \to 0$ *and* $\{\log n/(n^{1-2\epsilon} b_n)\}/\pi_n \to 0$ *as* $n \to \infty$*, we have*

$$\sqrt{n}\left(\widehat{\theta}^* - \theta_0^*\right) \to_d \mathcal{N}\left(0, \Sigma_X^{*-1} \Omega^* \Sigma_X^{*-1}\right) \tag{15}$$

as $n \to \infty$, where

$$\Sigma_X^* = \begin{bmatrix} \mathbb{E}\left[x_i x_i^\top \mathbf{1}_i^+\right] & 0 \\ 0 & \mathbb{E}\left[x_i x_i^\top \mathbf{1}_i^-\right] \end{bmatrix} \quad \text{and} \quad \Omega^* = \lim_{n\to\infty} \frac{1}{n} Var \begin{bmatrix} \sum_{i\in\Lambda_n} x_i u_i \mathbf{1}_i^+ \\ \sum_{i\in\Lambda_n} x_i u_i \mathbf{1}_i^- \end{bmatrix}$$

with $\mathbf{1}_i^+ = \mathbf{1}[q_i > \gamma_0(s_i)]\mathbf{1}[s_i \in \mathcal{S}_0]$ and $\mathbf{1}_i^- = \mathbf{1}[q_i \le \gamma_0(s_i)]\mathbf{1}[s_i \in \mathcal{S}_0]$.

For the second-step estimator $\widehat{\theta}^*$, we use (6) and (7), instead of the conventional plug-in estimator, say $\arg\min_{\beta,\delta} \sum_{i\in\Lambda_n} (y_i - x_i^\top \beta - x_i^\top \delta \mathbf{1}[q_i \le \widehat{\gamma}(s_i)])^2 \mathbf{1}[s_i \in \mathcal{S}_0]$. The reason is that the first-step nonparametric estimator $\widehat{\gamma}(\cdot)$ may not be asymptotically orthogonal to the second-step estimator. Unlike the standard semiparametric literature (e.g., Assumption N(c) in Andrews (1994)), the asymptotic effect of $\widehat{\gamma}(s)$ to the second-step estimation is not easily derived due to the discontinuity. The new estimation idea above, however, only uses the observations that are little influenced by the estimation error in the first step to achieve asymptotic orthogonality. As we verify in Lemma A.15 in the Appendix, this is done by choosing a large enough π_n in (6) and (7) such that the observations that are included in the second step are outside the uniform convergence bound of $\sup_{s\in\mathcal{S}_0} |\widehat{\gamma}(s) - \gamma_0(s)|$. Thanks to the threshold regression structure, we can estimate the parameters on each side of the threshold even using these subsamples. Meanwhile, we also want $\pi_n \to 0$ fast enough to include more observations. By doing so, though we lose some efficiency in finite samples, we can derive the asymptotic normality of $\widehat{\theta} = (\widehat{\beta}^\top, \widehat{\delta}^\top)^\top$ that has zero mean and achieves the same asymptotic variance as if $\gamma_0(\cdot)$ were known.

By the delta method, Theorem 5 readily yields the limiting distribution of $\widehat{\theta} = (\widehat{\beta}^\top, \widehat{\delta}^\top)^\top$ as

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \to_d \mathcal{N}\left(0, \Sigma_X^{-1}\Omega\Sigma_X^{-1}\right) \quad \text{as } n \to \infty, \tag{16}$$

where

$$\Sigma_X = \mathbb{E}\left[z_i z_i^\top \mathbf{1}\left[s_i \in \mathcal{S}_0\right]\right] \quad \text{and} \quad \Omega = \lim_{n\to\infty} \frac{1}{n} Var\left[\sum_{i\in\Lambda_n} z_i u_i \mathbf{1}\left[s_i \in \mathcal{S}_0\right]\right]$$

with $z_i = [x_i^\top, x_i^\top \mathbf{1}\left[q_i \le \gamma_0(s_i)\right]]^\top$. The asymptotic variance expressions in (15) and (16) allow for cross-sectional dependence as they use the long-run variances (LRV) Ω^* and Ω. We can estimate the LRV by the robust estimator developed by Conley and Molinari (2007) using $\widehat{u}_i = (y_i - x_i^\top \widehat{\beta} - x_i^\top \widehat{\delta} \mathbf{1}[q_i \le \widehat{\gamma}(s_i)])\mathbf{1}[s_i \in \mathcal{S}_0]$. The terms Σ_X^* and Σ_X can be estimated by their sample analogues.

4 Threshold Contour

The threshold model (1) can be generalized to estimate a nonparametric contour threshold model:

$$y_i = x_i^\top \beta_0 + x_i^\top \delta_0 \mathbf{1}\left[g_0\left(q_i, s_i\right) \leq 0\right] + u_i,$$

where the unknown function $g_0 : \mathcal{Q} \times \mathcal{S} \mapsto \mathbb{R}$ determines the threshold contour on a random field that yields sample splitting. An interesting example includes identifying an unknown closed boundary over the map, such as a city boundary, and an area of a disease outbreak or airborne pollution. In social science, it can identify a group boundary or a region in which the agents share common demographic, political, or economic characteristics.

To relate this generalized form to the original threshold model (1), we suppose there exists a known center at (q_i^*, s_i^*) such that $g_0\left(q_i^*, s_i^*\right) < 0$. Without loss of generality, we can normalize (q_i^*, s_i^*) to be $(0, 0)$ and re-center the original location variables (q_i, s_i) accordingly. In addition, we define the radius distance l_i and angle a_i° of the ith observation relative to the origin as

$$\begin{aligned} l_i &= (q_i^2 + s_i^2)^{1/2}, \\ a_i^\circ &= \bar{a}_i^\circ \mathbf{I}_i + \left(180^\circ - \bar{a}_i^\circ\right) \mathbf{II}_i + \left(180^\circ + \bar{a}_i^\circ\right) \mathbf{III}_i + \left(360^\circ - \bar{a}_i^\circ\right) \mathbf{IV}_i, \end{aligned}$$

where $\bar{a}_i^\circ = \arctan\left(|q_i/s_i|\right)$, and each of $(\mathbf{I}_i, \mathbf{II}_i, \mathbf{III}_i, \mathbf{IV}_i)$ respectively denotes the indicator that the ith observation locates in the first, second, third, and forth quadrant.

We suppose that there is only one threshold at any angle and the threshold contour is star-shaped.[4] For each chosen angle $a^\circ \in [0^\circ, 360^\circ)$, we rotate the original coordinate counterclockwise and implement the estimation in (5) only using the observations in the first two quadrants after rotation. It will ensure that the threshold mapping after rotation is a well-defined function.

In particular, the angle relative to the origin is $a_i^\circ - a^\circ$ after rotating the coordinate by a° degrees counterclockwise, and the new location after the rotation is given as $\left(q_i\left(a^\circ\right), s_i\left(a^\circ\right)\right)$, where

$$\begin{pmatrix} q_i\left(a^\circ\right) \\ s_i\left(a^\circ\right) \end{pmatrix} = \begin{pmatrix} q_i \cos\left(a^\circ\right) - s_i \sin\left(a^\circ\right) \\ s_i \cos\left(a^\circ\right) + q_i \sin\left(a^\circ\right) \end{pmatrix}.$$

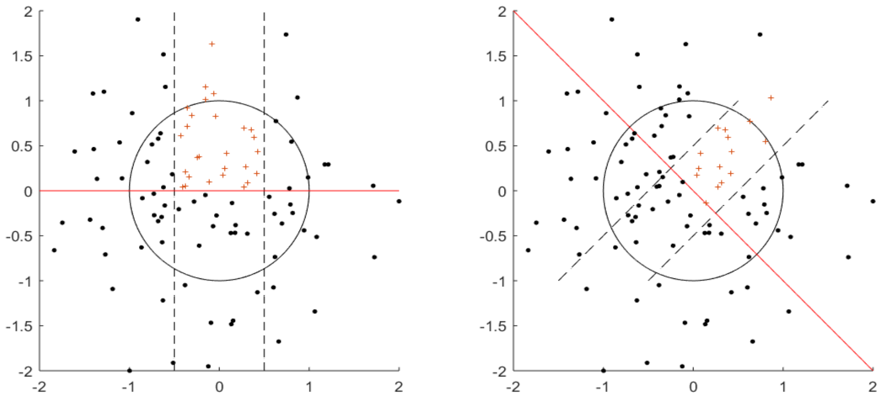
After this rotation, we estimate the following nonparametric threshold model:

$$y_i = x_i^\top \beta_0 + x_i^\top \delta_0 \mathbf{1}\left[q_i\left(a^\circ\right) \leq \gamma_{a^\circ}\left(s_i\left(a^\circ\right)\right)\right] + u_i \tag{17}$$

using only the observations i satisfying $q_i\left(a^\circ\right) \geq 0$ and in the neighborhood of $s_i\left(a^\circ\right) = 0$, where $\gamma_{a^\circ}(\cdot)$ is the unknown threshold curve as in the original model (1) on the a°-degree-rotated coordinate plane. Such reparametrization guarantees that $\gamma_{a^\circ}(\cdot)$ is always positive and it is

[4] This assumption implicitly depends on the choice of origin and rotation, which is a common problem in directional data analysis. We leave this for future research and thank an anonymous referee for pointing this out.

Figure 2: Illustration of rotation (color online)



estimated at $s_i(a^\circ) = 0$. Figure 2 illustrates the idea of such rotation and pointwise estimation over a bounded support so that only the red cross points are included for estimation at different angles. Thus, the estimation and inference procedures developed in the previous sections are directly applicable, though we expect some efficiency loss as we only use the subsample with $q_i(a^\circ) \geq 0$ at each a°.

5 Monte Carlo Experiments

We examine the small sample performance of the semiparametric threshold regression estimator by Monte Carlo simulations. We generate n draws from

$$y_i = x_i^\top \beta_0 + x_i^\top \delta_0 \mathbf{1}\left[q_i \leq \gamma_0(s_i)\right] + u_i, \tag{18}$$

where $x_i = (1, x_{2i})^\top$ and $x_{2i} \in \mathbb{R}$. We let $\beta_0 = (\beta_{10}, \beta_{20})^\top = (0,0)^\top$ and consider three different values of $\delta_0 = (\delta_{10}, \delta_{20})^\top = (\delta, \delta)^\top$ with $\delta = 1, 2, 3, 4$. For the threshold function, we let $\gamma_0(s) = \cos(\pi s)/2$. The supplementary material contains results with other specifications. The findings are similar to those presented in this section.

We consider the cross-sectional dependence structure in $(x_{2i}, q_i, s_i, u_i)^\top$ as follows:

$$\begin{cases} (q_i, s_i)^\top \sim iid\mathcal{N}(0, I_2); \\ x_{2i}|\,(q_i, s_i) \sim iid\mathcal{N}\left(0, (1 + \rho\left(s_i^2 + q_i^2\right))^{-1}\right); \\ \underline{\mathbf{u}}|\{(x_i, q_i, s_i)\}_{i=1}^n \sim \mathcal{N}(0, \Sigma), \end{cases} \tag{19}$$

where $\underline{\mathbf{u}} = (u_1, \ldots, u_n)^\top$. The (i,j)th element of Σ is $\Sigma_{ij} = \rho^{\lfloor \ell_{ij} n \rfloor} \mathbf{1}[\ell_{ij} < m/n]$, where $\ell_{ij} = \{(s_i - s_j)^2 + (q_i - q_j)^2\}^{1/2}$ is the L^2-distance between the ith and jth observations. The diagonal

Table 2: Bias, RMSE, and Rej. Prob. of the LR Test with i.i.d. Data

	$s = 0.0$				$s = 0.5$				$s = 1.0$			
$n\backslash\delta$	1	2	3	4	1	2	3	4	1	2	3	4
	Bias											
100	-0.42	-0.29	-0.23	-0.22	-0.05	-0.06	-0.14	-0.11	0.41	0.28	0.23	0.20
200	-0.36	-0.21	-0.13	-0.12	-0.06	-0.03	-0.08	-0.06	0.35	0.20	0.14	0.10
500	-0.25	-0.09	-0.06	-0.05	-0.01	-0.02	-0.03	-0.02	0.31	0.12	0.06	0.01
	RMSE											
100	0.51	0.28	0.18	0.16	0.27	0.18	0.11	0.08	0.46	0.33	0.27	0.23
200	0.42	0.20	0.10	0.06	0.27	0.13	0.08	0.05	0.44	0.26	0.19	0.14
500	0.32	0.09	0.03	0.02	0.21	0.06	0.03	0.02	0.40	0.18	0.10	0.06
	Rej. Prob. of the LR test											
100	0.16	0.09	0.08	0.08	0.18	0.13	0.14	0.15	0.28	0.18	0.15	0.13
200	0.11	0.05	0.05	0.02	0.12	0.09	0.11	0.15	0.20	0.12	0.08	0.05
500	0.06	0.03	0.02	0.02	0.09	0.07	0.11	0.14	0.10	0.06	0.03	0.02

Note: Entries are bias and root mean squared error (RMSE) of the estimator $\hat{\gamma}(s)$ and rejection probabilities of the LR test (13) when data are generated from (18) with $\gamma_0(s) = \cos(\pi s)/2$. The dependence structure is given in (19) with $\rho = 0$. The significance level is 5% and the results are based on 1000 simulations.

elements of Σ are normalized as $\Sigma_{ii} = 1$. This m-dependent setup follows from the Monte Carlo experiment in Conley and Molinari (2007) in the sense that each unit can be cross-sectionally correlated with at most $2m^2$ observations. Within the m distance, the dependence decays at a rate of $\rho^{\lfloor \ell_{ij} n \rfloor}$. The parameter ρ describes the strength of cross-sectional dependence in the way that a larger ρ leads to stronger dependence relative to the unit standard deviation. We consider the sample size $n = 100$, 200, and 500, and set $\mathcal{S}_0$ to include the middle 70% observations of s_i, which is roughly $[-1, 1]$ since we generate s_i from the standard normal in (19).[5]

First, Tables 2 and 3 report the bias and root mean squared error (RMSE) of $\widehat{\gamma}(s)$ as well as the small sample rejection probabilities of the LR test in (13) for $H_0 : \gamma_0(s) = \cos(\pi s)/2$ against $H_1 : \gamma_0(s) \neq \cos(\pi s)/2$ at three different locations $s = 0.0$, 0.5, and 1.0. The nominal level is 5%. In particular, Table 2 examines the case with no cross-sectional dependence ($\rho = 0$), while Table 3 examines the case with cross-sectional dependence whose dependence decays slowly with $\rho = 1$ and $m = 10$. We normalize s_i and q_i to have zero mean and unit standard deviation, and choose the bandwidth as $b_n = 0.5n^{-1/2}$ in the main regression.[6] This choice is for undersmoothing so that

[5] In fact, estimation of $\gamma(\cdot)$ does not require such trimming. Once $\widehat{\gamma}(s_i)$ is constructed for all $i \in \Lambda_n$, the second-step estimator of (β_0, δ_0) uses the observations within $\mathcal{S}_0 \subset \mathcal{S}$ since $\widehat{\gamma}(s_i)$ might perform poorly if s_i is close to the boundary of its support $\mathcal{S}$. Following the Associate Editor's suggestion, we also implemented the same simulation with the middle 80% and 90% observations of s_i, but the results are very similar and hence not reported.

[6] We can alternatively choose the bandwidth (or the constant c in $b_n = cn^{-1/2}$) by the leave-one-out cross-validation. In particular, given a candidate bandwidth b_n, we first construct the leave-one-out estimate $\widehat{\gamma}_{-i}(s_i)$ from (4) for each $i \in \Lambda_n$ without using the ith observation. Second, leaving the ith observation out, we construct $\widehat{\beta}_{-i}$ and $\widehat{\delta}_{-i}$ as in (6) and (7) with $\pi_n = (nb_n)^{-1/2}$ using the bandwidth b_n under consideration. Finally, we choose the bandwidth that minimizes $\sum_{i \in \Lambda_n} (y_i - x_i^\top \widehat{\beta}_{-i} - x_i^\top \widehat{\delta}_{-i} \mathbf{1}[q_i \leq \widehat{\gamma}_{-i}(s_i)])^2 \mathbf{1}[s_i \in \mathcal{S}_0]$. However, when the sample

Table 3: Bias, RMSE, and Rej. Prob. of the LR Test with Cross-sectionally Correlated Data

	$s=0.0$				$s=0.5$				$s=1.0$			
$n\backslash\delta$	1	2	3	4	1	2	3	4	1	2	3	4
	Bias											
100	-0.47	-0.33	-0.28	-0.24	-0.05	-0.04	-0.04	-0.05	0.40	0.31	0.21	0.17
200	-0.39	-0.22	-0.16	-0.13	-0.04	-0.05	-0.03	-0.04	0.38	0.23	0.16	0.14
500	-0.31	-0.09	-0.07	-0.04	-0.02	-0.01	-0.01	-0.02	0.35	0.15	0.06	0.01
	RMSE											
100	0.55	0.32	0.24	0.19	0.30	0.18	0.13	0.11	0.48	0.36	0.26	0.19
200	0.45	0.22	0.12	0.08	0.28	0.14	0.08	0.05	0.48	0.31	0.20	0.15
500	0.38	0.10	0.04	0.02	0.24	0.08	0.04	0.02	0.45	0.22	0.11	0.06
	Rej. Prob. of the LR test											
100	0.21	0.11	0.10	0.09	0.20	0.15	0.13	0.14	0.30	0.21	0.15	0.14
200	0.13	0.07	0.04	0.03	0.13	0.09	0.12	0.12	0.22	0.13	0.10	0.07
500	0.08	0.04	0.02	0.02	0.11	0.07	0.09	0.13	0.14	0.08	0.04	0.02

Note: Entries are bias and root mean squared error (RMSE) of the estimator $\hat{\gamma}(s)$ and rejection probabilities of the LR when data are generated from (18) with $\gamma_0(s) = \cos(\pi s)/2$. The dependence structure is given in (19) with $\rho = 1$ and $m = 10$. The significance level is 5% and the results are based on 1000 simulations.

$n^{1-2\epsilon}b_n^2 \to 0$. To estimate $D(\gamma_0(s), s)$ and $V(\gamma_0(s), s)$, we use the rule-of-thumb bandwidths from the standard kernel regression satisfying $b_n' = O(n^{-1/5})$ and $b_n'' = O(n^{-1/6})$. All the results are based on 1000 simulations. In general, the estimator $\widehat{\gamma}(s)$ and the test for $\gamma_0(s)$ perform better as the sample size gets larger and as the coefficient change gets more significant. When δ_0 and n are large, the LR test can be conservative, which is also found in the classical constant threshold regression (e.g., Hansen (2000)). The overall performance remains quite similar whether the cross-sectional dependence is present or not.

Second, Table 4 reports the bias and the RMSE of the coefficient estimators. As expected, both the bias and the RMSE decrease as the sample size increases.[7] Table 5 shows the finite sample coverage properties of the 95% confidence intervals for the parametric components β_{20}, $\delta_{20}^* = \beta_{20} + \delta_{20}$, and δ_{20}. The results are based on the same simulation design as above with mild cross-section dependence (i.e., $\rho = 0.5$ and $m = 3$), which stands between the cases of Tables 2 and 3. Regarding the tuning parameters, we use the same bandwidth choice $b_n = 0.5n^{-1/2}$ as before and set the trimming parameter $\pi_n = (nb_n)^{-1/2}$. Such a choice satisfies the conditions $\pi_n \to 0$ and $\{\log n/(n^{1-2\epsilon}b_n)\}/\pi_n \to 0$ in our design. Unreported results suggest that choice of the constant in the bandwidth matters particularly with small samples like $n = 100$, but

size is large, the estimation result is not sensitive to the choice of bandwidth, so that the cross-validation is not much helpful. In this case, the simple rule-of-thumb bandwidth choice seems reasonable, such as $b_n = cn^{-1/2}$ for some constant c, say 0.5 or 1, that satisfies all the required regularity conditions.

[7] We also compared the parameter estimators with and without trimming. We find that the proposed trimming idea substantially reduces the bias without increasing much sample variance. This comparison is coherent with our expectation that the first step estimation error could influence the second step and leads to some bias.

Table 4: Bias and RMSE of Coefficient Estimates

$n\backslash\delta$	β_{20}				$\beta_{20}+\delta_{20}$				δ_{20}			
	1	2	3	4	1	2	3	4	1	2	3	4
	Bias											
100	0.12	0.11	0.08	0.09	-0.06	-0.08	-0.07	-0.07	-0.18	-0.19	-0.16	-0.17
200	0.11	0.09	0.05	0.04	-0.07	-0.07	-0.04	-0.04	-0.18	-0.16	-0.09	-0.08
500	0.08	0.04	0.02	0.01	-0.05	-0.03	-0.02	-0.02	-0.13	-0.07	-0.03	-0.02
	RMSE											
100	0.37	0.40	0.43	0.44	0.34	0.38	0.40	0.40	0.53	0.57	0.61	0.60
200	0.26	0.26	0.25	0.24	0.22	0.23	0.22	0.23	0.36	0.36	0.34	0.34
500	0.16	0.13	0.12	0.12	0.13	0.12	0.12	0.11	0.23	0.19	0.17	0.17

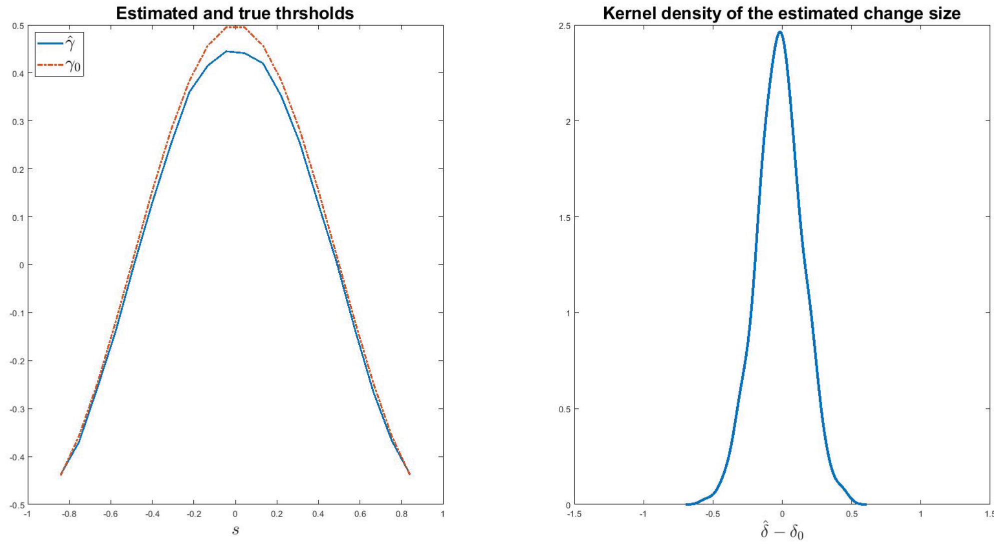
Note: Entries are bias and root mean squared error (RMSE) of the proposed two-step estimators for β_{20}, $\beta_{20}+\delta_{20}$, and δ_{20}. Data are generated from (18) with $\gamma_0(s) = \cos(\pi s)/2$, where the dependence structure is given in (19) with $\rho = 0.5$ and $m = 3$. The results are based on 1000 simulations.

Table 5: Coverage Prob. of the Confidence Intervals

$n\backslash\delta$	β_{20}				$\beta_{20}+\delta_{20}$				δ_{20}			
	1	2	3	4	1	2	3	4	1	2	3	4
	Coverage without small sample LRV adjustment											
100	0.82	0.86	0.88	0.89	0.84	0.86	0.88	0.89	0.84	0.85	0.88	0.89
200	0.87	0.90	0.91	0.91	0.88	0.91	0.93	0.93	0.87	0.91	0.92	0.93
500	0.86	0.92	0.94	0.94	0.89	0.93	0.94	0.94	0.84	0.91	0.92	0.94
	Coverage with small sample LRV adjustment											
100	0.92	0.94	0.94	0.94	0.92	0.94	0.94	0.94	0.92	0.94	0.95	0.96
200	0.92	0.96	0.96	0.95	0.94	0.95	0.96	0.96	0.93	0.96	0.97	0.96
500	0.92	0.96	0.97	0.97	0.94	0.96	0.97	0.97	0.89	0.95	0.96	0.97

Note: Entries are coverage probabilities of 95% confidence intervals for β_{20}, $\beta_{20}+\delta_{20}$, and δ_{20} with and without a small sample adjustment of the LRV estimator. Data are generated from (18) with $\gamma_0(s) = \cos(\pi s)/2$, where the dependence structure is given in (19) with $\rho = 0.5$ and $m = 3$. The results are based on 1000 simulations.

Figure 3: The Average of Threshold Estimates and Kernel Density of Coefficient Estimates



Note: The left panel depicts the average of $\hat{\gamma}(s)$ and the right panel depicts the kernel density of $\hat{\delta}_2 - \delta_{20}$ from 1000 simulations. Data are generated from (18) with $\gamma_0(s) = \cos(\pi s)/2$, where the dependence structure is given in (19) with $\rho = 0.5$ and $m = 3$.

such effect quickly decays as the sample size gets larger. For the estimator of the LRV, we use the spatial lag order of 5 following Conley and Molinari (2007). Results with other lag choices are similar and hence omitted. The result suggests that the asymptotic normality is better approximated with larger samples and larger change sizes. Table 5 shows the same results with a small sample adjustment of the LRV estimator for Ω^* by dividing it by the sample trimming fraction, $\sum_{i \in \Lambda_n} (\mathbf{1}[q_i > \widehat{\gamma}(s_i) + \pi_n] + \mathbf{1}[q_i < \widehat{\gamma}(s_i) - \pi_n]) \mathbf{1}[s_i \in \mathcal{S}_0] / \sum_{i \in \Lambda_n} \mathbf{1}[s_i \in \mathcal{S}_0]$. This ratio enlarges the LRV estimator and improves the coverage probabilities, especially when the change size is small. It only affects the finite sample performance as it approaches one in probability as $n \to \infty$.

Third, Figure 3 depicts the averaged $\widehat{\gamma}(s)$ over $s \in \mathcal{S}_0$ across simulation draws on the left panel and the density estimator of $\widehat{\delta}_2 - \delta_{20}$ on the right panel. Data are generated from the same model as in Table 5 with $\delta = 4$ and $n = 500$. From the left panel, we see that $\widehat{\gamma}(s)$ is uniformly close to $\gamma_0(s)$ though it shows some small sample downward bias near $s = 0$.[8] This finding is coherent with the results in Tables 2 and 3. From the right panel, we see that $\widehat{\delta}_2 - \delta_{20}$ is approximately normal with zero mean, which is coherent with Theorem 5.

[8] Such downward bias is also found in the standard local constant estimators, where the bias is a positive function of the second derivative of $\gamma_0(s)$ when $\dot{\gamma}_0(s) = 0$.

6 Applications

6.1 Tipping point and social segregation

The first application is about the tipping point problem in social segregation, which stimulates a vast literature in labor, public, and political economics. Schelling (1971) initially proposes the tipping point model to study the fact that the white population decreases substantially once the minority share exceeds a certain tipping point. Card, Mas, and Rothstein (2008) empirically estimate this model and find strong evidence for such a tipping point phenomenon. In particular, they specify the threshold regression model as

$$y_i = \beta_{10} + \delta_{10}\mathbf{1}\left[q_i \leq \gamma_0\right] + x_{2i}^\top \beta_{20} + u_i,$$

where for tract i in a certain city, q_i is the minority share in percentage at the beginning of a certain decade, y_i is the normalized white population change in percentage within this decade, and x_{2i} is a vector of control variables. They apply the least squares method to estimate the tipping point γ_0. For most cities and for the periods 1970-80, 1980-90, and 1990-2000, they find that white population flows exhibit the tipping-like behavior, with the estimated tipping points ranging from 5% to 20% across cities.

In Section VII of Card, Mas, and Rothstein (2008), they also find that the location of the estimated tipping point substantially depends on white residents' attitudes toward the minority. Specifically, they first construct a city-level index that measures the racial attitudes of whites and regress the estimated tipping point of each city on this index. The regression coefficient is significantly different from zero, suggesting that the tipping point is heterogeneous across cities. See Lee and Wang (2022) for a formal test of homogeneous tipping points.

We go one step further by considering a more flexible model in the tract level given as

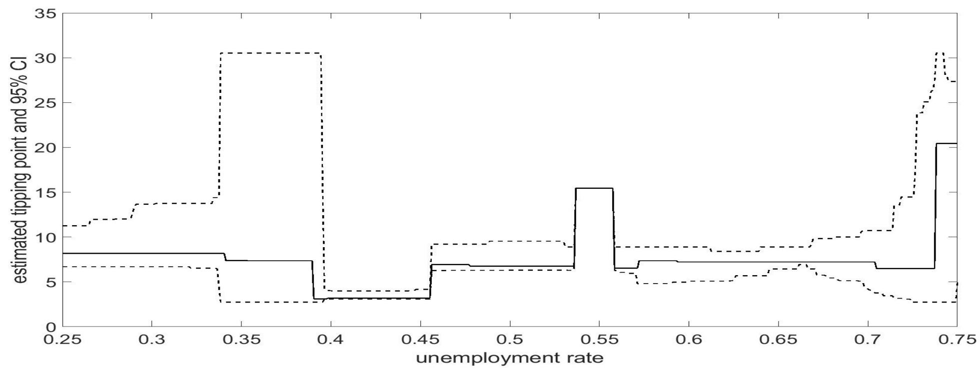
$$y_i = \beta_{10} + \delta_{10}\mathbf{1}\left[q_i \leq \gamma_0(s_i)\right] + x_{2i}^\top \beta_{20} + u_i,$$

where $\gamma_0(\cdot)$ denotes an unknown tipping point function and s_i denotes the attitude index. The nonparametric function $\gamma_0(\cdot)$ here allows for heterogeneous tipping points across tracts depending on the level of the attitude index s_i in tract i. Unfortunately, the attitude index by Card, Mas, and Rothstein (2008) is only available at the aggregated city-level, and hence we cannot use it to analyze the census tract-level observations. For this reason, we instead use the tract-level unemployment rate as s_i to illustrate the nonparametric threshold function, which is readily available in the original dataset. Such a compromise is far from being perfect but can be partially justified since race discrimination has been widely documented to be correlated with employment (e.g., Darity and Mason (1998)).

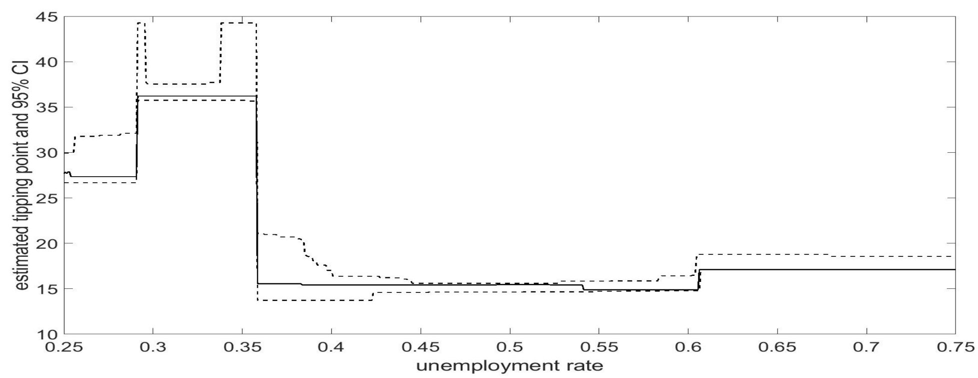
We use the data provided by Card, Mas, and Rothstein (2008) and estimate the tipping point

Figure 4: Estimate of the tipping point as a function of the unemployment rate

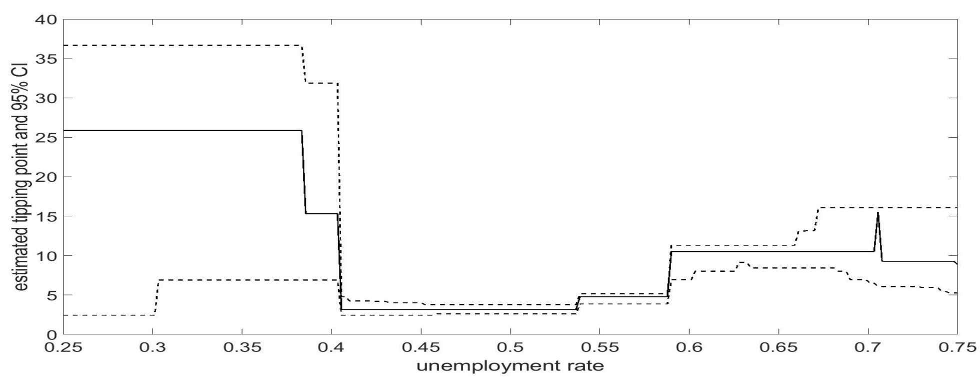
Panel A: Estimated tipping point function in Chicago 1980-90



Panel B: Estimated tipping point function in Los Angeles 1980-90



Panel C: Estimated tipping point function in New York City 1980-90



Note: The figure depicts the point estimates (solid) and the 95% pointwise confidence intervals (dash) of the tipping points as a function of the unemployment rate. The vertical axis is the estimated tipping point in percentage and the horizontal axis is the tract-level unemployment normalized to quantile level. Data are available from Card, Mas, and Rothstein (2008).

function $\gamma_0(\cdot)$ over census tracts by the method introduced in Section 2. As in their work, we drop the tracts where the minority shares are above 60 percentage points and use five control variables as x_{2i}, including the logarithm of mean family income, the fractions of single-unit, vacant, and renter-occupied housing units, and the fraction of workers who use public transport to travel to work.

Figure 4 depicts the estimated tipping points and the 95% pointwise confidence intervals by inverting the likelihood ratio test statistic (13) in the years 1980-90 in Chicago, Los Angeles, and New York City, whose sample sizes are relatively large. For each city, the bandwidth is set as $b_n = cn^{-1/2}$, where the constant $c > 0$ is chosen by cross-validation as described in the footnote 6, which is 3.20, 4.87, and 3.42, respectively. We make the following comments. First, the estimates of the tipping points vary substantially with the unemployment rate within all three cities. Therefore, the standard constant tipping point model is insufficient to characterize the segregation fully. Second, the tipping points as functions of the unemployment rate do not exhibit the same pattern across cities, reinforcing the heterogeneous tipping points in the city-level as found in Card, Mas, and Rothstein (2008). Finally, the estimated tipping point $\widehat{\gamma}(s)$ as a function of s can be discontinuous, which does not contrast with Assumption A-(vi), that is, the true function $\gamma_0(\cdot)$ is smooth. The discontinuity comes from the fact that $\widehat{\gamma}(s)$ is obtained by grid search and can only take values among the discrete points $\{q_1, ..., q_n\}$ in finite samples.

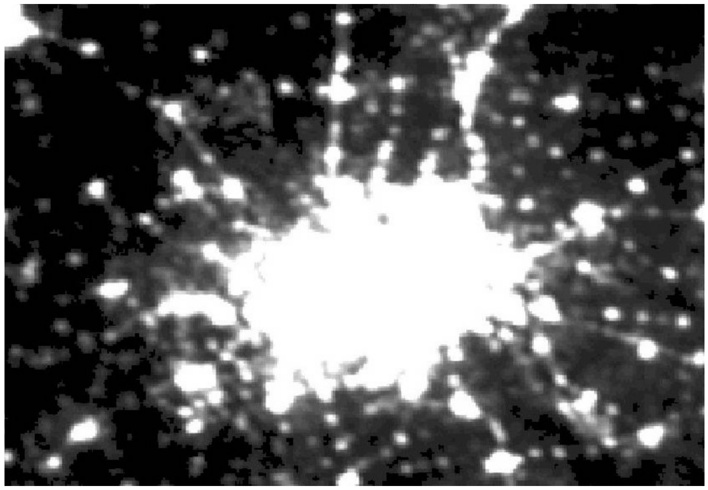
6.2 Metropolitan area determination

The second application is about determining the boundary of a metropolitan area, which is one of the fundamental questions in urban economics. Recently, researchers propose to use night-time light intensity obtained by satellite imagery to define metropolitan areas. The intuition is straightforward: metropolitan areas are bright at night while rural areas are dark.

Specifically, the National Oceanic and Atmospheric Administration (NOAA) collects satellite imagery of nighttime lights at approximately 1-kilometer resolution since 1992. NOAA further constructs several indices measuring the annual light intensity. Following the literature (e.g., Dingel, Miscio, and Davis (2021)), we choose the "average visible, stable lights" index that ranges from 0 (dark) to 63 (bright). For illustration, we focus on Dallas, Texas and use the data in the years 1995, 2000, 2005, and 2010. In each year, the data are recorded as a 240×360 grid that covers the latitudes from 32°N to 34°N and the longitudes from 98.5°W to 95.5°W. The total sample size is 240×360=86400 each year. These data are available at NOAA's website. Figure 5 depicts the intensity of the stable nighttime light of the Dallas area in 2010 as an example.

Let y_i be the level of nighttime light intensity and (q_i, s_i) be the latitude and longitude of the ith pixel, which is normalized into the equally-spaced grids on $[0, 1]^2$. To define the metropolitan area, existing literature in urban economics first chooses an *ad hoc* intensity threshold, say 95% quantile of y_i, and categorizes the ith pixel as a part of the metropolitan area if y_i is larger than the threshold. See Dingel, Miscio, and Davis (2021), Baragwanath, Goldblatt, Hanson,

Figure 5: Nighttime light intensity in Dallas, Texas, in 2010



Note: The figure depicts the intensity of the stable nighttime light in Dallas, TX 2010. Data are available from https://www.ncei.noaa.gov/.

and Khandelwal (2021), and references therein. In particular, in Section 2.1 of Dingel, Miscio, and Davis (2021), they note that "The choice of the light-intensity threshold, which governs the definitions of the resulting metropolitan areas, is not pinned down by economic theory or prior empirical research." Our new approach can provide a data-driven guidance of choosing the intensity threshold from the econometric perspective.

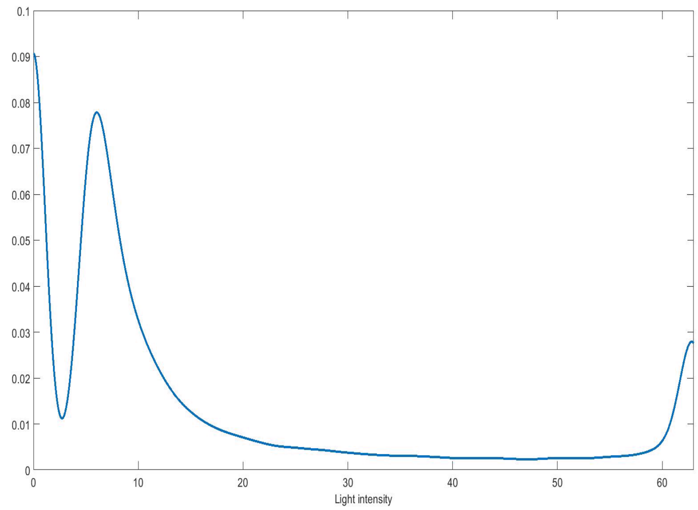
To this end, we first examine whether the light intensity data exhibits a clear threshold pattern. We plot the kernel density estimates of y_i in the year 2010 in Figure 6. The bandwidth is the standard rule-of-thumb one. The estimated density exhibits three peaks at around the intensity levels 0, 8, and 63. They respectively correspond to the rural area, small towns, and the central metropolitan area. It shows that the threshold model is appropriate in characterizing such a mean-shift pattern.

We consider

$$y_i = \beta_0 + \delta_0 \mathbf{1} \left[g_0 \left(q_i, s_i \right) \leq 0 \right] + u_i$$

and implement the rotation and estimation method described in Section 4. In particular, we pick the center point in the bright middle area as the Dallas metropolitan center, which corresponds to the pixel point in the 181st column from the left and the 100th row from the bottom. Then for each a° over the 500 equally-spaced grid on $[0^\circ, 360^\circ]$, we rotate the data by a° degrees counterclockwise and estimate the model (17) with $x_i = 1$. The bandwidth is chosen as $b_n = cn^{-1/2}$ with $c = 1$. Other choices of c lead to almost identical results, given the large sample size. Figure 7 presents the estimated metropolitan areas using our nonparametric approach (red) and the area determined by the *ad hoc* threshold of the 95% quantile of y_i (black) in the years 1995, 2000, 2005, and 2010.

Figure 6: Kernel density estimate of nighttime light intensity, Dallas 2010

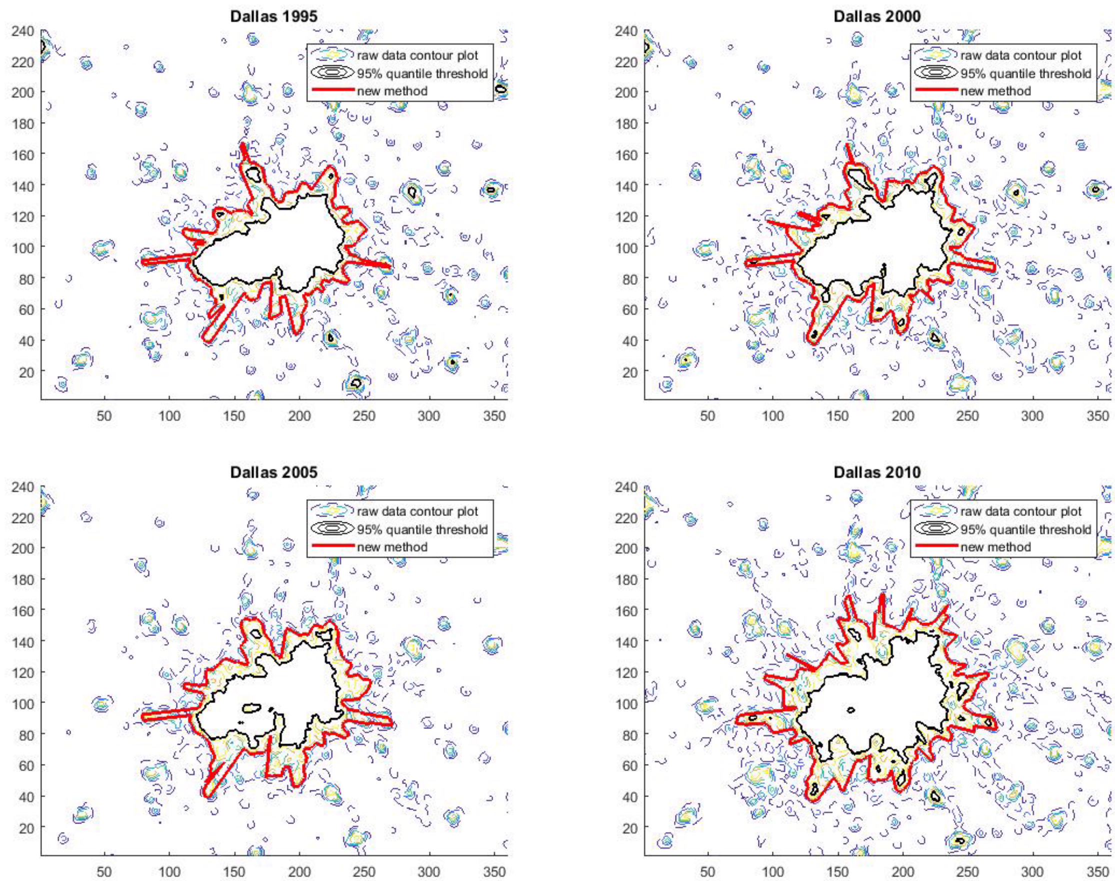


Note: The figure depicts the kernel density estimate of the strength of the stable nighttime light in Dallas, TX 2010. Data are available from https://www.ncei.noaa.gov/.

It clearly shows the expansion of the Dallas metropolitan area over the 15 years of the sample period.

Several interesting findings are summarized as follows. First, the estimated boundary is highly nonlinear as a function of the angle. Therefore, any parametric threshold model could lead to a substantially misleading result. Second, our estimated area is larger than that determined by the *ad hoc* threshold, by 80.31%, 81.56%, 106.46%, and 102.09% in the years 1995, 2000, 2005, and 2010, respectively. In particular, our nonparametric estimates tend to include some suburban areas that exhibit strong light intensity and that are geographically close to the city center. For example, the very left stretch-out area in the estimated boundary corresponds to Fort Worth, which is 30 miles from downtown Dallas. Residents can easily commute by train or driving on the Interstate 30. It is then reasonable to include Fort Worth as a part of the metropolitan Dallas area for economic analysis. Third, given the large sample size, the 95% confidence intervals of the boundary are too narrow to be distinguished from the estimates and therefore omitted from the figure. Such narrow intervals apparently exclude the boundary determined by the *ad hoc* method. Finally, the estimated value of $\beta_0 + \delta_0$ is approximately 53 in these sample periods, which corresponds to the 89% quantile of y_i in the sample. This suggests that a more proper choice of the level of light intensity threshold is the 89% quantile of y_i, instead of the 95% quantile, if one needs to choose the light-intensity threshold to determine the Dallas metropolitan area.

Figure 7: Metropolitan area determination in Dallas (color online)



Note: The figure depicts the city boundary determined by either the new method or by taking the 0.95 quantile of nighttime light strength as the threshold, using the satellite imagery data for Dallas, TX in the years 1995, 2000, 2005, and 2010. Data are available from https://www.ncei.noaa.gov/.

7 Concluding Remarks

This paper proposes a novel approach to conduct sample splitting. In particular, we develop a nonparametric threshold regression model where two variables can jointly determine a unknown threshold boundary. Our approach can be easily generalized so that the sample splitting depends on more numbers of variables, though such an extension is subject to the curse of dimensionality, as usually observed in the kernel regression literature. The main interest is in identifying the threshold function that determines how to split the sample. Thus our model should be distinguished from the smoothed threshold regression model or the random coefficient regression model.

This new approach is empirically relevant in broad areas studying sample splitting (e.g., segregation and group-formation) and heterogeneous effects over different subsamples. We illustrate some of them with the tipping point problem in social segregation and metropolitan area determination using satellite imagery datasets. Though we omit in this paper, we also estimate the economic border between Brooklyn and Queens boroughs in New York City using housing prices.[9] The estimated border is substantially different from the existing administrative border, which was determined in 1931 and cannot reflect the dramatic city development. Interestingly, the estimated border coincides with the Jackson Robinson Parkway and the Long Island Railroad. This finding provides new evidence that local transportation corridors could increase community segregation (cf. Ananat (2011) and Heilmann (2018)).

We list some related works, which could motivate potential theoretical extensions. First, while we focus on the local constant estimation in this paper, one could consider the local linear estimation. Although grid search can be very demanding in determining the two local parameters in this case, we could use the MCMC algorithm by Yu and Fan (2021) and the mixed integer optimization (MIO) algorithms by Lee, Liao, Seo, and Shin (2021). Besides the computational challenge, however, the asymptotic derivation is more involved since we need to consider higher-order expansions of the objective function. Second, while our nonparametric setup is on the threshold function $\gamma_0(\cdot)$, some recent literature studies the nonparametric regression model with a parametric threshold, such as $y_i = m_1(x_i) + m_2(x_i)\mathbf{1}[q_i \leq \gamma_0] + u_i$, where $m_1(\cdot)$ and $m_2(\cdot)$ are different nonparametric functions. See, for example, Henderson, Parmeter, and Su (2017), Chiou, Chen, and Chen (2018), Yu and Phillips (2018), Yu, Liao, and Phillips (2019), and Delgado and Hidalgo (2000).

[9] The result is available upon request.

A Appendix

Throughout the proof, we denote $K_i(s) = K((s_i - s)/b_n)$ and $\mathbf{1}_i(\gamma) = \mathbf{1}[q_i \leq \gamma]$. We let C and its variants such as C_1 and C_1' stand for generic positive finite constants that may vary across lines. We also let $a_n = n^{1-2\epsilon} b_n$. All the additional lemmas in the proof assume that Assumptions ID and A hold. Omitted proofs for technical lemmas are collected in the online supplementary appendix.

A.1 Proof of Theorem 1 (Identification)

Proof of Theorem 1 The proof follows similarly as Theorem A.1 in Lee, Liao, Seo, and Shin (2021). For any constant $\gamma \in \Gamma$ with given $s \in \mathcal{S}$, we define a conditional L_2-loss as

$$
\begin{aligned}
& R(\beta, \delta, \gamma | s) \\
= \; & \mathbb{E}\left[\left\{ y_i - x_i^\top \beta - x_i^\top \delta \mathbf{1}_i(\gamma) \right\}^2 \middle| s_i = s \right] - \mathbb{E}\left[\left\{ y_i - x_i^\top \beta_0 - x_i^\top \delta_0 \mathbf{1}_i(\gamma_0(s)) \right\}^2 \middle| s_i = s \right] \\
= \; & \mathbb{E}\left[\left\{ x_i^\top (\beta_0 - \beta) + x_i^\top (\delta_0 - \delta) \mathbf{1}_i(\gamma_0(s)) + x_i^\top \delta \left(\mathbf{1}_i(\gamma_0(s)) - \mathbf{1}_i(\gamma) \right) \right\}^2 \middle| s_i = s \right],
\end{aligned}
$$

which is continuous in $(\beta^\top, \delta^\top, \gamma)^\top$. By construction, $R(\beta, \delta, \gamma | s) \geq 0$ for any $(\beta^\top, \delta^\top, \gamma)^\top \in \mathbb{R}^{2\dim(x)} \times \Gamma$ and $R(\beta_0, \delta_0, \gamma_0(s) | s) = 0$. Hence, it suffices to show that $R(\beta, \delta, \gamma | s) > 0$ for any vector $(\beta^\top, \delta^\top, \gamma)^\top \neq (\beta_0^\top, \delta_0^\top, \gamma_0(s))^\top$ given $s \in \mathcal{S}$. To this end, we split the event $(\beta^\top, \delta^\top, \gamma)^\top \neq (\beta_0^\top, \delta_0^\top, \gamma_0(s))^\top$ into two disjoint cases: (i) $\gamma \neq \gamma_0(s)$ but $(\beta^\top, \delta^\top)^\top = (\beta_0^\top, \delta_0^\top)^\top$; or (ii) $(\beta^\top, \delta^\top)^\top \neq (\beta_0^\top, \delta_0^\top)^\top$ for any $\gamma \in \Gamma$.

For (i), note that

$$
\begin{aligned}
R(\beta_0, \delta_0, \gamma | s) &= \delta_0^\top \mathbb{E}\left[x_i x_i^\top \left(\mathbf{1}_i(\gamma_0(s)) - \mathbf{1}_i(\gamma) \right)^2 \middle| s_i = s \right] \delta_0 \\
&= \delta_0^\top \mathbb{E}\left[x_i x_i^\top \mathbf{1}\left[\min\{\gamma, \gamma_0(s)\} < q_i \leq \max\{\gamma, \gamma_0(s)\} \right] \middle| s_i = s \right] \delta_0 \\
&= \int_{\min\{\gamma, \gamma_0(s)\}}^{\max\{\gamma, \gamma_0(s)\}} \delta_0^\top \mathbb{E}\left[x_i x_i^\top \middle| q_i = q, s_i = s \right] \delta_0 f(q|s) dq \\
&\geq C(s) \mathbb{P}\left(\min\{\gamma, \gamma_0(s)\} < q_i \leq \max\{\gamma, \gamma_0(s)\} \middle| s_i = s \right) \\
&> 0
\end{aligned}
$$

from Assumptions ID-(i), (iii), and (iv), where $C(s) = \inf_{q \in \mathcal{Q}} \delta_0^\top \mathbb{E}[x_i x_i^\top | q_i = q, s_i = s] \delta_0 > 0$. The last probability is strictly positive because we assume $f(q|s) > 0$ for any $(q, s) \in \mathcal{Q} \times \mathcal{S}$ and $\gamma_0(s)$ is not located on the boundary of $\mathcal{Q}$ as $\varepsilon(s) < \mathbb{P}(q_i \leq \gamma_0(s) | s_i = s) < 1 - \varepsilon(s)$ for some $\varepsilon(s) > 0$.

For (ii), let $I_{i\gamma}(s) = \{q_i \leq \min\{\gamma, \gamma_0(s)\}\} \cup \{q_i \geq \max\{\gamma, \gamma_0(s)\}\}$ and note that

$$
\begin{aligned}
& R(\beta, \delta, \gamma | s) \\
\geq \; & \mathbb{E}\left[\mathbf{1}[I_{i\gamma}(s)] \left\{ x_i^\top (\beta_0 - \beta) + x_i^\top (\delta_0 - \delta) \mathbf{1}_i(\gamma_0(s)) + x_i^\top \delta \left(\mathbf{1}_i(\gamma_0(s)) - \mathbf{1}_i(\gamma) \right) \right\}^2 \middle| s_i = s \right] \\
= \; & \mathbb{E}\left[\left\{ x_i^\top (\beta_0 - \beta) + x_i^\top (\delta_0 - \delta) \right\}^2 \mathbf{1}\left[q_i \leq \min\{\gamma, \gamma_0(s_i)\} \right] \middle| s_i = s \right] \\
& + \mathbb{E}\left[\left\{ x_i^\top (\delta_0 - \delta) \right\}^2 \mathbf{1}\left[q_i > \max\{\gamma, \gamma_0(s_i)\} \right] \middle| s_i = s \right]
\end{aligned}
$$

$> \quad 0$

when $(\beta^\top, \delta^\top)^\top \neq (\beta_0^\top, \delta_0^\top)^\top$ from Assumption ID-(ii), for any $\gamma \in \Gamma$. ■

A.2 Proof of Theorem 2 (Pointwise Convergence) and Key Lemmas

We first present a covariance inequality for strong mixing random field. Suppose Λ_1 and Λ_2 are finite subsets in Λ_n with $|\Lambda_1| = k_x, |\Lambda_2| = l_x$, and let X_1 and X_2 be random variables respectively measurable with respect to the σ-algebra's generated by Λ_1 and Λ_2. If $\mathbb{E}[|X_1|^{p_x}] < \infty$ and $\mathbb{E}[|X_2|^{q_x}] < \infty$ with $1/p_x + 1/q_x + 1/r_x = 1$ for some constants $p_x, q_x > 1$ and $r_x > 0$, then

$$|Cov[X_1, X_2]| < 8\alpha_{k_x,l_x}(\lambda(\Lambda_1, \Lambda_2))^{1/r_x}\mathbb{E}[|X_1|^{p_x}]^{1/p_x}\,\mathbb{E}[|X_2|^{q_x}]^{1/q_x} \tag{A.1}$$

under Assumptions A-(i) and A-(iii). This covariance inequality is presented as Lemma 1 in the working paper version of Jenish and Prucha (2009). The proof is also available in Hall and Heyde (1980), p.277.

For a given $s \in \mathcal{S}_0$, we define

$$\begin{aligned} M_n(\gamma; s) &= \frac{1}{nb_n}\sum_{i\in\Lambda_n} x_i x_i^\top \mathbf{1}_i(\gamma) K_i(s) \\ J_n(\gamma; s) &= \frac{1}{\sqrt{nb_n}}\sum_{i\in\Lambda_n} x_i u_i \mathbf{1}_i(\gamma) K_i(s). \end{aligned}$$

The following four lemmas give the asymptotic behavior of $M_n(\gamma; s)$ and $J_n(\gamma; s)$.

Lemma A.1 *For any given $s \in \mathcal{S}_0$, there exist finite constants C^*, C^{**} and $\varpi \geq (n^{(2+\varphi)/(2+2\varphi)}b_n)^{-1}$ such that for any $\gamma_1 \in \Gamma$ and $\eta > 0$,*

$$\mathbb{P}\left(\sup_{\gamma\in[\gamma_1,\gamma_1+\varpi]} \|J_n(\gamma; s) - J_n(\gamma_1; s)\| > \eta\right) \leq \frac{C^*\varpi^2}{\eta^4}$$

with sufficiently large n if $\eta \geq C^{**}(n^{1/(1+\varphi)}b_n)^{-1/2}$, where $\varphi > 0$ is specified in Assumption A-(vi).

Lemma A.2 *For any fixed $s \in \mathcal{S}_0$,*

$$J_n(\gamma; s) \Rightarrow J(\gamma; s),$$

where $J(\gamma; s)$ is a mean-zero Gaussian process indexed by γ as $n \to \infty$.

Lemma A.3

$$\sup_{(\gamma,s)\in\Gamma\times\mathcal{S}_0} \|M_n(\gamma; s) - M(\gamma; s)\| \to_p 0,$$

$$\sup_{(\gamma,s)\in\Gamma\times\mathcal{S}_0} (nb_n)^{-1/2}\|J_n(\gamma; s)\| \to_p 0$$

as $n \to \infty$, where

$$M(\gamma; s) = \int_{-\infty}^{\gamma} D(q,s)f(q,s)\,dq. \tag{A.2}$$

Lemma A.4 *Uniformly over $s \in \mathcal{S}_0$,*

$$\Delta M_n(s) \equiv \frac{1}{nb_n}\sum_{i\in\Lambda_n} x_i x_i^\top \{\mathbf{1}_i(\gamma_0(s_i)) - \mathbf{1}_i(\gamma_0(s))\} K_i(s) = O_{a.s.}(b_n). \tag{A.3}$$

The following lemma establishes the pointwise consistency of $\widehat{\gamma}(s)$.

Lemma A.5 *For a given $s \in \mathcal{S}_0$, $\widehat{\gamma}(s) \to_p \gamma_0(s)$ as $n \to \infty$.*

Proof of Lemma A.5 For given $s \in \mathcal{S}_0$, we let $\widetilde{y}_i(s) = K_i(s)^{1/2} y_i$, $\widetilde{x}_i(s) = K_i(s)^{1/2} x_i$, $\widetilde{u}_i(s) = K_i(s)^{1/2} u_i$, $\widetilde{x}_i(\gamma; s) = K_i(s)^{1/2} x_i \mathbf{1}_i(\gamma)$, and $\widetilde{x}_i(\gamma_0(s_i); s) = K_i(s)^{1/2} x_i \mathbf{1}_i(\gamma_0(s_i))$. We denote $\widetilde{y}(s)$, $\widetilde{X}(s)$, $\widetilde{u}(s)$, $\widetilde{X}(\gamma; s)$, and $\widetilde{X}(\gamma_0(s_i); s)$ as their corresponding matrices of n-stacks. Then $\widehat{\theta}(\gamma; s) = (\widehat{\beta}(\gamma; s)^\top, \widehat{\delta}(\gamma; s)^\top)^\top$ in (2) is given as

$$\widehat{\theta}(\gamma; s) = (\widetilde{Z}(\gamma; s)^\top \widetilde{Z}(\gamma; s))^{-1} \widetilde{Z}(\gamma; s)^\top \widetilde{y}(s), \tag{A.4}$$

where $\widetilde{Z}(\gamma; s) = [\widetilde{X}(s), \widetilde{X}(\gamma; s)]$. Therefore, since $\widetilde{y}(s) = \widetilde{X}(s)\beta_0 + \widetilde{X}(\gamma_0(s_i); s)\delta_0 + \widetilde{u}(s)$ and $\widetilde{X}(s)$ lies in the space spanned by $\widetilde{Z}(\gamma; s)$, we have

$$\begin{aligned} Q_n(\gamma; s) - \widetilde{u}(s)^\top \widetilde{u}(s) &= \widetilde{y}(s)^\top \left(I_n - P_{\widetilde{Z}}(\gamma; s)\right) \widetilde{y}(s) - \widetilde{u}(s)^\top \widetilde{u}(s) \\ &= -\widetilde{u}(s)^\top P_{\widetilde{Z}}(\gamma; s) \widetilde{u}(s) + 2\delta_0^\top \widetilde{X}(\gamma_0(s_i); s)^\top \left(I_n - P_{\widetilde{Z}}(\gamma; s)\right) \widetilde{u}(s) \\ &\quad + \delta_0^\top \widetilde{X}(\gamma_0(s_i); s)^\top \left(I_n - P_{\widetilde{Z}}(\gamma; s)\right) \widetilde{X}(\gamma_0(s_i); s)\delta_0, \end{aligned}$$

where $P_{\widetilde{Z}}(\gamma; s) = \widetilde{Z}(\gamma; s)(\widetilde{Z}(\gamma; s)^\top \widetilde{Z}(\gamma; s))^{-1} \widetilde{Z}(\gamma; s)^\top$ and I_n is the identity matrix of rank n. Note that $P_{\widetilde{Z}}(\gamma; s)$ is the same as the projection onto $[\widetilde{X}(s) - \widetilde{X}(\gamma; s), \widetilde{X}(\gamma; s)]$, where $\widetilde{X}(\gamma; s)^\top (\widetilde{X}(s) - \widetilde{X}(\gamma; s)) = 0$. Furthermore, for $\gamma \geq \gamma_0(s_i)$, $\widetilde{x}_i(\gamma_0(s_i); s)^\top (\widetilde{x}_i(s) - \widetilde{x}_i(\gamma; s)) = 0$ and hence $\widetilde{X}(\gamma_0(s_i); s)^\top \widetilde{X}(\gamma; s) = \widetilde{X}(\gamma_0(s_i); s)^\top \widetilde{X}(\gamma_0(s_i); s)$. Since we can rewrite

$$\begin{aligned} M_n(\gamma; s) &= \frac{1}{nb_n} \sum_{i \in \Lambda_n} \widetilde{x}_i(\gamma; s) \widetilde{x}_i(\gamma; s)^\top \quad \text{and} \\ J_n(\gamma; s) &= \frac{1}{\sqrt{nb_n}} \sum_{i \in \Lambda_n} \widetilde{x}_i(\gamma; s) \widetilde{u}_i(s), \end{aligned}$$

Lemma A.3 yields that

$$\begin{aligned} \widetilde{Z}(\gamma; s)^\top \widetilde{u}(s) &= [\widetilde{X}(s)^\top \widetilde{u}(s), \widetilde{X}(\gamma; s)^\top \widetilde{u}(s)] = O_p\left((nb_n)^{1/2}\right) \\ \widetilde{Z}(\gamma; s)^\top \widetilde{X}(\gamma_0(s_i); s) &= [\widetilde{X}(s)^\top \widetilde{X}(\gamma_0(s_i); s), \widetilde{X}(\gamma; s)^\top \widetilde{X}(\gamma_0(s_i); s)] \\ &= [\widetilde{X}(s)^\top \widetilde{X}(\gamma_0(s_i); s), \widetilde{X}(\gamma_0(s_i); s)^\top \widetilde{X}(\gamma_0(s_i); s)] = O_p(nb_n) \end{aligned}$$

for given s. It follows that

$$\begin{aligned} \Upsilon_n(\gamma; s) &\equiv \frac{1}{a_n}\left(Q_n(\gamma; s) - \widetilde{u}(s)^\top \widetilde{u}(s)\right) \\ &= O_p\left(\frac{1}{a_n}\right) + O_p\left(\frac{1}{a_n^{1/2}}\right) + \frac{1}{nb_n} c_0^\top \widetilde{X}(\gamma_0(s_i); s)^\top \left(I_n - P_{\widetilde{Z}}(\gamma; s)\right) \widetilde{X}(\gamma_0(s_i); s) c_0 \\ &= \frac{1}{nb_n} c_0^\top \widetilde{X}(\gamma_0(s_i); s)^\top \left(I - P_{\widetilde{Z}}(\gamma; s)\right) \widetilde{X}(\gamma_0(s_i); s) c_0 + o_p(1) \end{aligned} \tag{A.5}$$

for $a_n = n^{1-2\epsilon} b_n \to \infty$ as $n \to \infty$. Moreover, we have

$$\begin{aligned} M_n(\gamma_0(s_i); s) &= \frac{1}{nb_n} \sum_{i \in \Lambda_n} \widetilde{x}_i(\gamma_0(s_i); s) \widetilde{x}_i(\gamma_0(s_i); s)^\top \\ &= M_n(\gamma_0(s); s) + \Delta M_n(s) \end{aligned} \tag{A.6}$$

$$= \quad M_n\left(\gamma_0(s);s\right) + o_{a.s.}(1)$$

from Lemma A.4, where $\Delta M_n(s)$ is defined in (A.3). It follows that the last expression in (A.5) satisfies

$$\begin{aligned} &\frac{1}{nb_n} c_0^\top \widetilde{X}(\gamma_0(s_i);s)^\top \left(I_n - P_{\widetilde{Z}}(\gamma;s)\right) \widetilde{X}(\gamma_0(s_i);s)c_0 \\ &\to_p c_0^\top M(\gamma_0(s);s)c_0 - c_0^\top M(\gamma_0(s);s)^\top M(\gamma;s)^{-1} M(\gamma_0(s);s)c_0 \equiv \Upsilon_0(\gamma;s) < \infty \end{aligned} \tag{A.7}$$

uniformly over $\gamma \in \Gamma \cap [\gamma_0(s), \infty)$ as $n \to \infty$, from Lemma A.3 and Assumptions ID-(ii) and A-(viii). However,

$$\partial \Upsilon_0(\gamma;s)/\partial\gamma = c_0^\top M(\gamma_0(s);s)^\top M(\gamma;s)^{-1} D(\gamma,s) f(\gamma,s) M(\gamma;s)^{-1} M(\gamma_0(s);s)c_0 \geq 0$$

and

$$\partial \Upsilon_0(\gamma_0(s);s)/\partial\gamma = c_0^\top D(\gamma_0(s),s) f(\gamma_0(s),s)c_0 > 0 \tag{A.8}$$

from Assumption A-(viii), which implies that $\Upsilon_0(\gamma;s)$ is continuous, non-decreasing, and uniquely minimized at $\gamma_0(s)$ given $s \in \mathcal{S}_0$.

We can symmetrically show that, uniformly over $\gamma \in \Gamma \cap (-\infty, \gamma_0(s)]$, $\Upsilon_0(\gamma;s)$ in (A.7) is continuous, non-increasing, and uniquely minimized at $\gamma_0(s)$ as well. Therefore, given $s \in \mathcal{S}_0$, $\sup_{\gamma\in\Gamma} |\Upsilon_n(\gamma;s) - \Upsilon_0(\gamma;s)| = o_p(1)$; $\Upsilon_0(\gamma;s)$ is continuous and uniquely minimized at $\gamma_0(s)$. Since Γ is compact and $\widehat{\gamma}(s)$ is the minimizer of $\Upsilon_n(\gamma;s)$, the pointwise consistency follows as Theorem 2.1 of Newey and McFadden (1994). ■

We let $\phi_{1n} = a_n^{-1}$, where $a_n = n^{1-2\epsilon} b_n$ and ϵ is given in Assumption A-(ii). For a given $s \in \mathcal{S}_0$ and any $\gamma : \mathcal{S}_0 \mapsto \Gamma$, we define

$$T_n(\gamma;s) = \frac{1}{nb_n} \sum_{i\in\Lambda_n} \left(c_0^\top x_i\right)^2 \left|\mathbf{1}_i(\gamma(s)) - \mathbf{1}_i(\gamma_0(s))\right| K_i(s), \tag{A.9}$$

$$\overline{T}_n(\gamma,s) = \frac{1}{nb_n} \sum_{i\in\Lambda_n} \|x_i\|^2 \left|\mathbf{1}_i(\gamma(s)) - \mathbf{1}_i(\gamma_0(s))\right| K_i(s), \tag{A.10}$$

$$L_{nj}(\gamma;s) = \frac{1}{\sqrt{nb_n}} \sum_{i\in\Lambda_n} x_{ij} u_i \left\{\mathbf{1}_i(\gamma(s)) - \mathbf{1}_i(\gamma_0(s))\right\} K_i(s) \tag{A.11}$$

for $j = 1, \ldots, \dim(x)$, where x_{ij} denotes the jth element of x_i.

Lemma A.6 *For a given $s \in \mathcal{S}_0$, for any $\gamma(\cdot) : \mathcal{S}_0 \mapsto \Gamma$, $\eta(s) > 0$, and $\varepsilon(s) > 0$, there exist constants $0 < C_T(s), C_{\overline{T}}(s), \overline{C}(s), \overline{r}(s) < \infty$ such that if n is sufficiently large,*

$$\mathbb{P}\left(\inf_{\overline{r}(s)\phi_{1n} < |\gamma(s)-\gamma_0(s)| < \overline{C}(s)} \frac{T_n(\gamma;s)}{|\gamma(s) - \gamma_0(s)|} < C_T(1-\eta(s))\right) \leq \varepsilon(s), \tag{A.12}$$

$$\mathbb{P}\left(\sup_{\overline{r}(s)\phi_{1n} < |\gamma(s)-\gamma_0(s)| < \overline{C}(s)} \frac{\overline{T}_n(\gamma;s)}{|\gamma(s) - \gamma_0(s)|} > C_{\overline{T}}(1+\eta(s))\right) \leq \varepsilon(s), \tag{A.13}$$

$$\mathbb{P}\left(\sup_{\overline{r}(s)\phi_{1n} < |\gamma(s)-\gamma_0(s)| < \overline{C}(s)} \frac{|L_{nj}(\gamma;s)|}{\sqrt{a_n}\,|\gamma(s) - \gamma_0(s)|} > \eta(s)\right) \leq \varepsilon(s) \tag{A.14}$$

for $j = 1, \ldots, \dim(x)$.

For a given $s \in \mathcal{S}_0$, we let $\widehat{\theta}(\widehat{\gamma}(s)) = (\widehat{\beta}(\widehat{\gamma}(s))^\top, \widehat{\delta}(\widehat{\gamma}(s))^\top)^\top$ and $\theta_0 = (\beta_0^\top, \delta_0^\top)^\top$.

Lemma A.7 *For a given $s \in \mathcal{S}_0$, $n^\epsilon(\widehat{\theta}(\widehat{\gamma}(s)) - \theta_0) = o_p(1)$.*

Proof of Theorem 2 The consistency is proved in Lemma A.5 above. For given $s \in \mathcal{S}_0$, we let

$$\begin{aligned} Q_n^*(\gamma(s); s) &= Q_n(\widehat{\beta}(\widehat{\gamma}(s)), \widehat{\delta}(\widehat{\gamma}(s)), \gamma(s); s) \\ &= \sum_{i \in \Lambda_n} \left\{ y_i - x_i^\top \widehat{\beta}(\widehat{\gamma}(s)) - x_i^\top \widehat{\delta}(\widehat{\gamma}(s)) \mathbf{1}_i(\gamma(s)) \right\}^2 K_i(s) \end{aligned} \quad \text{(A.15)}$$

for any $\gamma(\cdot)$, where $Q_n(\beta, \delta, \gamma; s)$ is the sum of squared errors function in (3). Consider $\gamma(s)$ such that $\gamma(s) \in [\gamma_0(s) + \overline{r}(s)\phi_{1n}, \gamma_0(s) + \overline{C}(s)]$ for some $0 < \overline{r}(s), \overline{C}(s) < \infty$ that are chosen in Lemma A.6. We let $\Delta_i(\gamma; s) = \mathbf{1}_i(\gamma(s)) - \mathbf{1}_i(\gamma_0(s))$. Then, since $y_i = x_i^\top \beta_0 + x_i^\top \delta_0 \mathbf{1}_i(\gamma_0(s_i)) + u_i$,

$$\begin{aligned} & Q_n^*(\gamma(s); s) - Q_n^*(\gamma_0(s); s) \\ = & \sum_{i \in \Lambda_n} \left\{ y_i - x_i^\top \widehat{\beta}(\widehat{\gamma}(s)) - x_i^\top \widehat{\delta}(\widehat{\gamma}(s)) \mathbf{1}_i(\gamma_0(s)) - x_i^\top \widehat{\delta}(\widehat{\gamma}(s)) \Delta_i(\gamma; s) \right\}^2 K_i(s) \\ & - \sum_{i \in \Lambda_n} \left\{ y_i - x_i^\top \widehat{\beta}(\widehat{\gamma}(s)) - x_i^\top \widehat{\delta}(\widehat{\gamma}(s)) \mathbf{1}_i(\gamma_0(s)) \right\}^2 K_i(s) \\ = & \sum_{i \in \Lambda_n} \left(x_i^\top \widehat{\delta}(\widehat{\gamma}(s)) \right)^2 \Delta_i(\gamma; s) K_i(s) \\ & -2 \sum_{i \in \Lambda_n} \left\{ u_i + x_i^\top \left(\beta_0 - \widehat{\beta}(\widehat{\gamma}(s)) \right) + x_i^\top \left(\delta_0 \mathbf{1}_i(\gamma_0(s_i)) - \widehat{\delta}(\widehat{\gamma}(s)) \mathbf{1}_i(\gamma_0(s)) \right) \right\} \\ & \qquad \times x_i^\top \widehat{\delta}(\widehat{\gamma}(s)) \Delta_i(\gamma; s) K_i(s) \\ = & \sum_{i \in \Lambda_n} \delta_0^\top x_i x_i^\top \delta_0 \Delta_i(\gamma; s) K_i(s) - \sum_{i \in \Lambda_n} \left\{ \left(x_i^\top \delta_0 \right)^2 - \left(x_i^\top \widehat{\delta}(\widehat{\gamma}(s)) \right)^2 \right\} \Delta_i(\gamma; s) K_i(s) \\ & -2 \sum_{i \in \Lambda_n} \widehat{\delta}(\widehat{\gamma}(s))^\top x_i u_i \Delta_i(\gamma; s) K_i(s) \\ & -2 \sum_{i \in \Lambda_n} \left(\beta_0 - \widehat{\beta}(\widehat{\gamma}(s)) \right)^\top x_i x_i^\top \widehat{\delta}(\widehat{\gamma}(s)) \Delta_i(\gamma; s) K_i(s) \\ & -2 \sum_{i \in \Lambda_n} \delta_0^\top x_i x_i^\top \delta_0 \left\{ \mathbf{1}_i(\gamma_0(s_i)) - \mathbf{1}_i(\gamma_0(s)) \right\} \Delta_i(\gamma; s) K_i(s) \qquad \text{(A.16)} \\ & -2 \sum_{i \in \Lambda_n} \delta_0^\top x_i x_i^\top \left(\widehat{\delta}(\widehat{\gamma}(s)) - \delta_0 \right) \left\{ \mathbf{1}_i(\gamma_0(s_i)) - \mathbf{1}_i(\gamma_0(s)) \right\} \Delta_i(\gamma; s) K_i(s) \\ & -2 \sum_{i \in \Lambda_n} \left(\delta_0 - \widehat{\delta}(\widehat{\gamma}(s)) \right)^\top x_i x_i^\top \widehat{\delta}(\widehat{\gamma}(s)) \mathbf{1}_i(\gamma_0(s)) \Delta_i(\gamma; s) K_i(s), \end{aligned}$$

where the second equality is because $\Delta_i(\gamma; s)^2 = \Delta_i(\gamma; s)$ as we consider the case $\gamma(s) > \gamma_0(s)$.

For any vector $v = (v_1, \ldots, v_{\dim(v)})^\top$, we let $\|v\|_\infty = \max_{1 \le j \le \dim(v)} |v_j|$. From Lemma A.7, we also let a sufficiently small $\kappa_n(s)$ such that $n^\epsilon \|\widehat{\theta}(\widehat{\gamma}(s)) - \theta_0\| \le \kappa_n(s)$ and $\kappa_n(s) \to 0$ as $n \to \infty$ for any s. We denote $\widehat{c}(\widehat{\gamma}(s))$ such that $\widehat{\delta}(\widehat{\gamma}(s)) = \widehat{c}(\widehat{\gamma}(s)) n^{-\epsilon}$, where $\delta_0 = c_0 n^{-\epsilon}$. Then, $\|\widehat{c}(\widehat{\gamma}(s)) - c_0\| \le \kappa_n(s)$, $\|\widehat{c}(\widehat{\gamma}(s))\| \le \|c_0\| + \kappa_n(s)$, and $\|\widehat{c}(\widehat{\gamma}(s)) + c_0\| \le 2\|c_0\| + \kappa_n(s)$. In addition, given Lemma A.6, there exist $0 < C(s), \overline{C}(s), \overline{r}(s), \eta(s), \varepsilon(s) < \infty$ such that

$$\begin{aligned}
\mathbb{P}\left(\inf_{\overline{r}(s)\phi_{1n} < |\gamma(s)-\gamma_0(s)| < \overline{C}(s)} \frac{T_n(\gamma; s)}{|\gamma(s) - \gamma_0(s)|} < C(s)(1-\eta(s)) \right) &\leq \frac{\varepsilon(s)}{3}, \\
\mathbb{P}\left(\sup_{\overline{r}(s)\phi_{1n} < |\gamma(s)-\gamma_0(s)| < \overline{C}(s)} \frac{\overline{T}_n(\gamma; s)}{|\gamma(s) - \gamma_0(s)|} > C_{\overline{T}}(1+\eta(s)) \right) &\leq \frac{\varepsilon(s)}{3}, \\
\mathbb{P}\left(\sup_{\overline{r}(s)\phi_{1n} < |\gamma(s)-\gamma_0(s)| < \overline{C}(s)} \frac{2\dim(x)\|c_0\|_\infty \|L_n(\gamma; s)\|_\infty}{\sqrt{a_n}|\gamma(s) - \gamma_0(s)|} > \eta(s) \right) &\leq \frac{\varepsilon(s)}{3}
\end{aligned}$$

for $\|c_0\|_\infty < \infty$. Furthermore, the term in line (A.16) satisfies

$$\begin{aligned}
& \frac{1}{a_n} \sum_{i \in \Lambda_n} \delta_0^\top x_i x_i^\top \delta_0 \left\{ \mathbf{1}_i(\gamma_0(s_i)) - \mathbf{1}_i(\gamma_0(s)) \right\} \Delta_i(\gamma; s) K_i(s) \\
\leq\ & \frac{1}{a_n} \sum_{i \in \Lambda_n} \delta_0^\top x_i x_i^\top \delta_0 \left| \mathbf{1}_i(\gamma_0(s_i)) - \mathbf{1}_i(\gamma_0(s)) \right| K_i(s) = C_n^*(s) b_n \qquad \text{(A.17)}
\end{aligned}$$

for some $C_n^*(s) = O_{a.s.}(1)$ as in Lemma A.4. For $\gamma(s) \in \left[\gamma_0(s) + \overline{r}(s)\phi_{1n}, \gamma_0(s) + \overline{C}(s)\right]$, we also have

$$\mathbb{P}\left(\sup_{\overline{r}(s)\phi_{1n} < |\gamma(s)-\gamma_0(s)| < \overline{C}(s)} \frac{2C_n^*(s) b_n}{|\gamma(s) - \gamma_0(s)|} > \eta(s) \right) \leq \frac{\varepsilon(s)}{3}$$

by choosing $\overline{r}(s)$ large enough, since

$$\sup_{\overline{r}(s)\phi_{1n} < |\gamma(s)-\gamma_0(s)| < \overline{C}(s)} \frac{C_n^*(s) b_n}{|\gamma(s) - \gamma_0(s)|} \leq \frac{C_n^*(s) b_n}{\overline{r}(s)\phi_{1n}} = a_n b_n \frac{C_n^*(s)}{\overline{r}(s)} < \infty$$

almost surely, where $a_n b_n = n^{1-2\epsilon} b_n^2 \to \varrho < \infty$.[10]

It follows that, with probability approaching to one,

$$\begin{aligned}
& \frac{Q_n^*(\gamma(s); s) - Q_n^*(\gamma_0(s); s)}{a_n(\gamma(s) - \gamma_0(s))} \qquad \text{(A.18)} \\
\geq\ & \frac{T_n(\gamma; s)}{\gamma(s) - \gamma_0(s)} - \|c_0 - \widehat{c}(\widehat{\gamma}(s))\| \|c_0 - \widehat{c}(\widehat{\gamma}(s))\| \frac{\overline{T}_n(\gamma, s)}{\gamma(s) - \gamma_0(s)} \\
& -2\dim(x) \|\widehat{c}(\widehat{\gamma}(s))\|_\infty \frac{\|L_n(\gamma; s)\|_\infty}{\sqrt{a_n}(\gamma(s) - \gamma_0(s))} \\
& -2 \left\| n^\epsilon(\beta_0 - \widehat{\beta}(\widehat{\gamma}(s))) \right\| \|\widehat{c}(\widehat{\gamma}(s))\| \frac{\overline{T}_n(\gamma, s)}{\gamma(s) - \gamma_0(s)} \\
& -2 \frac{C_n^*(s) b_n}{\gamma(s) - \gamma_0(s)} \\
& -2 \|c_0\| \|\widehat{c}(\widehat{\gamma}(s)) - c_0\| \frac{\overline{T}_n(\gamma, s)}{\gamma(s) - \gamma_0(s)} \qquad \text{(A.19)} \\
& -2 \left\| n^\epsilon(\delta_0 - \widehat{\delta}(\widehat{\gamma}(s))) \right\| \|\widehat{c}(\widehat{\gamma}(s))\| \frac{\overline{T}_n(\gamma, s)}{\gamma(s) - \gamma_0(s)} \qquad \text{(A.20)}
\end{aligned}$$

[10] Note that the term in line (A.16) is the source of the $O(b_n)$ bias of $\widehat{\gamma}(s)$, whereas $a_n = n^{1-2\epsilon} b_n$ is the order of the variance from Theorem 3. Therefore, the condition $a_n b_n = n^{1-2\epsilon} b_n^2 \to \varrho < \infty$ is to balance the bias-variance trade-off so that the bias term does not dominate in the limit.

$$
\begin{aligned}
\geq\ & C_T(s)(1-\eta(s)) - \kappa_n(s)\{2||c_0|| + \kappa_n(s)\} C_{\overline{T}}(s)(1+\eta(s)) \\
& -2\dim(x)\{\|c_0\|_\infty + \kappa_n(s)\}\eta(s) \\
& -2\kappa_n(s)\{\|c_0\| + \kappa_n(s)\} C_{\overline{T}}(s)(1+\eta(s)) \\
& -2\eta(s) - 2\|c_0\|\kappa_n(s) C_{\overline{T}}(s)(1+\eta(s)) \\
& -2\kappa_n(s)\{||c_0|| + \kappa_n(s)\} C_{\overline{T}}(s)(1+\eta(s)) \\
>\ & 0
\end{aligned}
$$

by choosing sufficiently small $\kappa_n(s)$ and $\eta(s)$, where the expressions in lines (A.19) and (A.20) are because $|\mathbf{1}_i(\gamma_0(s_i)) - \mathbf{1}_i(\gamma_0(s))| \leq 1$ and $|\mathbf{1}_i(\gamma_0(s))| \leq 1$.

Since we suppose $a_n(\gamma(s) - \gamma_0(s)) > 0$, it implies that, for any $\varepsilon(s) \in (0,1)$ and $\eta(s) > 0$,

$$
\mathbb{P}\left(\inf_{\overline{r}(s)\phi_{1n} < |\gamma(s)-\gamma_0(s)| < \overline{C}(s)} \{Q_n^*(\gamma(s);s) - Q_n^*(\gamma_0(s);s)\} > \eta(s) \right) \geq 1 - \varepsilon(s),
$$

which yields $\mathbb{P}(Q_n^*(\gamma(s);s) - Q_n^*(\gamma_0(s);s) > 0) \to 1$ as $n \to \infty$ for given $s \in \mathcal{S}_0$. We can similarly show the same result when $\gamma(s) \in \left[\gamma_0(s) - \overline{C}(s), \gamma_0(s) - \overline{r}(s)\phi_{1n}\right]$. Therefore, because $Q_n^*(\widehat{\gamma}(s);s) - Q_n^*(\gamma_0(s);s) \leq 0$ for any $s \in \mathcal{S}_0$ by construction, it should hold that $|\widehat{\gamma}(s) - \gamma_0(s)| \leq \overline{r}(s)\phi_{1n}$ with probability approaching to one; or for any $\varepsilon(s) > 0$ and $s \in \mathcal{S}_0$, there exists $\overline{r}(s) > 0$ such that

$$
\mathbb{P}(a_n|\widehat{\gamma}(s) - \gamma_0(s)| > \overline{r}(s)) < \varepsilon(s)
$$

for sufficiently large n, as required, since $\phi_{1n} = a_n^{-1}$. ■

A.3 Proof of Theorem 3 and Corollary 1 (Asymptotic Distribution)

For a given $s \in \mathcal{S}_0$,we define

$$
\begin{aligned}
A_n^*(r,s) &= \sum_{i\in\Lambda_n} \left(\delta_0^\top x_i\right)^2 |\mathbf{1}_i(\gamma_0(s) + r/a_n) - \mathbf{1}_i(\gamma_0(s))| K_i(s) \\
B_n^*(r,s) &= \sum_{i\in\Lambda_n} \delta_0^\top x_i u_i \{\mathbf{1}_i(\gamma_0(s) + r/a_n) - \mathbf{1}_i(\gamma_0(s))\} K_i(s)
\end{aligned}
$$

for some $0 < |r| < \infty$, where $a_n = n^{1-2\epsilon} b_n$ and ϵ is given in Assumption A-(ii).

Lemma A.8 *For fixed $s \in \mathcal{S}_0$, uniformly over r in any compact set,*

$$
A_n^*(r,s) \to_p |r|\, c_0^\top D(\gamma_0(s), s) c_0 f(\gamma_0(s), s)
$$

and

$$
B_n^*(r,s) \Rightarrow W(r) \sqrt{c_0^\top V(\gamma_0(s), s) c_0 f(\gamma_0(s), s) \kappa_2}
$$

as $n \to \infty$, where $\kappa_2 = \int K^2(v)dv$ and $W(r)$ is the two-sided Brownian Motion defined in (10).

For a given $s \in \mathcal{S}_0$, we let $\widehat{\theta}(\gamma_0(s)) = (\widehat{\beta}(\gamma_0(s))^\top, \widehat{\delta}(\gamma_0(s))^\top)^\top$. Recall that $\theta_0 = (\beta_0^\top, \delta_0^\top)^\top$ and $\widehat{\theta}(\widehat{\gamma}(s)) = (\widehat{\beta}(\widehat{\gamma}(s))^\top, \widehat{\delta}(\widehat{\gamma}(s))^\top)^\top$.

Lemma A.9 *For a given $s \in \mathcal{S}_0$, $\sqrt{nb_n}(\widehat{\theta}(\widehat{\gamma}(s)) - \theta_0) = O_p(1)$ and $\sqrt{nb_n}(\widehat{\theta}(\widehat{\gamma}(s)) - \widehat{\theta}(\gamma_0(s))) = o_p(1)$, if $n^{1-2\epsilon} b_n^2 \to \varrho < \infty$ as $n \to \infty$.*

Proof of Theorem 3 From Theorem 2, we define a random variable $r^*(s)$ such that

$$r^*(s) = a_n(\widehat{\gamma}(s) - \gamma_0(s)) = \arg\max_{r\in\mathbb{R}} \left\{ Q_n^*(\gamma_0(s); s) - Q_n^*\left(\gamma_0(s) + \frac{r}{a_n}; s\right) \right\},$$

where $Q_n^*(\gamma(s); s)$ is defined in (A.15). We let $\Delta_i(r,s) = \mathbf{1}_i(\gamma_0(s) + r/a_n) - \mathbf{1}_i(\gamma_0(s))$ as in Lemma A.8. We then have

$$\begin{aligned}
& \Delta Q_n^*(r; s) \\
= \; & Q_n^*(\gamma_0(s); s) - Q_n^*\left(\gamma_0(s) + \frac{r}{a_n}; s\right) \\
= \; & -\sum_{i\in\Lambda_n} \left(\widehat{\delta}(\widehat{\gamma}(s))^\top x_i\right)^2 |\Delta_i(r,s)| K_i(s) \\
& +2\sum_{i\in\Lambda_n} \left(y_i - \widehat{\beta}(\widehat{\gamma}(s))^\top x_i - \widehat{\delta}(\widehat{\gamma}(s))^\top x_i \mathbf{1}_i(\gamma_0(s))\right) \left(\widehat{\delta}(\widehat{\gamma}(s))^\top x_i\right) \Delta_i(r,s) K_i(s) \\
\equiv \; & -A_n(r; s) + 2B_n(r; s).
\end{aligned}$$

For $A_n(r; s)$, Lemmas A.8 and A.9 yield

$$\begin{aligned}
& A_n(r; s) \\
= \; & \sum_{i\in\Lambda_n} \left(\delta_0^\top x_i\right)^2 |\Delta_i(r,s)| K_i(s) + \sum_{i\in\Lambda_n} \left(\left(\widehat{\delta}(\widehat{\gamma}(s)) - \delta_0\right)^\top x_i\right)^2 |\Delta_i(r,s)| K_i(s) \\
= \; & \sum_{i\in\Lambda_n} \left(\delta_0^\top x_i\right)^2 |\Delta_i(r,s)| K_i(s) + \frac{1}{nb_n} \sum_{i\in\Lambda_n} \left(n^{-\epsilon} c^\top x_i\right)^2 |\Delta_i(r,s)| K_i(s) \\
= \; & A_n^*(r,s) + o_p(1)
\end{aligned}$$

for some finite vector c, since we have $\widehat{\delta}(\widehat{\gamma}(s)) - \delta_0 = n^{-\epsilon}(\widehat{c}(\widehat{\gamma}(s)) - c_0) = O_p((nb_n)^{-1/2})$ and $\sum_{i\in\Lambda_n} \left(n^{-\epsilon} c^\top x_i\right)^2 |\Delta_i(r,s)| K_i(s) = O_p(1)$ as $A_n^*(r,s)$. Similarly, for $B_n(r; s)$, since $y_i = \beta_0^\top x_i + \delta_0^\top x_i \mathbf{1}_i(\gamma_0(s_i)) + u_i$ and $\widehat{\beta}(\widehat{\gamma}(s)) - \beta_0 = O_p((nb_n)^{-1/2})$, we have

$$\begin{aligned}
& B_n(r; s) \\
= \; & \sum_{i\in\Lambda_n} \Bigg(u_i + \delta_0^\top x_i \left\{\mathbf{1}_i(\gamma_0(s_i)) - \mathbf{1}_i(\gamma_0(s))\right\} - \left(\widehat{\beta}(\widehat{\gamma}(s)) - \beta_0\right)^\top x_i \\
& \qquad - \left(\widehat{\delta}(\widehat{\gamma}(s)) - \delta_0\right)^\top x_i \mathbf{1}_i(\gamma_0(s)) \Bigg) \widehat{\delta}(\widehat{\gamma}(s))^\top x_i \Delta_i(r,s) K_i(s) \\
= \; & \sum_{i\in\Lambda_n} u_i \widehat{\delta}(\widehat{\gamma}(s))^\top x_i \Delta_i(r,s) K_i(s) \\
& + \sum_{i\in\Lambda_n} \delta_0^\top x_i \left\{\mathbf{1}_i(\gamma_0(s_i)) - \mathbf{1}_i(\gamma_0(s))\right\} \widehat{\delta}(\widehat{\gamma}(s))^\top x_i \Delta_i(r,s) K_i(s) \\
& - \sum_{i\in\Lambda_n} \left\{ \left(\widehat{\beta}(\widehat{\gamma}(s)) - \beta_0\right)^\top x_i + \left(\widehat{\delta}(\widehat{\gamma}(s)) - \delta_0\right)^\top x_i \mathbf{1}_i(\gamma_0(s)) \right\} \widehat{\delta}(\widehat{\gamma}(s))^\top x_i \Delta_i(r,s) K_i(s) \\
= \; & \sum_{i\in\Lambda_n} u_i \delta_0^\top x_i \Delta_i(r,s) K_i(s) + \sum_{i\in\Lambda_n} \delta_0^\top x_i \left\{\mathbf{1}_i(\gamma_0(s_i)) - \mathbf{1}_i(\gamma_0(s))\right\} \delta_0^\top x_i \Delta_i(r,s) K_i(s) + o_p(1) \\
= \; & B_n^*(r,s) + B_n^{**}(r,s) + o_p(1),
\end{aligned}$$

where we let

$$B_n^{**}(r,s) \equiv \sum_{i\in\Lambda_n} \delta_0^\top x_i \left\{\mathbf{1}_i\left(\gamma_0\left(s_i\right)\right) - \mathbf{1}_i\left(\gamma_0\left(s\right)\right)\right\} \delta_0^\top x_i \Delta_i(s) K_i\left(s\right).$$

In Lemma A.10 below, we show that, if $n^{1-2\epsilon}b_n^2 \to \varrho \in (0,\infty)$,

$$\begin{aligned} B_n^{**}(r,s) \to_p & \left|r\right| c_0^\top D\left(\gamma_0\left(s\right),s\right) c_0 f\left(\gamma_0\left(s\right),s\right) \left\{\frac{1}{2} - \mathcal{K}_0\left(r,\varrho;s\right)\right\} \\ & + \varrho c_0^\top D\left(\gamma_0\left(s\right),s\right) c_0 f\left(\gamma_0\left(s\right),s\right) \left|\dot\gamma_0(s)\right| \mathcal{K}_1\left(r,\varrho;s\right) \end{aligned}$$

as $n\to\infty$, where $\dot\gamma_0\left(\cdot\right)$ is the first derivative of $\gamma_0(\cdot)$ and $\mathcal{K}_j\left(r,\varrho;s\right) = \int_0^{|r|/(\varrho|\dot\gamma_0(s)|)} t^j K\left(t\right) dt$ for $j=0,1$.

From Lemma A.8, it follows that

$$\begin{aligned} \Delta Q_n^*(r;s) &= -A_n^*\left(r,s\right) + 2B_n^{**}(r,s) + 2B_n^*\left(r,s\right) \\ &\Rightarrow -\left|r\right| c_0^\top D\left(\gamma_0\left(s\right),s\right) c_0 f\left(\gamma_0\left(s\right),s\right) \\ &\quad + \left|r\right| c_0^\top D\left(\gamma_0\left(s\right),s\right) c_0 f\left(\gamma_0\left(s\right),s\right) \left\{1 - 2\mathcal{K}_0\left(r,\varrho;s\right)\right\} \\ &\quad + 2\varrho c_0^\top D\left(\gamma_0\left(s\right),s\right) c_0 f\left(\gamma_0\left(s\right),s\right) \left|\dot\gamma_0(s)\right| \mathcal{K}_1\left(r,\varrho;s\right) \\ &\quad + 2W\left(r\right)\sqrt{c_0^\top V\left(\gamma_0\left(s\right),s\right) c_0 f\left(\gamma_0\left(s\right),s\right)\kappa_2} \\ &= -2\left|r\right| \ell_D(s)\mathcal{K}_0\left(r,\varrho;s\right) + 2\varrho\ell_D(s)\left|\dot\gamma_0(s)\right| \mathcal{K}_1\left(r,\varrho;s\right) \\ &\quad + 2W\left(r\right)\sqrt{\ell_V(s)} \equiv \Delta Q^*(r;s), \end{aligned} \tag{A.21}$$

where

$$\begin{aligned} \ell_D(s) &= c_0^\top D\left(\gamma_0\left(s\right),s\right) c_0 f\left(\gamma_0\left(s\right),s\right), \\ \ell_V(s) &= c_0^\top V\left(\gamma_0\left(s\right),s\right) c_0 f\left(\gamma_0\left(s\right),s\right)\kappa_2. \end{aligned}$$

Note that $n^{1-2\epsilon}b_n\left(\widehat\gamma\left(s\right) - \gamma_0\left(s\right)\right) = \arg\max_{r\in\mathbb{R}} \Delta Q_n^*(r;s) = O_p(1)$ from Theorem 2 and we showed $\Delta Q_n^*(r;s) \Rightarrow \Delta Q^*(r;s)$ for any $s\in\mathcal{S}_0$, which is continuous in r, has a unique maximum, and $\lim_{|r|\to\infty}\Delta Q^*(r;s) = -\infty$ almost surely. Similar to the proof of Theorem 1 in Hansen (2000), if we let $\xi(s) = \ell_V(s)/\ell_D^2(s) > 0$ and $r = \xi(s)\nu$, we have

$$\begin{aligned} & \arg\max_{r\in\mathbb{R}} \Delta Q^*(r;s) \\ = \;& \arg\max_{r\in\mathbb{R}} \Big(2W\left(r\right)\sqrt{\ell_V(s)} - 2\left|r\right|\ell_D(s)\mathcal{K}_0\left(r,\varrho;s\right) + 2\varrho\ell_D(s)\left|\dot\gamma_0(s)\right|\mathcal{K}_1\left(r,\varrho;s\right)\Big) \\ = \;& \xi(s)\arg\max_{\nu\in\mathbb{R}} \Big(W\left(\xi(s)\nu\right)\sqrt{\ell_V(s)} - \left|\xi(s)\nu\right|\ell_D(s)\mathcal{K}_0\left(\xi(s)\nu,\varrho;s\right) + \varrho\ell_D(s)\left|\dot\gamma_0(s)\right|\mathcal{K}_1\left(\xi(s)\nu,\varrho;s\right)\Big) \\ = \;& \xi(s)\arg\max_{\nu\in\mathbb{R}} \left(W\left(\nu\right)\frac{\ell_V(s)}{\ell_D(s)} - \left|\nu\right|\frac{\ell_V(s)}{\ell_D(s)}\mathcal{K}_0\left(\xi(s)\nu,\varrho;s\right) + \varrho\frac{\ell_V(s)}{\ell_D(s)}\cdot\frac{\left|\dot\gamma_0(s)\right|}{\xi(s)}\mathcal{K}_1\left(\xi(s)\nu,\varrho;s\right)\right) \\ = \;& \xi(s)\arg\max_{\nu\in\mathbb{R}} \left(W\left(\nu\right) - \left|\nu\right|\mathcal{K}_0\left(\xi(s)\nu,\varrho;s\right) + \varrho\frac{\left|\dot\gamma_0(s)\right|}{\xi(s)}\mathcal{K}_1\left(\xi(s)\nu,\varrho;s\right)\right). \end{aligned}$$

By Theorem 2.7 of Kim and Pollard (1990), it thus follows that (rewriting ν as r)

$$n^{1-2\epsilon}b_n\left(\widehat\gamma\left(s\right) - \gamma_0\left(s\right)\right) \to_d \xi\left(s\right)\arg\max_{r\in\mathbb{R}} \left(W\left(r\right) - \left|r\right|\psi_0\left(r,\varrho;s\right) + \varrho\frac{\left|\dot\gamma_0(s)\right|}{\xi(s)}\psi_1\left(r,\varrho;s\right)\right)$$

as $n \to \infty$, where

$$\psi_j(r, \varrho; s) = \int_0^{|r|\xi(s)/(\varrho|\dot{\gamma}_0(s)|)} t^j K(t)\, dt$$

for $j = 0, 1$. Finally, letting

$$\mu(r, \varrho; s) = -|r|\, \psi_0(r, \varrho; s) + \varrho \frac{|\dot{\gamma}_0(s)|}{\xi(s)} \psi_1(r, \varrho; s), \tag{A.22}$$

$\mathbb{E}[\arg\max_{r\in\mathbb{R}} (W(r) + \mu(r, \varrho; s))] = 0$ follows from Lemmas A.11 and A.12 below. ■

Lemma A.10 *For a given $s \in \mathcal{S}_0$, let r be the same term used in Lemma A.8. If $n^{1-2\epsilon} b_n^2 \to \varrho \in (0, \infty)$, uniformly over r in any compact set,*

$$\begin{aligned} B_n^{**}(r, s) &\equiv \sum_{i\in\Lambda_n} \delta_0^\top x_i \{\mathbf{1}_i(\gamma_0(s_i)) - \mathbf{1}_i(\gamma_0(s))\} \delta_0^\top x_i \Delta_i(s) K_i(s) \\ &\to_p |r|\, c_0^\top D(\gamma_0(s), s)\, c_0 f(\gamma_0(s), s) \left\{\frac{1}{2} - \mathcal{K}_0(r, \varrho; s)\right\} \\ &\quad + \varrho c_0^\top D(\gamma_0(s), s)\, c_0 f(\gamma_0(s), s)\, |\dot{\gamma}_0(s)|\, \mathcal{K}_1(r, \varrho; s) \end{aligned}$$

as $n \to \infty$, where $\dot{\gamma}_0(\cdot)$ is the first derivatives of $\gamma_0(\cdot)$ and

$$\mathcal{K}_j(r, \varrho; s) = \int_0^{|r|/(\varrho|\dot{\gamma}_0(s)|)} t^j K(t)\, dt$$

for $j = 0, 1$.

Lemma A.11 *Let $\tau = \arg\max_{r\in\mathbb{R}} (W(r) + \mu(r))$, where $W(r)$ is a two-sided Brownian motion in (10) and $\mu(r)$ is a continuous and symmetric function satisfying: $\mu(0) = 0$, $\mu(-r) = \mu(r)$, $\mu(r)/r^{1/2+\varepsilon}$ is monotonically decreasing to $-\infty$ on $[\underline{r}, \infty)$ for some $\underline{r} > 0$ and $\varepsilon > 0$. Then, $\mathbb{E}[\tau] = 0$.*

Lemma A.12 *For any given $\varrho < \infty$ and $s \in \mathcal{S}_0$, $\mu(r, \varrho; s)$ in (A.22) satisfies conditions in Lemma A.11.*

Proof of Corollary 1 Under $H_0 : \gamma_0(s) = \gamma_*(s)$, we write

$$LR_n(s) = \frac{1}{nb_n} \sum_{i\in\Lambda_n} K\left(\frac{s_i - s}{b_n}\right) \times \frac{\{Q_n(\gamma_*(s), s) - Q_n(\widehat{\gamma}(s), s)\}}{(nb_n)^{-1} Q_n(\widehat{\gamma}(s), s)},$$

where $Q_n(\gamma; s) = Q_n(\widehat{\beta}(\gamma; s), \widehat{\delta}(\gamma; s), \gamma; s)$ defined in (5). From (A.5) and (A.7), we have

$$\frac{1}{nb_n} Q_n(\widehat{\gamma}(s), s) = \frac{1}{nb_n} \sum_{i\in\Lambda_n} u_i^2 K_i(s) + O_p(n^{-2\epsilon}) \to_p \mathbb{E}[u_i^2 | s_i = s]\, f_s(s)$$

as $n \to \infty$, where $f_s(s)$ is the marginal density of s_i. In addition, from Lemmas A.3 and A.9, we have

$$\begin{aligned} & Q_n(\gamma_0(s), s) - Q_n(\widehat{\gamma}(s), s) \\ = \;& Q_n^*(\gamma_0(s), s) - Q_n^*(\widehat{\gamma}(s), s) \\ & + \left(\widehat{\theta}(\widehat{\gamma}(s)) - \widehat{\theta}(\gamma_0(s))\right)^\top \widetilde{Z}(\gamma_0(s); s) \widetilde{Z}(\gamma_0(s); s)^\top \left(\widehat{\theta}(\widehat{\gamma}(s)) - \widehat{\theta}(\gamma_0(s))\right) \\ = \;& Q_n^*(\gamma_0(s), s) - Q_n^*(\widehat{\gamma}(s), s) + o_p(1), \end{aligned}$$

where $Q_n^*(\gamma, s) = Q_n(\widehat{\beta}(\widehat{\gamma}(s)), \widehat{\delta}(\widehat{\gamma}(s)), \gamma; s)$ defined in (A.15) and $\widetilde{Z}(\gamma; s)$ is defined in Lemma A.5. Similar to Theorem 2 of Hansen (2000), the rest of the proof follows from the change of variables and the continuous mapping theorem using the limiting expression in (A.21) and $(nb_n)^{-1}\sum_{i\in\Lambda_n} K_i(s) \to_p f_s(s)$ by the standard result of the kernel density estimator. ■

A.4 Proof of Theorem 4 (Uniform Convergence)

We let $\phi_{2n} = \log n/a_n$, where $a_n = n^{1-2\epsilon}b_n$ and ϵ is given in Assumption A-(ii). We also define $\mathcal{G}_n(\mathcal{S}_0; \Gamma)$ as a class of cadlag and piecewise constant functions $\mathcal{S}_0 \mapsto \Gamma$ with at most n discontinuity points. Recall that $T_n(\gamma; s)$, $\overline{T}_n(\gamma; s)$, and $L_{nj}(\gamma; s)$ are defined in (A.9), (A.10), and (A.11), respectively; $\sup_{s\in\mathcal{S}_0} |\gamma(s) - \gamma_0(s)|$ is bounded since $\gamma(s) \in \Gamma$, a compact set, for any $s \in \mathcal{S}_0$.

Lemma A.13 *For any $\eta > 0$, and $\varepsilon > 0$, there exist constants $\overline{C}$, $\overline{r}$, C_T, and $C_{\overline{T}}$ such that if $n^{1-2\epsilon}b_n^2 \to \varrho < \infty$ and n is sufficiently large,*

$$\mathbb{P}\left(\inf_{\substack{\{\gamma(\cdot)\in\mathcal{G}_n(\mathcal{S}_0;\Gamma):\\ \overline{r}\phi_{2n}<\sup_{s\in\mathcal{S}_0}|\gamma(s)-\gamma_0(s)|<\overline{C}\}}} \frac{\sup_{s\in\mathcal{S}_0} T_n(\gamma; s)}{\sup_{s\in\mathcal{S}_0} |\gamma(s) - \gamma_0(s)|} < C_T(1-\eta)\right) \leq \varepsilon, \qquad \text{(A.23)}$$

$$\mathbb{P}\left(\sup_{\substack{\{\gamma(\cdot)\in\mathcal{G}_n(\mathcal{S}_0;\Gamma):\\ \overline{r}\phi_{2n}<\sup_{s\in\mathcal{S}_0}|\gamma(s)-\gamma_0(s)|<\overline{C}\}}} \frac{\sup_{s\in\mathcal{S}_0} \overline{T}_n(\gamma; s)}{\sup_{s\in\mathcal{S}_0} |\gamma(s) - \gamma_0(s)|} > C_{\overline{T}}(1+\eta)\right) \leq \varepsilon, \qquad \text{(A.24)}$$

$$\mathbb{P}\left(\sup_{\substack{\{\gamma(\cdot)\in\mathcal{G}_n(\mathcal{S}_0;\Gamma):\\ \overline{r}\phi_{2n}<\sup_{s\in\mathcal{S}_0}|\gamma(s)-\gamma_0(s)|<\overline{C}\}}} \frac{\sup_{s\in\mathcal{S}_0} |L_{nj}(\gamma; s)|}{\sqrt{a_n}\sup_{s\in\mathcal{S}_0} |\gamma(s) - \gamma_0(s)|} > \eta\right) \leq \varepsilon \qquad \text{(A.25)}$$

for $j = 1, \ldots, \dim(x)$.

Lemma A.14 $\sup_{s\in\mathcal{S}_0} |\widehat{\gamma}(s) - \gamma_0(s)| = o_p(1)$ *and* $n^\epsilon \sup_{s\in\mathcal{S}_0} ||\widehat{\theta}(\widehat{\gamma}(s)) - \theta_0|| = o_p(1)$.

Proof of Theorem 4 Note that $\widehat{\gamma}(\cdot)$ belongs to $\mathcal{G}_n(\mathcal{S}_0; \Gamma)$. For $Q_n^*(\cdot; \cdot)$ defined in (A.15), since $\sup_{s\in\mathcal{S}_0} (Q_n^*(\widehat{\gamma}(s); s) - Q_n^*(\gamma_0(s); s)) \leq 0$ by construction, it suffices to show that as $n \to \infty$,

$$\mathbb{P}\left(\inf_{\substack{\{\gamma(\cdot)\in\mathcal{G}_n(\mathcal{S}_0;\Gamma):\\ \overline{r}\phi_{2n}\leq\sup_{s\in\mathcal{S}_0}|\gamma(s)-\gamma_0(s)|\leq\overline{C}\}}} \sup_{s\in\mathcal{S}_0} \{Q_n^*(\gamma(s); s) - Q_n^*(\gamma_0(s); s)\} > 0\right) \to 1,$$

where $\overline{r}$ is chosen in Lemma A.13.

To this end, consider $\gamma(\cdot)$ such that $\overline{r}\phi_{2n} \leq \sup_{s\in\mathcal{S}_0} |\gamma(s) - \gamma_0(s)| \leq \overline{C}$ for some $0 < \overline{r}, \overline{C} < \infty$. Then, similarly as (A.18) and using Lemmas A.13 and A.14, we have that for a sufficiently large n

$$\frac{Q_n^*(\gamma(s); s) - Q_n^*(\gamma_0(s); s)}{a_n \sup_{s\in\mathcal{S}_0} |\gamma(s) - \gamma_0(s)|}$$

$$
\begin{aligned}
&\geq \frac{T_n(\gamma; s)}{\sup_{s\in\mathcal{S}_0}|\gamma(s)-\gamma_0(s)|} - \kappa_n(s)\{2||c_0|| + \kappa_n(s)\}\frac{\overline{T}_n(\gamma, s)}{\sup_{s\in\mathcal{S}_0}|\gamma(s)-\gamma_0(s)|} \\
&\quad -2\dim(x)\{\|c_0\|_\infty + \kappa_n(s)\}\frac{\|L_n(\gamma; s)\|_\infty}{\sqrt{a_n}\sup_{s\in\mathcal{S}_0}|\gamma(s)-\gamma_0(s)|} \\
&\quad -2\kappa_n(s)\{\|c_0\| + \kappa_n(s)\}\frac{\overline{T}_n(\gamma, s)}{\sup_{s\in\mathcal{S}_0}|\gamma(s)-\gamma_0(s)|} - \frac{2C_n^*(s)b_n}{\sup_{s\in\mathcal{S}_0}|\gamma(s)-\gamma_0(s)|} \\
&\quad -2\|c_0\|\kappa_n(s)\frac{\overline{T}_n(\gamma, s)}{\sup_{s\in\mathcal{S}_0}|\gamma(s)-\gamma_0(s)|} - 2\kappa_n(s)\{||c_0|| + \kappa_n(s)\}\frac{\overline{T}_n(\gamma, s)}{\sup_{s\in\mathcal{S}_0}|\gamma(s)-\gamma_0(s)|} \\
&> 0,
\end{aligned}
$$

where $n^\epsilon||\widehat{\theta}(\widehat{\gamma}(s)) - \theta_0|| \leq \kappa_n(s)$ and $\kappa_n(s) \to 0$ as $n \to \infty$ uniformly in s given Lemma A.14 and all the notations are the same as in (A.18). Note that the $C_n^*(s)$ term in (A.17) satisfies $\sup_{s\in\mathcal{S}_0} C_n^*(s) = O_{a.s.}(1)$ from Lemma A.4, and

$$
\begin{aligned}
\sup_{\overline{r}\phi_{2n}<|\gamma(s)-\gamma_0(s)|<\overline{C}} \frac{\sup_{s\in\mathcal{S}_0} C_n^*(s)\, b_n}{\sup_{s\in\mathcal{S}_0}|\gamma(s)-\gamma_0(s)|} &< \frac{\sup_{s\in\mathcal{S}_0} C_n^*(s) b_n}{\overline{r}\phi_{2n}} \\
&= \frac{\sup_{s\in\mathcal{S}_0} C_n^*(s)}{\overline{r}}\left(\frac{a_n b_n}{\log n}\right) \\
&= o_{a.s.}(1)
\end{aligned}
$$

given $a_n b_n \to \varrho < \infty$. Thus, we have

$$
\mathbb{P}\left(\sup_{\overline{r}\phi_{2n}<|\gamma(s)-\gamma_0(s)|<\overline{C}} \frac{2\sup_{s\in\mathcal{S}_0} C^*(s) b_n}{\sup_{s\in\mathcal{S}_0}|\gamma(s)-\gamma_0(s)|} > \eta\right) \leq \frac{\varepsilon}{3}
$$

when n is sufficiently large. Therefore, for any $\varepsilon \in (0, 1)$ and $\eta > 0$,

$$
\mathbb{P}\left(\inf_{\overline{r}\phi_{2n}<\sup_{s\in\mathcal{S}_0}|\gamma(s)-\gamma_0(s)|<\overline{C}} \sup_{s\in\mathcal{S}_0}\{Q_n^*(\gamma(s); s) - Q_n^*(\gamma_0(s); s)\} > \eta\right) \geq 1 - \varepsilon,
$$

which completes the proof by the same argument as Theorem 2. ■

A.5 Proof of Theorem 5 (Asymptotic Normality of $\widehat{\theta}$)

Proof of Theorem 5 We let $\mathbf{1}_{\mathcal{S}_0} = \mathbf{1}[s_i \in \mathcal{S}_0]$ and consider a sequence of positive constants $\pi_n \to 0$ as $n \to \infty$. Then, since $y_i = x_i^\top\beta_0 + x_i^\top\delta_0\mathbf{1}[q_i \leq \gamma_0(s_i)] + u_i = x_i^\top\delta_0^* - x_i^\top\delta_0\mathbf{1}[q_i > \gamma_0(s_i)] + u_i$ for $\delta_0^* = \beta_0 + \delta_0$,

$$
\begin{aligned}
\sqrt{n}\left(\widehat{\beta} - \beta_0\right) &= \left(\frac{1}{n}\sum_{i\in\Lambda_n} x_i x_i^\top \mathbf{1}[q_i > \widehat{\gamma}(s_i) + \pi_n]\mathbf{1}_{\mathcal{S}_0}\right)^{-1} \\
&\quad \times\Bigg\{\frac{1}{\sqrt{n}}\sum_{i\in\Lambda_n} x_i u_i \mathbf{1}[q_i > \gamma_0(s_i) + \pi_n]\mathbf{1}_{\mathcal{S}_0} \\
&\quad\quad + \frac{1}{\sqrt{n}}\sum_{i\in\Lambda_n} x_i u_i\{\mathbf{1}[q_i > \widehat{\gamma}(s_i) + \pi_n] - \mathbf{1}[q_i > \gamma_0(s_i) + \pi_n]\}\mathbf{1}_{\mathcal{S}_0}
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \frac{1}{\sqrt{n}} \sum_{i \in \Lambda_n} x_i x_i^\top \delta_0 \mathbf{1}\left[q_i \le \gamma_0(s_i)\right] \mathbf{1}\left[q_i > \widehat{\gamma}(s_i) + \pi_n\right] \mathbf{1}_{\mathcal{S}_0} \Bigg\} \\
&\equiv \Xi_{\beta 0}^{-1} \left\{ \Xi_{\beta 1} + \Xi_{\beta 2} + \Xi_{\beta 3} \right\} \qquad \text{(A.26)}
\end{aligned}
$$

and

$$
\begin{aligned}
\sqrt{n}\left(\widehat{\delta}^* - \delta_0^*\right) &= \left(\frac{1}{n} \sum_{i \in \Lambda_n} x_i x_i^\top \mathbf{1}\left[q_i < \widehat{\gamma}(s_i) - \pi_n\right] \mathbf{1}_{\mathcal{S}_0} \right)^{-1} \\
&\quad \times \Bigg\{ \frac{1}{\sqrt{n}} \sum_{i \in \Lambda_n} x_i u_i \mathbf{1}\left[q_i < \gamma_0(s_i) - \pi_n\right] \mathbf{1}_{\mathcal{S}_0} \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i \in \Lambda_n} x_i u_i \left\{ \mathbf{1}\left[q_i < \widehat{\gamma}(s_i) - \pi_n\right] - \mathbf{1}\left[q_i < \gamma_0(s_i) - \pi_n\right] \right\} \mathbf{1}_{\mathcal{S}_0} \\
&\quad - \frac{1}{\sqrt{n}} \sum_{i \in \Lambda_n} x_i x_i^\top \delta_0 \mathbf{1}\left[q_i > \gamma_0(s_i)\right] \mathbf{1}\left[q_i < \widehat{\gamma}(s_i) - \pi_n\right] \mathbf{1}_{\mathcal{S}_0} \Bigg\} \\
&\equiv \Xi_{\delta 0}^{-1} \left\{ \Xi_{\delta 1} + \Xi_{\delta 2} - \Xi_{\delta 3} \right\}, \qquad \text{(A.27)}
\end{aligned}
$$

where $\Xi_{\beta 2}$, $\Xi_{\beta 3}$, $\Xi_{\delta 2}$, and $\Xi_{\delta 3}$ are all $o_p(1)$ from Lemma A.15 below, provided $\phi_{2n}/\pi_n \to 0$ as $n \to \infty$. Therefore,

$$
\sqrt{n}\left(\widehat{\theta}^* - \theta_0^*\right) = \begin{pmatrix} \Xi_{\beta 0} & 0 \\ 0 & \Xi_{\delta 0} \end{pmatrix}^{-1} \begin{pmatrix} \Xi_{\beta 1} \\ \Xi_{\delta 1} \end{pmatrix} + o_p(1)
$$

and the desired result follows once we establish that

$$
\Xi_{\beta 0} \to_p \mathbb{E}\left[x_i x_i^\top \mathbf{1}\left[q_i > \gamma_0(s_i)\right] \mathbf{1}_{\mathcal{S}_0} \right], \qquad \text{(A.28)}
$$

$$
\Xi_{\delta 0} \to_p \mathbb{E}\left[x_i x_i^\top \mathbf{1}\left[q_i \le \gamma_0(s_i)\right] \mathbf{1}_{\mathcal{S}_0} \right], \qquad \text{(A.29)}
$$

and

$$
\begin{pmatrix} \Xi_{\beta 1} \\ \Xi_{\delta 1} \end{pmatrix} \to_d \mathcal{N}\left(0, \lim_{n \to \infty} \frac{1}{n} Var\left[\begin{pmatrix} \sum_{i \in \Lambda_n} x_i u_i \mathbf{1}\left[q_i > \gamma_0(s_i)\right] \mathbf{1}_{\mathcal{S}_0} \\ \sum_{i \in \Lambda_n} x_i u_i \mathbf{1}\left[q_i \le \gamma_0(s_i)\right] \mathbf{1}_{\mathcal{S}_0} \end{pmatrix} \right] \right) \qquad \text{(A.30)}
$$

as $n \to \infty$.

First, by Assumptions A-(v) and (ix), (A.28) can be readily verified since we have

$$
\begin{aligned}
&\frac{1}{n} \sum_{i \in \Lambda_n} x_i x_i^\top \mathbf{1}\left[q_i > \widehat{\gamma}(s_i) + \pi_n\right] \mathbf{1}_{\mathcal{S}_0} \\
&= \frac{1}{n} \sum_{i \in \Lambda_n} x_i x_i^\top \mathbf{1}\left[q_i > \gamma_0(s_i) + \pi_n\right] \mathbf{1}_{\mathcal{S}_0} \\
&\quad + \frac{1}{n} \sum_{i \in \Lambda_n} x_i x_i^\top \left\{ \mathbf{1}\left[q_i > \widehat{\gamma}(s_i) + \pi_n\right] - \mathbf{1}\left[q_i > \gamma_0(s_i) + \pi_n\right] \right\} \mathbf{1}_{\mathcal{S}_0} \\
&= \frac{1}{n} \sum_{i \in \Lambda_n} x_i x_i^\top \mathbf{1}\left[q_i > \gamma_0(s_i) + \pi_n\right] \mathbf{1}_{\mathcal{S}_0} + O_p(\phi_{2n})
\end{aligned}
$$

with $\pi_n \to 0$ as $n \to \infty$. More precisely, given Theorem 4, we consider $\widehat{\gamma}(s)$ in a neighborhood of $\gamma_0(s)$ with uniform distance at most $\overline{r}\phi_{2n}$ for some large enough constant $\overline{r}$. We define a non-random function $\widetilde{\gamma}(s) = \gamma_0(s) + \overline{r}\phi_{2n}$. Then, on the event $E_n^* = \{\sup_{s\in\mathcal{S}_0} |\widehat{\gamma}(s) - \gamma_0(s)| \leq \overline{r}\phi_{2n}\}$,

$$\begin{aligned}
& \mathbb{E}\left[x_i x_i^\top \{\mathbf{1}[q_i > \widehat{\gamma}(s_i) + \pi_n] - \mathbf{1}[q_i > \gamma_0(s_i) + \pi_n]\} \mathbf{1}_{\mathcal{S}_0}\right] \\
\leq \; & \mathbb{E}\left[x_i x_i^\top \{\mathbf{1}[q_i > \widetilde{\gamma}(s_i) + \pi_n] - \mathbf{1}[q_i > \gamma_0(s_i) + \pi_n]\} \mathbf{1}_{\mathcal{S}_0}\right] \\
= \; & \int_{\mathcal{S}_0} \int_{\gamma_0(v)+\pi_n}^{\widetilde{\gamma}(v)+\pi_n} D(q,v) f(q,v)\, dq dv \\
= \; & \int_{\mathcal{S}_0} \{D(\gamma_0(v), v) f(\gamma_0(v), v)(\widetilde{\gamma}(v) - \gamma_0(v)) + o_p(\phi_{2n})\} dv \\
\leq \; & \overline{r}\phi_{2n} \int D(\gamma_0(v), v) f(\gamma_0(v), v)\, dv \\
= \; & O_p(\phi_{2n}) = o_p(1)
\end{aligned}$$

from Theorem 4, Assumptions A-(v), (vii), and (ix). (A.29) can be verified symmetrically. Using a similar argument, since $\mathbb{E}[x_i u_i \mathbf{1}[q_i > \gamma_0(s_i)] \mathbf{1}_{\mathcal{S}_0}] = \mathbb{E}[x_i u_i \mathbf{1}[q_i \leq \gamma_0(s_i)] \mathbf{1}_{\mathcal{S}_0}] = 0$ from Assumption ID-(i), the asymptotic normality in (A.30) follows by the Theorem of Bolthausen (1982) under Assumption A-(iii), which completes the proof. ■

Lemma A.15 *When $\phi_{2n} \to 0$ as $n \to \infty$, if we let $\pi_n > 0$ such that $\pi_n \to 0$ and $\phi_{2n}/\pi_n \to 0$ as $n \to \infty$, then $\Xi_{\beta 2}$, $\Xi_{\beta 3}$, $\Xi_{\delta 2}$, and $\Xi_{\delta 3}$ in (A.26) and (A.27) are all $o_p(1)$.*

Acknowledgement

We are grateful to Xiaohong Chen, Jonathan Dingel, Bo Honoré, Sokbae Lee, Yuan Liao, Francesca Molinari, Ingmar Prucha, Myung Hwan Seo, Ping Yu, and participants at numerous seminar and conference presentations for very helpful comments. This work was supported by the Appleby-Mosher grant and the CUSE grant, Syracuse University.

References

Ananat, E. O. (2011): "The Wrong Side(s) of the Tracks: The Causal Effects of Racial Segregation on Urban Poverty and Inequality," *American Economic Journal: Applied Economics*, 3(2), 34–66.

Andrews, D. W. K. (1994): "Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity," *Econometrica*, 62(1), 43–72.

Bai, J. (1997): "Estimation of a Change Point in Multiple Regressions," *Review of Economics and Statistics*, 79(4), 551–563.

Bai, J., and P. Perron (1998): "Estimating and Testing Linear Models with Multiple Structural Changes," *Econometrica*, 66(1), 47–78.

Baragwanath, K., R. Goldblatt, G. Hanson, and A. K. Khandelwal (2021): "Detecting Urban Markets with Satellite Imagery: An Application to India," *Journal of Urban Economics*, 125, 103–173.

Bhattacharya, P. K., and P. J. Brockwell (1976): "The Minimum of an Additive Process with Applications to Signal Estimation and Storage Theory," *Z. Wahrsch. Verw. Gebiete,*, 37, 51–75.

Billingsley, P. (1968): *Convergence of Probability Measure*. Wiley, New York.

Bolthausen, E. (1982): "On the central limit theorem for stationary mixing random fields," *Annuals of Probability*, 10(4), 1047–1050.

Caner, M., and B. E. Hansen (2004): "Instrumental Variable Estimation of a Threshold Model," *Econometric Theory*, 20(5), 813–843.

Carbon, M., C. Francq, and L. T. Tran (2007): "Kernel Regression Estimation for Random Fields," *Journal of Statistical Planning and Inference*, 137(3), 778–798.

Card, D., A. Mas, and J. Rothstein (2008): "Tipping and the Dynamics of Segregation," *Quarterly Journal of Economics*, 123(1), 177–218.

Chan, K. S. (1993): "Consistency and Limiting Distribution of the Least Squares Estimator of a Threshold Autoregressive Model," *Annals of Statistics*, 21(1), 520–533.

Chiou, Y., M. Chen, and J. Chen (2018): "Nonparametric Regression with Multiple Thresholds: Estimation and Inference," *Journal of Econometrics*, 206(2), 472–514.

Conley, T. G. (1999): "GMM Estimation with Cross Sectional Dependence," *Journal of Econometrics*, 92(1), 1–45.

Conley, T. G., and F. Molinari (2007): "Spatial Correlation Robust Inference with Errors in Location or Distance," *Journal of Econometrics*, 140(1), 76–96.

Darity, W. A., and P. L. Mason (1998): "Evidence on Discrimination in Employment: Codes of Color, Codes of Gender," *Journal of Economic Perspectives*, 12(2), 63–90.

Delgado, M. A., and J. Hidalgo (2000): "Nonparametric Inference on Structural Break," *Journal of Econometrics*, 96(1), 113–144.

Dingel, J. I., A. Miscio, and D. R. Davis (2021): "Cities, Lights, and Skills in Developing Economics," *Journal of Urban Economics*, 125, 103–174.

Hall, P., and C. C. Heyde (1980): *Martingale Limit Theory and its Applications*. Academic Press, New York.

Hansen, B. E. (2000): "Sample Splitting and Threshold Estimation," *Econometrica*, 68(3), 575–603.

Heilmann, K. (2018): "Transit Access and Neighborhood Segregation. Evidence from the Dallas Light Rail System," *Regional Science and Urban Economics*, 73, 237–250.

Henderson, D. J., C. F. Parmeter, and L. Su (2017): "Nonparametric Threshold Regression: Estimation and Inference," *Working Paper*.

Henderson, J. V., A. Storeygard, and D. N. Weil (2012): "Measuring Economic Growth from Outer Space," *American Economic Review*, 102(2), 994–1028.

Hidalgo, J., J. Lee, and M. H. Seo (2019): "Robust Inference for Threshold Regression Models," *Journal of Econometrics*, 210(2), 291–309.

Jenish, N., and I. R. Prucha (2009): "Central Limit Theorems and Uniform Laws of Large Numbers for Arrays of Random Fields," *Journal of Econometrics*, 150(1), 86–98.

Kim, J., and D. Pollard (1990): "Cube Root Asymptotics," *Annals of Statistics*, 18(1), 191–219.

Lee, S., Y. Liao, M. H. Seo, and Y. Shin (2021): "Factor-driven Two-regime Regression," *Annuals of Statistics*, 49(3), 1656–1678.

Lee, S., M. H. Seo, and Y. Shin (2011): "Testing for Threshold Effects in Regression Models," *Journal of the American Statistical Association*, 106(493), 220–231.

Lee, Y., and Y. Wang (2022): "Testing for Homogeneous Thresholds in Threshold Regression Models," *Working Paper*.

Li, D., and S. Ling (2012): "On the Least Squares Estimation of Multiple-Regime Threshold Autoregressive Models," *Journal of Econometrics*, 167(1), 240–253.

Newey, W. K., and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of Econometrics*, 4, 2111–2245.

Rozenfeld, H. D., D. Rybski, X. Gabaix, and H. A. Makse (2011): "The Area and Population of Cities: New Insights from a Different Perspective on Cities," *American Economic Review*, 101(5), 2205–2225.

Schelling, T. C. (1971): "Dynamic Models of Segregation," *Journal of Mathematical Sociology*, 1(2), 143–186.

Seo, M. H., and O. Linton (2007): "A Smooth Least Squares Estimator for Threshold Regression Models," *Journal of Econometrics*, 141(2), 704–735.

Tong, H. (1983): *Threshold models in nonlinear time series analysis (Lecture Notes in Statistics No. 21)*. New York: Springer-Verlag.

Yu, P. (2012): "Likelihood Estimation and Inference in Threshold Regression," *Journal of Econometrics*, 167(1), 274–294.

Yu, P., and X. Fan (2021): "Threshold Regression With a Threshold Boundary," *Journal of Business and Economic Statistics*, 39(4), 953–971.

Yu, P., Q. Liao, and P. Phillips (2019): "Inferences and Specification Testing in Threshold Regression with Endogeneity," *Working Paper*.

Yu, P., and P. Phillips (2018): "Threshold Regression with Endogeneity," *Journal of Econometrics*, 203(1), 50–68.