# Partially Polynomial Estimation in Regression Discontinuity

Ping Yu*
University of Auckland

First Version: February 2010
This Draft: June 2010
(Very Preliminary, Comments Welcome)

**Abstract**

This paper proposes a new estimator of the treatment effect, called the partially polynomial estimator (PPE), in the regression discontinuity framework by extending the partially linear estimator (PLE) in Porter (2003). By treating regression discontinuity as threshold regression with a known threshold point, we interpret the PPE as a reparametrization of the local polynomial estimator (LPE) in the neighborhood of the discontinuity point. As a result, the PPE can achieve the optimal rate of convergence which the PLE can not attain under the broader conditions specified by Porter (2003). Furthermore, we show the PLE is indeed special in the sense that the form of its bias can not be extended to the general PPE case. A further advantage of the PPE is that the bandwidth can be easily selected by cross-validation since the discontinuity point is treated as an interior point instead of the boundary in the LPE.

KEYWORDS: Regression Discontinuity Design, Threshold Regression, Partially Polynomial Estimator, Partially Linear Estimator, Local Polynomial Estimator, Optimal Convergence Rate, Cross Validation
JEL-CLASSIFICATION: C13, C14, C21

*Email: p.yu@auckland.ac.nz. Special thanks go to Jack Porter for insightful discussions.

# 1 Introduction

The regression discontinuity (RD) design has got much popularity in applied econometric practice for identifying the treatment effects since Thistlewaite and Campbell (1960). An incomplete list of applications includes Angrist and Lavy (1999), Battistin and Rettore (2002), Black (1999), Card et al. (2006), Chay and Greenstone (2005), Chay et al. (2005), DesJardins and McCall (2008), DiNardo and Lee (2004), Jacob and Lefgren (2004), Lee (2008), Ludwig and Miller (2007), Pence (2002), and Van der Klaauw (2002). See Imbens and Lemieux (2008), Lee and Lemieux (2009), and Van der Klaauw (2008) for excellent surveys on this topic.

In estimating the causal effects of the treatment in the RD design, there are two key theoretical contributions among others in the literature. Hahn et al (2001) provide sufficient conditions for identification and use the local linear estimator (LLE) for estimating the treatment effect to overcome the boundary bias. Porter (2003) reveals that the optimal rate of convergence for estimation of the RD treatment effect is the same as that in the usual nonparametric conditional mean estimation problem by using a similar argument in Stone (1980). Porter (2003) provides two estimators to attain the optimal rate. The first estimator is based on Robinson's (1988) partially linear estimator (PLE). This estimator can achieve the optimal rate only under the more stringent assumption (Assumption 2(b) of Porter (2003)) on the data generating process (DGP). The second estimator is based on the local polynomial estimator (LPE) at the boundary which generalizes the LLE of Hahn et al (2001). This estimator can achieve the optimal rate under a broader assumption (Assumption 2(a) of Porter (2003)) on the DGP. There is a gap in logic: what is the relationship between the PLE and the LPE? Why the PLE can not achieve the optimal rate under the broader assumption? In this paper, we propose a new estimator called the partially polynomial estimator (PPE) which generalizes the PLE and builds a connection between the PLE and LPE. By including the differences in the derivatives (besides the size) of the conditional mean of the response on the two sides of the discontinuity point, the PPE is shown to be able to obtain the optimal rate. Actually, the PPE can be treated as an alternative estimator of the LPE, and is motivated by reparametrizing the threshold regression formulation of the RD design. As a result, the rate of its bias is same as the interior point in the local polynomial estimation.

Before presenting the main results on the PPE, we first define the basic structure of the RD design. Human behavior always evolves smoothly unless an abrupt change happens exogenously. This observation lies in the heart of RD design. Suppose a treatment $T$ is given based on a forcing (selection or assignment) variable $x$ by

$$T = \begin{cases} T_1, & \text{if } x \geq \pi, \\ T_0, & \text{if } x < \pi, \end{cases}$$

where $x$ is observed, the cut-off point $\pi$ is known, and both $T_0$ and $T_1$ follow the Bernoulli distribution while have different conditional means. Trochim (1984) divides the RD design into the sharp design and fuzzy design depending on $T$ is a deterministic function of $x$ or not. In the sharp design, the treatment assignment $T_1 = 1$ and $T_0 = 0$ almost surely. Let $Y_1$ and $Y_0$ be the potential outcomes corresponding to the two treatment assignments, then the observed outcome is $y = TY_1 + (1-T)Y_0$. Hahn et al (2001) shows that when $E[Y_0|x]$ and $E[Y_1|x]$ are continuous at $\pi$, the expected causal effect of the treatment on the outcome can be identified as

$$\alpha \equiv E[Y_1 - Y_0|\pi] = E[y|x = \pi+] - E[y|x = \pi-],$$

where $E[y|x = \pi+] = \lim_{x \downarrow \pi} E[y|x]$, and $E[y|x = \pi-] = \lim_{x \uparrow \pi} E[y|x]$. In the fuzzy design, $T_1$ and $T_0$ are random, but the propensity scores $E[T_1|x = \pi+] \neq E[T_0|x = \pi-]$. In this case, Hahn et al (2001) shows that $\alpha$ can

be identified under the local unconfoundedness condition. Specifically,

$$\alpha \equiv E\left[Y_1 - Y_0 | \pi\right] = \frac{E\left[y|x = \pi+\right] - E\left[y|x = \pi-\right]}{E\left[T|x = \pi+\right] - E\left[T|x = \pi-\right]}.$$

In both cases, $\alpha$ only involves the difference of two conditional means. We will concentrate on the sharp design, since the estimation scheme has no essential change in the fuzzy design.

In section 2, we construct the PPE, derive its asymptotic distribution, and discuss the relationship with the LPE. In section 3, we discuss some practical issues and extensions of the PPE, and Section 5 concludes. The proof of the main theorem of this paper and related lemmas are given in Appendices A and B, respectively.

A word on notation: $\otimes$ is the Kronecker product. Any object with a subbar generally denotes its "demeaned" counterpart as explained in the main text. $1(A)$ is the indicator function with value 1 when the event $A$ is true and 0 when it is false. $\approx$ means the higher-order terms are omitted or a constant term is omitted (depending on the context). Because the LPE at an interior point is used throughout this paper, we define some notations used in its construction before closing this section. Suppose we observed a dataset $\{\{\omega_i\}_{i=1}^n, \{x_i\}_{i=1}^n\} \equiv \{\boldsymbol{\omega}, \mathbf{x}\}$, and we want to estimate the conditional mean $m(x) \equiv E\left[\omega_i | x_i = x\right]$. From Fan and Gijbels (1996), the $p$th order LPE is a linear functional of $\boldsymbol{\omega}$:

$$\begin{aligned}
\widehat{m}(x) &= P_x^n(\boldsymbol{\omega}) = e_1' \left(X(x)' K(x) X(x)\right)^{-1} X(x)' K(x) \boldsymbol{\omega}, \\
&= e_1' \left(H^{-1} X(x)' K_h(x) X(x) H^{-1}\right)^{-1} H^{-1} X(x)' K_h(x) \boldsymbol{\omega}, \\
&\equiv e_1' \left(Z(x)' K_h(x) Z(x)\right)^{-1} Z(x)' K_h(x) \boldsymbol{\omega}, \\
&= e_1' \left(\frac{1}{n}\sum_{j=1}^n Z_j(x) Z_j'(x) k_h(x_j - x)\right)^{-1} \frac{1}{n}\sum_{j=1}^n Z_j(x) k_h(x_j - x)\omega_j, \\
&\equiv e_1' S_n^{-1}(x) \widetilde{r}(\omega(x)),
\end{aligned}$$

where

$$\begin{aligned}
X(x) &= \begin{pmatrix} 1 & x_1 - x & \cdots & (x_1 - x)^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n - x & \cdots & (x_n - x)^p \end{pmatrix}_{n\times(p+1)} \equiv \begin{pmatrix} X_1(x)' \\ \vdots \\ X_n(x)' \end{pmatrix} \equiv \left(X^0(x), \cdots, X^p(x)\right), \\
K(x) &= diag\left\{k\left(\frac{x_1 - x}{h}\right), \cdots, k\left(\frac{x_n - x}{h}\right)\right\}_{n\times n}, K_h(x) = diag\left\{k_h(x_1 - x), \cdots, k_h(x_n - x)\right\}_{n\times n}, \\
e_1 &= (1, 0, \cdots, 0)'_{(p+1)\times 1}, \; H = diag\left\{1, h, \cdots, h^p\right\}_{(p+1)\times(p+1)}, Z(x) = X(x) H^{-1}, \\
Z_j(x) &= \left(1, \frac{x_j - x}{h}, \cdots, \left(\frac{x_j - x}{h}\right)^p\right)'_{(p+1)\times 1} \equiv \left(1, Z_j^1(x), \cdots, Z_j^p(x)\right)',
\end{aligned}$$

with $k(\cdot)$ being a kernel function with a compact support $[-M, M]$, $k_h(\cdot) = \frac{1}{h}k\left(\frac{\cdot}{h}\right)$, and $h$ being the bandwidth. The dimensions of $e_1$ and $H$ are determined by the context without further explanations. Denote $e_1' \left(X(x)' K(x) X(x)\right)^{-1} X(x)' K(x)$ as $W^n(x) = (W_1^n(x), \cdots, W_n^n(x))$, then

$$P_x^n(\omega) = \sum_{j=1}^n W_j^n(x)\omega_j.$$

As shown in Lemma 2.1 of Fan et al (1997), $P_x^n$ is equivalent to a linear functional $P_x$ on $\mathbb{R}^n$ asymptotically:

$$P_x(\boldsymbol{\omega}) = \frac{1}{nhf(x)} \sum_{j=1}^{n} K_p^* \left( \frac{x_j - x}{h} \right) \omega_j,$$

where $f(x)$ is the density of $x_i$,

$$K_p^*(u) = e_1' \Gamma^{-1} (1, u, \cdots, u^p)' k(u) \equiv e_1' \Gamma^{-1} \delta(u),$$

is a kernel of order $p+1$ when $p$ is odd and of order $p+2$ when $p$ is even as defined by Gasser et al (1985),

$$\delta(u) = (k(u), uk(u), \cdots, u^p k(u))',$$

and $\Gamma = \left(\gamma_{i+j-2}\right)_{1 \le i,j \le p+1}$ is invertible with $\gamma_j = \int u^j k(u) du$.[1] When the arguments of $P_x^n(\cdot)$ and $P_x(\cdot)$ are matrices, we treat them as operating on each column of the matrices to get a row vector.

# 2 Partially Polynomial Estimation

This section presents the main results of this paper. It begins with the construction of the PPE, followed by the discussion of its connection with the LPE, and concludes with the asymptotic theory of the PPE.

## 2.1 Construction of the Estimator

Let us first review the motivation of the PLE in Porter (2003). Recall that the response is related to the one-dimensional covariate by the following form:

$$y = m(x) + \alpha d + \varepsilon, \text{ where } E[\varepsilon | x, d] = 0,\ d = 1(x \ge \pi), \tag{1}$$

where $m(x) \equiv E[y|x] - \alpha d$, so $\alpha$ can be treated as the parametric coefficient in the partially linear model. The PLE is defined as

$$\arg\min_{\alpha} \sum_{i=1}^{n} \left[ y_i - \alpha d_i - \sum_{j=1}^{n} w_j^i (y_j - \alpha d_j) \right]^2,$$

where $w_j^i = \frac{k_h(x_i - x_j)}{\sum_{l=1}^{n} k_h(x_i - x_l)}$. $\sum_{j=1}^{n} w_j^i (y_j - \alpha d_j)$ can be treated as an estimator of $m(x)$ at $x_i$. Actually, the PLE in Robinson (1988) can be equivalently redefined in the way above. $\alpha$ can be identified as $E[y|x = \pi+] - E[y|x = \pi-]$ because $m(x)$ is assumed to be continuous in the neighborhood of $\pi$. Note that $d_i - \sum_{j=1}^{n} w_j^i d_j = 0$ when $x_i$ is out of a $O(h)$ neighborhood of $\pi$, so only the information in the neighborhood of $\pi$ is used to estimate $\alpha$. As a result, the PLE only has a nonparametric convergence rate; see Section 3.3 of Porter (2003) for more discussions on this nonparametric rate.

Because the PLE only explores the information that $E[y|x]$ rather than its derivatives has a jump at $\pi$, it can not achieve the efficient rate when $m(\cdot)$ is known to be only continuous at $\pi$. In this paper, we generalize the PLE to the PPE by explicitly considering the jumps of the derivatives of $E[y|x = x]$ at $\pi$.

[1] Or equivalently, as shown in Ruppert and Wand (1994), $K_p^*(u) = |\Gamma(u)| / |\Gamma| k(u)$, where $\Gamma(u)$ is the same as $\Gamma$, but with the first column replaced by $(1, u, \cdots, u^p)$.

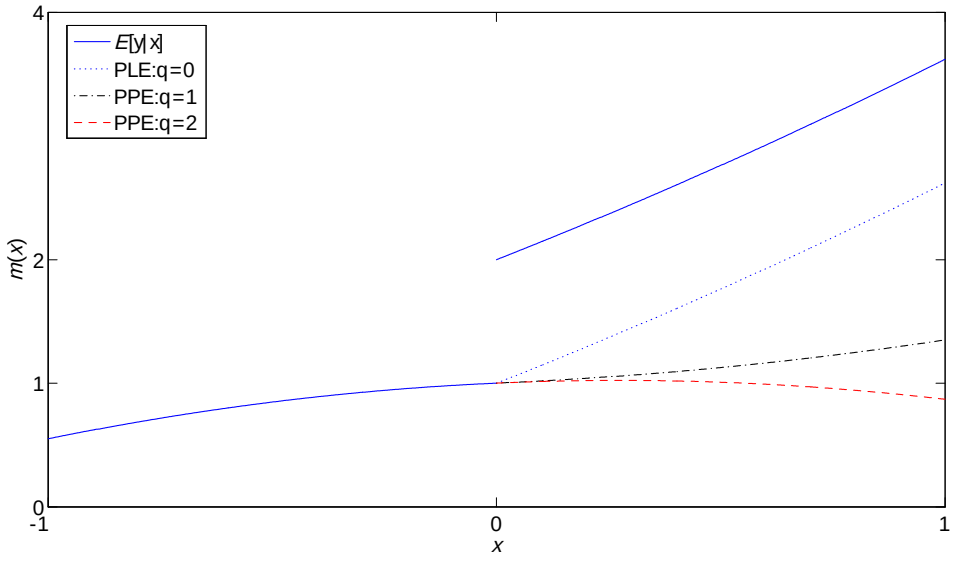


Figure 1: $m(x)$ in Partially Polynomial Estimation with Different Orders

Specifically, let

$$\begin{aligned} y &= m(x) + \left(\alpha + \beta_1 (x-\pi) + \cdots + \beta_q (x-\pi)^q\right) d + \varepsilon \\ &\equiv \widetilde{y} + \left(\alpha + \beta_1 (x-\pi) + \cdots + \beta_q (x-\pi)^q\right) d, \end{aligned} \tag{2}$$

where

$$m(x) \equiv E[y|x] - \left(\alpha + \beta_1 (x-\pi) + \cdots + \beta_p (x-\pi)^q\right) d$$

has continuous derivatives at $\pi$ to $q$th order, and $\beta_\nu = \left[\left.\frac{\partial^\nu E[y|x]}{\partial x^\nu}\right|_{x=\pi+} - \left.\frac{\partial^\nu E[y|x]}{\partial x^\nu}\right|_{x=\pi-}\right]/\nu!$, $\nu = 1, \cdots, q$, is the scaled difference of the $\nu$th derivative of $E[y|x]$ in the left and right neighborhoods of $\pi$.[2] $m(x)$ with $E[y|x] = \begin{cases} 1 + 0.16x - 0.29x^2, & \text{if } x < 0; \\ 2 + 1.43x + 0.19x^2, & \text{if } x \geq 0; \end{cases}$ is shown in Figure 1. In this special case, $\alpha = 1$, $\beta_1 = 1.27$ and $\beta_2 = 0.48$. Note that $q = 0$ corresponds to the PLE. Obviously, its $m(x)$ is not smooth at 0. As in the partially linear estimation, $\alpha$ is estimated by

$$\min_{\alpha, \beta_1, \cdots, \beta_p} \frac{1}{n} \sum_{i=1}^{n} \left[\overrightarrow{y}_i - \widehat{m}\left(x_i | \overrightarrow{\mathbf{y}}\right)\right]^2, \tag{3}$$

where

$$\overrightarrow{y}_i = y_i - \left(\alpha + \beta_1 (x_i - \pi) + \cdots + \beta_p (x_i - \pi)^q\right) d_i, \ \overrightarrow{\mathbf{y}} = \left(\overrightarrow{y}_1, \cdots, \overrightarrow{y}_n\right)',$$

and $\widehat{m}\left(x_i | \overrightarrow{\mathbf{y}}\right)$ is a nonparametric estimator of $m(x_i)$. A popular choice of $\widehat{m}\left(x | \overrightarrow{\mathbf{y}}\right)$ is the LPE, where

[2] (2) is a partially linear regression in Robinson (1988) because the parametric component of (2) is linear in the parameters. The term PPE is to distinguish (2) from the partially linear regression in Porter (2003).

$\widehat{m}\left(x|\overrightarrow{\mathbf{y}}\right)$ is determined by the minimizer $\widehat{a}$ in the following minimization problem:

$$\min_{a,b_1,\cdots,b_p} \frac{1}{n}\sum_{j=1}^{n} k_h\left(x_j - x\right)\left[\overrightarrow{y}_j - a - b_1\left(x_j - x\right) - \cdots - b_p\left(x_j - x\right)^p\right]^2. \tag{4}$$

To explore the $q$th order smoothness of $m(\cdot)$, we assume $p \geq q$, but $p$ and $q$ are not necessarily the same. Note that both $\overrightarrow{y}_i$ and $\widehat{m}\left(x_i|\overrightarrow{\mathbf{y}}\right)$ depend on $\theta \equiv \left(\alpha, \beta'\right)'$. From Lemma 2.1 of Fan et al. (1997), $\widehat{m}\left(x|\overrightarrow{\mathbf{y}}\right)$ is equivalent to the local constant estimator with a higher-order kernel. Because the kernel function in Porter (2003) is allowed to be higher order, the PPE distinguishes from the PLE mainly by considering the difference of derivatives at $\pi$ in (3) rather than using the LPE in estimating $m(x)$.

Some calculus shows that

$$\begin{pmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{pmatrix} = \left(\underline{X}^{d\prime}\underline{X}^{d}\right)^{-1}\underline{X}^{d\prime}\underline{\mathbf{y}} \text{ and } \widehat{\alpha} = e_1'\left(\underline{X}^{d\prime}\underline{X}^{d}\right)^{-1}\underline{X}^{d\prime}\underline{\mathbf{y}}, \tag{5}$$

where

$$\begin{aligned}
\underline{X}^{d} &= \begin{pmatrix} X_1^d(\pi) - P_{x_1}^n\left(X^d(\pi)\right) \\ \vdots \\ X_n^d(\pi) - P_{x_n}^n\left(X^d(\pi)\right) \end{pmatrix} \equiv \begin{pmatrix} \underline{X}_1^d(\pi)' \\ \vdots \\ \underline{X}_n^d(\pi)' \end{pmatrix}_{n\times(q+1)} \equiv \left(\underline{X}^{0d}(\pi), \cdots, \underline{X}^{qd}(\pi)\right)_{n\times(q+1)}, \\
&= X^d(\pi) - \mathbf{e}_1'\left(X'KX\right)^{-1}X'K\mathbf{I}X^d(\pi) = \left(I_n - \mathbf{e}_1'\left(X'KX\right)^{-1}X'K\mathbf{I}\right)X^d(\pi),
\end{aligned}$$

with

$$\begin{aligned}
X &= diag\left\{X(x_1), \cdots, X(x_n)\right\}_{n^2\times n(p+1)}, \\
X^d(x) &= \begin{pmatrix} 1(x_1 \geq x) & (x_1 - x)1(x_1 \geq x) & \cdots & (x_1 - x)^q 1(x_1 \geq x) \\ \vdots & \vdots & \vdots & \vdots \\ 1(x_n \geq x) & (x_n - x)1(x_n \geq x) & \cdots & (x_n - x)^q 1(x_n \geq x) \end{pmatrix} \\
&\equiv \begin{pmatrix} X_1^d(x)' \\ \vdots \\ X_n^d(x)' \end{pmatrix}_{n\times(q+1)} \equiv \left(X^{0d}(x), \cdots, X^{qd}(x)\right)_{n\times(q+1)},
\end{aligned}$$

$$\begin{aligned}
I_n &= diag\left\{1, \cdots, 1\right\}_{n\times n}, \mathbf{e}_1 = diag\left\{e_1, \cdots, e_1\right\}_{n(p+1)\times n} = I_n \otimes e_1, \\
e &= (1, 1, \cdots, 1)'_{n\times 1}, \mathbf{I} = (e \otimes I_n)_{n^2\times n},
\end{aligned}$$

$$K = diag\left\{K_h(x_1), \cdots, K_h(x_n)\right\}_{n^2\times n^2},$$

and

$$\underline{\mathbf{y}} = \begin{pmatrix} y_1 - P_{x_1}^n(\mathbf{y}) \\ \vdots \\ y_n - P_{x_n}^n(\mathbf{y}) \end{pmatrix} = \left(I_n - \mathbf{e}_1'\left(X'KX\right)^{-1}X'K\mathbf{I}\right)\mathbf{y},$$

with

$$\begin{aligned} \mathbf{y} &= m(\mathbf{x}) + X^d(\pi)\theta + \boldsymbol{\varepsilon} \equiv \widetilde{\mathbf{y}} + X^d(\pi)\theta, \\ m(\mathbf{x}) &= (m(x_1), \cdots, m(x_n))', \boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)', \\ \widetilde{\mathbf{y}} &= m(\mathbf{x}) + \boldsymbol{\varepsilon} = (\widetilde{y}_1, \cdots, \widetilde{y}_n). \end{aligned}$$

Some explanations on $\widehat{\theta}$ are in order. $\underline{X}^d$ and $\underline{\mathbf{y}}$ are the demeaned $X^d(\pi)$ and $\mathbf{y}$ by the "local polynomial operator" $P_x^n$. $I_n - \mathbf{e}_1'(X'KX)^{-1}X'K\mathbf{I} \equiv I_n - P_{\mathbf{x}}^n$ is like a demeaned operator on a vector in $\mathbb{R}^n$ at $\mathbf{x}$. Note that

$$\begin{aligned} \left(\underline{X}^{d\prime}\underline{X}^d\right)^{-1}\underline{X}^{d\prime}\underline{\mathbf{y}} &= \left(\underline{X}^{d\prime}\underline{X}^d\right)^{-1}\underline{X}^{d\prime}\left(\underline{X}^d\theta + \widetilde{\mathbf{y}} - P_{\mathbf{x}}^n(\widetilde{\mathbf{y}})\right) \\ &= \theta + H^{-1}\left(\frac{1}{nh}H^{-1}\underline{X}^{d\prime}\underline{X}^dH^{-1}\right)^{-1}\frac{1}{nh}H^{-1}\underline{X}^{d\prime}\underline{\widetilde{\mathbf{y}}} \\ &\equiv \theta + H^{-1}\left(\frac{1}{nh}\underline{Z}^{d\prime}\underline{Z}^d\right)^{-1}\frac{1}{nh}\underline{Z}^{d\prime}(m(\mathbf{x}) - \overline{m}(\mathbf{x}) + \boldsymbol{\varepsilon} - \overline{\boldsymbol{\varepsilon}}) \\ &= \theta + H^{-1}\left(\frac{1}{nh}\sum_{l=1}^n \underline{Z}_l^d(\pi)\underline{Z}_l^{d'}(\pi)\right)^{-1}\left(\frac{1}{nh}\sum_{l=1}^n \underline{Z}_l^d(\pi)((m(x_l) - \overline{m}(x_l) + \varepsilon_l - \overline{\varepsilon}_l))\right), \end{aligned} \tag{6}$$

where $\underline{Z}^d = \underline{X}^dH^{-1}$ is the normalized $\underline{X}^d$ like $Z(x)$ in $P_x^n$, $\underline{Z}_l^d(\pi) = H^{-1}\underline{X}_l^d(\pi)$, $l = 1, \cdots, n$, and $\underline{\widetilde{\mathbf{y}}} = \underline{\widetilde{\mathbf{y}}} - \underline{\widetilde{m}}(\mathbf{x})$ with

$$\begin{aligned} \widetilde{m}(\mathbf{x}) &= (\widetilde{m}(x_1), \cdots, \widetilde{m}(x_n))' = P_{\mathbf{x}}^n(\widetilde{\mathbf{y}}) \\ &= \left(P_{x_1}^n(m(\mathbf{x})), \cdots, P_{x_n}^n(m(\mathbf{x}))\right)' + \left(P_{x_1}^n(\boldsymbol{\varepsilon}), \cdots, P_{x_n}^n(\boldsymbol{\varepsilon})\right)' \\ &\equiv (\overline{m}(x_1), \cdots, \overline{m}(x_n))' + (\overline{\varepsilon}_1, \cdots, \overline{\varepsilon}_n) \\ &\equiv \overline{m}(\mathbf{x}) + \overline{\boldsymbol{\varepsilon}}. \end{aligned}$$

From Lemma 1 in Appendix B, $\underline{X}_l^d(\pi) = 0$ for $|x_l - \pi| > Mh$, $l = 1, \cdots, n$, so only the $x_l$'s in the $Mh$ neighborhood of $\pi$ will contribute to $\widehat{\alpha}$. In consequence, the convergence rate of $\widehat{\alpha}$ is $\sqrt{nh}$ instead of $\sqrt{n}$. In the proof of the appendices, we will see $\underline{Z}^{d\prime}(m(\mathbf{x}) - \overline{m}(\mathbf{x}))$ will contribute to the bias, and $\underline{Z}^{d\prime}(\boldsymbol{\varepsilon} - \overline{\boldsymbol{\varepsilon}})$ will contribute to the variance. Interestingly, $m(\mathbf{x}) - \widetilde{m}(\mathbf{x})$ will also contribute to the variance since $\overline{\boldsymbol{\varepsilon}}$ comes from $\widetilde{m}(\mathbf{x})$. This is different from the usual LPE at an interior point where only $\boldsymbol{\varepsilon}$ contributes to the variance; see Ruppert and Wand (1994) for the details.

## 2.2 Connection with the Local Polynomial Estimator

To further understand the PPE, let us compare it with the least squares estimator (LSE) in threshold regression; see Chan (1993), Hansen (2000) and Yu (2007) for more discussions about threshold regression. A typical setup of the PPE is $p = q$, and we only concentrate on this case. In threshold regression,

$$y = \begin{cases} x'\beta_1 + \varepsilon_1, & z < \pi; \\ x'\beta_2 + \varepsilon_2, & z \geq \pi; \end{cases} \tag{7}$$

where $z$ is the threshold variable used to split the sample, $x \in \mathbb{R}^{p+1}$ with the first element a constant, $\beta \equiv (\beta_1', \beta_2')' \in \mathbb{R}^{2(p+1)}$ and $\sigma \equiv (\sigma_1, \sigma_2)'$ are threshold parameters on mean and variance in the two regimes of (7), the error terms $\varepsilon_1$ and $\varepsilon_2$ adopt conditional heteroskedasticity and are not necessarily the

same, and all the other variables have the same definitions as in the linear regression framework. A useful reparametrization of (7) is

$$y = x'\beta_1 + x'\left(\beta_2 - \beta_1\right) 1\left(z \geq \pi\right) + \varepsilon, \tag{8}$$

where $\varepsilon = \varepsilon_1$ when $z < \pi$ and $\varepsilon = \varepsilon_2$ when $z \geq \pi$. Return to the regression discontinuity model, then (7) is only satisfied locally. (2) can be approximated in two equivalent ways:

$$y = \begin{cases} \beta_{10} + \beta_{11}\left(x - \pi\right) + \cdots + \beta_{1p}\left(x - \pi\right)^p + \varepsilon_1, & x < \pi; \\ \beta_{20} + \beta_{21}\left(x - \pi\right) + \cdots + \beta_{2p}\left(x - \pi\right)^p + \varepsilon_2, & x \geq \pi; \end{cases} \tag{9}$$

and

$$y = \beta_{10} + \beta_{11}\left(x - \pi\right) + \cdots + \beta_{1p}\left(x - \pi\right)^p + \left(\alpha + \beta_1\left(x - \pi\right) + \cdots + \beta_p\left(x - \pi\right)^p\right) 1\left(x \geq \pi\right) + \varepsilon, \tag{10}$$

where $\beta_{10} + \beta_{11}\left(x - \pi\right) + \cdots + \beta_{1p}\left(x - \pi\right)^p$ is the Taylor expansion of $m(x)$ to order $p$ in the left neighborhood of $\pi$, $\beta_{10} = m(\pi-)$,

$$\left(\beta_{20}, \beta_{21}, \cdots, \beta_{2p}\right) = \left(\beta_{10}, \beta_{11}, \cdots, \beta_{1p}\right) + \left(\alpha, \beta_1, \cdots, \beta_p\right),$$

and the threshold variable $z$ in (7) is just $x$. Obviously, $\left(\alpha, \beta_1, \cdots, \beta_p\right)$ plays the role of $\beta_2 - \beta_1$ in (8).

The main concern in threshold regression is the threshold point $\pi$. In contrast, in regression discontinuity, $\pi$ is generally known from the design, and the main concern is the mean difference $\alpha$ between the two regimes. In threshold regression, we can set up the objective functions of the least squares estimation for the two equivalent models (7) and (8) as follows:

$$\begin{aligned} \text{Obj1} &= \sum_{i=1}^{n} \left(y_i - x_i'\beta_1 1(z_i < \pi) - x_i'\beta_2 1(z_i \geq \pi)\right)^2, \\ \text{Obj2} &= \sum_{i=1}^{n} \left(y_i - x_i'\left(\beta_2 - \beta_1\right) 1\left(z_i \geq \pi\right) - x_i'\beta_1\right)^2. \end{aligned}$$

Suppose $\pi$ is known, then in Obj1, $\beta_2 - \beta_1$ is estimated in two steps. First, estimate $\beta_2$ using the data with $z_i \geq \pi$, and estimate $\beta_1$ using the data with $z_i < \pi$. Second, take difference of the estimates of $\beta_2$ and $\beta_1$ in step 1 as the estimator of $\beta_2 - \beta_1$. In contrast, Obj2 uses a profiled procedure. First fix $\beta_2 - \beta_1$ and regress $y_i - x_i'\left(\beta_2 - \beta_1\right) 1\left(z_i > \pi\right)$ on $x_i$ to get an estimate of $\beta_1$, then minimize Obj2 with respect to $\beta_2 - \beta_1$ to estimate $\beta_2 - \beta_1$. The estimators based on these two objective functions correspond to the LPE and PPE in regression discontinuity, respectively. The only difference is that we run the regressions using the local data around $\pi$. Suppose the uniform kernel is used and the bandwidth is $h$. In (9), we run the regression in the right $h$ neighborhood of $\pi$ to estimate $\beta_2$, and in the left neighborhood of $\pi$ to estimate $\beta_1$, and $\theta$ is then estimated by the difference between these two estimators. This is just the LPE. In contrast, the profiled procedure is used in (10) to construct the PPE. Now, $\beta_{10} + \beta_{11}\left(x_i - \pi\right) + \cdots + \beta_{1q}\left(x_i - \pi\right)^p$ plays the role of $x_i'\beta_1$ which corresponds to $m(x_i)$ in (2). A better approximation of $m(x_i)$ is using the Taylor expansion at $x_i$ instead of $\pi$. This is just what is done in the PPE. In threshold regression, because the conditional mean of $y$ in the regime $z < \pi$ is linear in $x$, the Taylor expansion around any point in the support of $x_i$ is the same, but in regression discontinuity, different expansions indeed introduce some differences in their asymptotic properties. This is understandable since the PPE is a nonparametric estimator which is designed to study the local properties. The nonparametric nature of the PPE is evident by checking (10). The differences in size and derivatives of $E[y|x]$ in the right regime from the left regime are not the same at any $x \geq \pi$. The approximation in (10) is valid only if $x$ is close to $\pi$. A key advantage of the PPE is that $m(x)$, $x < \pi$, is

estimated using the data in both neighborhoods of $\pi$, but in the LPE, $m^-(\pi)$ is estimated using only the data in the left neighborhood of $\pi$.

## 2.3 Asymptotic Theory of $\widehat{\alpha}$

First, we give out some regularity conditions required in deriving the asymptotic distribution of $\widehat{\alpha}$. These assumptions roughly corresponds to those in Section 3.1 of Porter (2003).

**Assumption K:** $K(\cdot)$ is a symmetric, bounded, Lipschitz function, zero outside a bounded set $[-M, M]$, and $\int K(u)du = 1$.

**Assumption F:** For some compact interval $N$ of $\pi$ with $\pi \in int(N)$, $f$ is $l_f$ times continuously differentiable and bounded away from zero.

**Assumption M:**

**(a)** $m(x)$ is $l_m$ times continuously differentiable for $x \in N \backslash \{\pi\}$, and $m(x)$ is continuous at $\pi$ with finite right and left-hand derivatives to order $l_m$.

**(b)** Right and left-hand derivatives of $m(x)$ to order $l_m$ are equal at $\pi$.

The typical case where Assumption M(b) holds is the common treatment effects model. In such a model, $Y_{1i} - Y_{0i} = \alpha$ is constant across individuals, and $m(\cdot)$ is smooth even in the PLE and we need not consider the derivative differences.

**Assumption E:**

**(a)** $\sigma^2(x) = E\left[\varepsilon^2 | x\right]$ is continuous for $x \neq \pi$, $x \in N$, and the right and left-hand limits at $\pi$ exist.

**(b)** For some $\zeta > 0$, $E\left[|\varepsilon|^{2+\zeta} | x\right]$ is uniformly bounded on $N$.

**Assumption B:** $\frac{n^{\zeta/(2+\zeta)}h}{\ln n} \to \infty$, $\frac{\sqrt{nh}}{\ln n} \to \infty$,

**(a)** $\sqrt{nh}h^{q+3} \to 0$, $\sqrt{nh}h^{q+1} \to C_a$, where $0 \leq C_a < \infty$.

**(b1)** $\sqrt{nh}h^{p+3} \to 0$, $\sqrt{nh}h^{p+1} \to C_{b1}$, where $0 \leq C_{b1} < \infty$.

**(b2)** $\sqrt{nh}h^{p+3} \to 0$, $\sqrt{nh}h^{p+2} \to C_{b2}$, where $0 \leq C_{b2} < \infty$.

The following theorem 1 provides the asymptotic results of the PPE under different set of regularity conditions.

**Theorem 1** *Suppose $p \geq q$, $q \geq 1$, and Assumptions E and K hold,*

**(a)** *If Assumption F holds with $l_f \geq 0$, Assumption M(a) holds with $l_m \geq q+1$, and Assumption B(a) holds, then*

$$\sqrt{nh}\left(\widehat{\alpha}-\alpha\right) \xrightarrow{d} N\left(-C_a B_a, \frac{V}{f(\pi)}\right),$$

*Here,*

$$\begin{aligned} B_a &= e_1' N_p^{-1}\left[\frac{m^{(q+1)}(\pi+)}{(q+1)!}Q_{pq}^{+}+\frac{m^{(q+1)}(\pi-)}{(q+1)!}Q_{pq}^{-}\right], \\ V &= e_1' N_p^{-1}\left[\sigma^2(\pi+)\Omega_p^{+}+\sigma^2(\pi-)\Omega_p^{-}\right] N_p^{-1} e_1 \end{aligned}$$

*with*

$$N_p(i,j)=\int_0^M K_p^*(\overline{\delta}_{i-1}(w+))K_p^*(\overline{\delta}_{j-1}(w+))dw+\int_{-M}^0 K_p^*(\overline{\delta}_{i-1}(w-))K_p^*(\overline{\delta}_{j-1}(w-))dw,$$

$$\begin{aligned} Q_{pq}^{+}(i) &= \int_0^M K_p^*(\overline{\delta}_{i-1}(w+))\left(\int_{-w}^M K_p^*(u)\left(w+u\right)^{q+1}du - w^{q+1}\right)dw \\ &\quad + \int_{-M}^0 K_p^*(\overline{\delta}_{i-1}(w-))\left(\int_{-w}^M K_p^*(u)\left(w+u\right)^{q+1}du\right)dw, \\ Q_{pq}^{-}(i) &= \int_0^M K_p^*(\overline{\delta}_{i-1}(w+))\left(\int_{-M}^{-w} K_p^*(u)\left(w+u\right)^{q+1}du\right)dw \\ &\quad + \int_{-M}^0 K_p^*(\overline{\delta}_{i-1}(w-))\left(\int_{-M}^{-w} K_p^*(u)\left(w+u\right)^{q+1}du - w^{q+1}\right)dw, \end{aligned}$$

$$\begin{aligned} \Omega_p^{+}(i,j) &= \int_0^M \left[K_p^*(\overline{\delta}_{i-1}(w+)) - \left(\int_0^M K_p^*(\overline{\delta}_{i-1}(v+))K_p^*(w-v)\,dv + \int_{-M}^0 K_p^*(\overline{\delta}_{i-1}(v-))K_p^*(w-v)\,dv\right)\right] \\ &\quad \left[K_p^*(\overline{\delta}_{j-1}(w+)) - \left(\int_0^M K_p^*(\overline{\delta}_{j-1}(v+))K_p^*(w-v)\,dv + \int_{-M}^0 K_p^*(\overline{\delta}_{j-1}(v-))K_p^*(w-v)\,dv\right)\right]dw, \\ \Omega_p^{-}(i,j) &= \int_{-M}^0 \left[K_p^*(\overline{\delta}_{i-1}(w-)) - \left(\int_0^M K_p^*(\overline{\delta}_{i-1}(v+))K_p^*(w-v)\,dv + \int_{-M}^0 K_p^*(\overline{\delta}_{i-1}(v-))K_p^*(w-v)\,dv\right)\right] \\ &\quad \left[K_p^*(\overline{\delta}_{j-1}(w-)) - \left(\int_0^M K_p^*(\overline{\delta}_{j-1}(v+))K_p^*(w-v)\,dv + \int_{-M}^0 K_p^*(\overline{\delta}_{j-1}(v-))K_p^*(w-v)\,dv\right)\right]dw, \end{aligned}$$

*and*

$$\begin{aligned} K_p^*(\overline{\delta}_{i-1}(w+)) &= w^{i-1} - \int_{-w}^M K_p^*(u)(w+u)^{i-1}du, \\ K_p^*(\overline{\delta}_{i-1}(w-)) &= -\int_{-w}^M K_p^*(u)(w+u)^{i-1}du, \end{aligned}$$

$i,j = 1,\cdots,q+1$.

**(b1)** *If Assumption F holds with $l_f \geq 0$, Assumption M(b) holds $l_m \geq p+1$, and Assumption B(b1) holds,*

*when $p$ is odd, then*

$$\sqrt{nh}\left(\widehat{\alpha}-\alpha\right)\xrightarrow{d}N\left(-C_{b1}B_{b1},\frac{V}{f(\pi)}\right),$$

*where*

$$B_{b1}=\left(\int_{-M}^{M}K_p^*\left(u\right)u^{p+1}du\right)\frac{m^{(p+1)}(\pi)}{(p+1)!}e_1'N_p^{-1}Q_p$$

*with*

$$Q_p(i)=\int_0^M K_p^*(\overline{\delta}_{i-1}(w+))dw+\int_{-M}^0 K_p^*(\overline{\delta}_{i-1}(w-))dw,$$

$i=1,\cdots,q+1$.

**(b2)** *If Assumption F holds with $l_f\geq 1$, Assumption M(b) holds with $l_m\geq p+2$, and Assumption B(b2) holds, when $p$ is even, then*

$$\sqrt{nh}\left(\widehat{\alpha}-\alpha\right)\xrightarrow{d}N\left(-C_{b2}B_{b2},\frac{V}{f(\pi)}\right),$$

*where*

$$B_{b2}=\left(\int_{-M}^{M}K_p^*\left(u\right)u^{p+2}du\right)\left(\frac{m^{(p+2)}(\pi)f'(\pi)}{(p+1)!f(\pi)}+\frac{m^{(p+2)}(\pi)}{(p+2)!}\right)e_1'N_p^{-1}Q_p.$$

Theorem 1 is surprising in two aspects. First, under Assumption M, the PPE can achieve the efficient rate by including the differences between derivatives in the left and right neighborhoods of $\pi$. For example, if $m(x)$ is in $\overline{C}_r$, $r>1$, of Porter (2003), where $\overline{C}_r$ is the set of functions satisfying Assumption M(a) with $l_m=r$, then the PPE with $p\geq q=r-1$ can achieve the optimal convergence rate. If $m(x)$ is in $C_r$, $r>2$, of Porter (2003), where $C_r$ is the set of functions satisfying Assumption M(b) with $l_m=r$, then the PPE with $0<q\leq p=r-1$ ($r-2$ when $r$ is even) can achieve the optimal convergence rate.[3] Second, the PLE is indeed very special. In our notation, when $q=0$, $Q_{pq}^+=Q_{pq}^-=0$, so the bias in (a) is $O\left(\sqrt{nh}h^2\right)$ instead of $O\left(\sqrt{nh}h\right)$ as illustrated in Theorem 2(a) of Porter (2003). In (b1) and (b2), $Q_p=0$, so a higher-order bias $O\left(\sqrt{nh}h^{p+2+1(p\text{ is even})}\right)$ appears as shown in Theorem 2(b) of Porter (2003). This is basically because $1\left(x_i\geq\pi\right)$ and $1\left(x_i<\pi\right)$ are symmetric, and the lower-order biases in the left and right neighborhoods of $\pi$ are canceled. In the PPE, $(x_i-\pi)^k1\left(x_i\geq\pi\right)$, $k\geq 1$, and $1\left(x_i<\pi\right)$ are not symmetric, so the lower-order bias remains. The order of the biases in the PLE, LPE and PPE is summarized in the following Table 1. Note that when the kernel is symmetric, $s$ in the partially linear estimation of Porter (2003) must be even. Roughly speaking, $s$ plays the similar role as $p+1$ when $p$ is odd and $p+2$ when $p$ is even in the partially polynomial estimation. In the LPE, when $p$ is odd and Assumption M(b) holds, the lower-order biases in the two neighborhoods of $\pi$ offset each other, and a higher-order bias appears.

| | Assumption M(a) | Assumption M(b) |
|---|---|---|
| Partially Linear ($q=0,p\geq q$) | 2 | $p+2+\mathbf{1}\left(p\text{ is even}\right)$ |
| Partially Polynomial ($q>0,p\geq q$) | $q+1$ | $p+1+\mathbf{1}\left(p\text{ is even}\right)$ |
| Local Polynomial ($p\geq 0$) | $p+1$ | $p+1+\mathbf{1}\left(p\text{ is odd}\right)$ |

[3] If $q=0$, then it is hard to achieve the optimal rate exactly when $r$ is even. If $p=r-2$, then the bias order is $r+1$; if $p=r-3$, then the bias order is $r-1$. This phenomenon also appears in the LPE at the interior point and the PPE when $r$ is odd. This is why Stone (1980) uses the nearest neighborhood estimator instead of the LPE to achieve the optimal rate. This is also why in the usual local polynomial literature, $r$ is assumed to be even.

Table 1: Biases in Three Estimation Methods (the $b$ in $\sqrt{nh}h^b$)

As discussed above, the PLE with a higher-order kernel is essentially equivalent to the PPE with $q = 0$ and some $p > q$. But there is indeed some subtle difference between them. Namely, Theorem 1 needs less stringent conditions on the smoothness of $f(x)$ than those in Theorem 2 of Porter (2003). For example, in (a), Porter (2003) requires $l_f \geq 2$ while Theorem 1 only requires $l_f \geq 0$; in (b1) and (b2), Porter (2003) requires $l_f \geq s$, while Theorem 1 only requires $l_f \geq 1$. This is the role played by the PPE more than the higher-order kernel estimator; that is, the PPE adapts automatically to the smoothness of the density of $x$.

Note that the first parts of $B_{b1}$ and $B_{b2}$ are the same as those appearing in Theorem 4.1 of Ruppert and Wand (1994), which confirms our intuition that $\pi$ can be treated as an interior point in the partially polynomial estimation. In case (a), the optimal bandwidth to minimize the MSE is $O\left(n^{-\frac{1}{2q+3}}\right)$; in case (b1), the optimal bandwidth is $O\left(n^{-\frac{1}{2p+3}}\right)$; in case (b2), the optimal bandwidth is $O\left(n^{-\frac{1}{2p+5}}\right)$. When we have more smoothness in $m(x)$, the optimal bandwidth is larger. Note also that $N_p$, $\Omega_p^+$, $\Omega_p^-$, $Q_{pq}^+$, $Q_{pq}^+$ and $Q_p$ only depend on the kernel function. This fact convinces the conventional insight that the bandwidth affects the convergence rate while the kernel only affects the efficiency constant. Also, $K_p^*(\cdot)$ instead of $k(\cdot)$ appears in these notations. This convinces the observation in the introduction that the LPE at a interior point is equivalent to the local constant estimator with a higher-order kernel. When $q = p = 0$, $K_p^*(u) = k(u)$, and $N_p$ reduces to $2\int_0^M K_0^2(w)dw$ in Porter (2003), where $K_0(w) = \int_w^M k(u)du$. Now, we check some special cases in (a) to show the results above are right. Suppose $m^{(q+1)}(\pi+) = m^{(q+1)}(\pi-)$, then

$$\begin{aligned}
& Q_{pq}^+(i) + Q_{pq}^-(i) \\
= & \int_0^M K_p^*(\overline{\delta}_{i-1}(w+)) \left( \int_{-M}^M K_p^*(u) (w+u)^{q+1} du - w^{q+1} \right) dw \\
& + \int_{-M}^0 K_p^*(\overline{\delta}_{i-1}(w-)) \left( \int_{-M}^M K_p^*(u) (w+u)^{q+1} du - w^{q+1} \right) dw \\
= & \begin{cases} 0, & \text{if } q < p; \\ \left( \int_{-M}^M K_p^*(u) u^{p+1} du \right) Q_p(i), & \text{if } q = p \text{ and } p \text{ odd}; \\ 0, & \text{if } q = p \text{ and } p \text{ even}; \end{cases}
\end{aligned}$$

which matches the asymptotic biases in (b1) and (b2).

# 3 Discussions

## 3.1 Practical Issues

### 3.1.1 Bandwidth Selection

The bandwidth is a necessary input in any kernel estimators. Section 5 of Imbens and Lemieux (2008) discusses this issue when the LLE is used to estimate the treatment effect. The least squares cross-validation (CV) approach is suggested, but as argued in Ludwig and Miller (2007), "there is currently no widely agreed-upon method for selection of optimal bandwidths in the nonparametric RD context". Ludwig and Miller (2005) analyze the causes for the bad performance of the bandwidth based on CV. First, the convergence rate of the CV bandwidth to the optimal bandwidth is very slow. This can be seen from Hardle et al (1988) who show that the convergence rate of the CV bandwidth is in the order of $n^{-1/10}$ when the covariate is

one-dimensional as in the RD design. Second, in the RD design, the conditional mean at a boundary point is estimated, while the usual CV procedure is designed for the interior point. Third, the usual CV method concentrates on the global feature of the data, while we are interested in the local nature of the data in the RD design.

Only the optimal bandwidth at a single point $\pi$ is of interest, so all suggested methods in the literature utilize only the data in the neighborhood of $\pi$. Based on the criterion used, the methods can be roughly divided into two classes. The first class targets to minimize an approximation to the MSE or MISE criterion on $\widehat{E}[y|\pi+]$ and $\widehat{E}[y|\pi-]$ separately. For example, DeJardins and McCall (2008) use the criterion

$$E\left[\left(\widehat{E}[y|\pi+]-E[y|\pi+]\right)^2+\left(\widehat{E}[y|\pi-]-E[y|\pi-]\right)^2\right],$$

where $E[y|x\pm]=E[y_i|x_i=x\pm]$ and $\widehat{E}[y|x\pm]$ is its estimator. Ludwig and Miller (2005) suggest the CV criterion

$$BCV_\tau(h)=\frac{1}{n}\sum_{i:x_i\in[Q^-_{1-\tau},Q^+_\tau]}\left[y_i-\widehat{m}_{-i}(x_i)\right]^2, \tag{11}$$

where $Q^+_\tau$ is the $\tau$th quantile of $\{x_i|x_i\geq\pi\}$, and $Q^-_{1-\tau}$ is the $(1-\tau)$th quantile of $\{x_i|x_i<\pi\}$ with $\tau$ converging to zero. $\widehat{m}_{-i}(x_i)=\begin{cases}\widehat{\alpha}_l(x_i), & \text{if } x_i<\pi;\\ \widehat{\alpha}_r(x_i), & \text{if } x_i\geq\pi.\end{cases}$ $\widehat{\alpha}_l(x)$ is determined by the minimizer $\widehat{a}$ in a similar minimization problem as (4):

$$\min_{a,b_1,\cdots,b_p}\frac{1}{n}\sum_{j:x_j<x}k_h(x_j-x)\left[y_j-a-b_1(x_j-x)-\cdots-b_p(x_j-x)^p\right]^2. \tag{12}$$

and $\widehat{\alpha}_r(x)$ is determined similarly as in (12) but the index in the summation is replaced by $\{j:x_j>x\}$. Note that $x_i$ is not used in the estimation of $\widehat{m}_{-i}(x_i)$ to match the principle of cross validation. Such a procedure essentially minimizes the criterion

$$E\left[\int_{\pi\leq x\leq\overline{Q}^+_\tau}\left(\widehat{E}[y|x+]-E[y|x+]\right)^2f(x)dx+\int_{\overline{Q}^-_{1-\tau}\leq x<\pi}\left(\widehat{E}[y|x-]-E[y|x-]\right)^2f(x)dx\right],$$

where $\overline{Q}^+_\tau$ is the $\tau$th quantile of the conditional random variable $x|x\geq\pi$, and $\overline{Q}^-_{1-\tau}$ is the $(1-\tau)$th quantile of $x|x<\pi$. The second class targets the difference between $\widehat{E}[y|\pi+]$ and $\widehat{E}[y|\pi-]$. For example, Imbens and Kalyanaraman (2009) use the criterion

$$E\left[\left(\left(\widehat{E}[y|\pi+]-\widehat{E}[y|\pi-]\right)-\left(E[y|\pi+]-E[y|\pi-]\right)\right)^2\right].$$

All the methods mentioned above seem to concern about the boundary problem in the RD design, while this is not a problem in the partially polynomial estimation since $\widehat{m}(\pi+)$ and $\widehat{m}(\pi-)$ do not appear explicitly. Given that only the conditional mean at an interior point is estimated in the partially polynomial estimation, the usual CV procedure can be applied here. Specifically, we minimize

$$CV_\tau(h;\alpha,\beta)=\frac{1}{n}\sum_{i:x_i\in[Q^-_{1-\tau},Q^+_\tau]}\left[\overrightarrow{y}_i-\widehat{m}_{-i}\left(x_i|\overrightarrow{\mathbf{y}}\right)\right]^2, \tag{13}$$

where $\widehat{m}_{-i}\left(x_i|\overrightarrow{\mathbf{y}}\right)$ is the same as $\widehat{m}\left(x_i|\overrightarrow{\mathbf{y}}\right)$ except that $x_i$ is not used in the estimation. As argued in Section

2.3, the PPE is applicable because $(\alpha, \beta)$ can be treated as the same in the right $h$ neighborhood of $\pi$. So (13) balances the similarity of the data in the right neighborhood of $\pi$ and the quality of the local polynomial fitting, while (11) only consider the latter criterion. In consequence, (13) essentially minimizes the criterion

$$E\left[\int_{\overline{Q}_{1-\tau}^{-} \leq x \leq \overline{Q}_{\tau}^{+}} \left(\widehat{m}(x) - m(x)\right)^2 f(x)dx\right] + E\left[\int_{\pi \leq x \leq \pi + Mh} \left(\widehat{\alpha} - \alpha\right)^2 f(x)dx\right].$$

Usually, $h$ can be estimated by a profiled procedure. For each $h$, $(\alpha, \beta)$ is calculated under this fixed $h$, so the intermediate estimate $\left(\widetilde{\alpha}, \widetilde{\beta}\right)$ is a function of $h$, denoted as $\left(\widetilde{\alpha}(h), \widetilde{\beta}(h)\right)$. Then search over $h$ to find the optimal $\widehat{h}$, and the PPE $\left(\widehat{\alpha}, \widehat{\beta}\right)$ is calculated using $\widehat{h}$ and denoted as $\left(\widehat{\alpha}\left(\widehat{h}\right), \widehat{\beta}\left(\widehat{h}\right)\right)$. We can select different $h$'s for $x_i \geq \pi$ and $x_i < \pi$, but as argued in Imbens and Lemieux (2008) and Imbens and Kalyanaraman (2009), a single $h$ is enough if $f(x)$ and the smoothness of $m(x)$ are similar on both sides of the cutoff point.[4]

In practice, we use the following Algorithm CV to select $h$ based on the PPE. Suppose $\mathbf{x}$ is sorted ascendingly, and $\mathbf{y}$ is arranged correspondingly. $n_{low} = \sum 1(x_i < \pi)$, and $n_{up} = n - n_{low}$.

**Algorithm CV**

1. Specify a $\tau \in (0, 1)$; e.g., $\tau = 0.5$. The range of the bandwidth is set as $R_h = [h_{low}, h_{up}]$, where

$$\begin{aligned} h_{low} &= \max\left\{x_{[n_{low}(1-\tau)]+i} - x_{[n_{low}(1-\tau)]+i-1} : i = 1, \cdots, n_{low} - [n_{low}(1-\tau)] + [n_{up}\tau]\right\}, \\ h_{up} &= x_{n_{low}+[n_{up}\tau]} - x_{[n_{low}(1-\tau)]}, \end{aligned}$$

   and $[z]$ for $z \in \mathbb{R}$ is the largest integer no greater than $z$.

2. Denote the profiled objective function in (13) as $CV(h)$, and minimize $CV(h)$ with respect to $h$ on $R_h$. In practice, we need only minimize $CV(h)$ on a discretized set $D_h$ of $h$; e.g., $D_h = \{h_{low} + i \cdot step : i = 1, \cdots, [n\tau]\}$, where $step = \frac{h_{up} - h_{low}}{[n\tau]}$.

In step 1, $h_{low}$ guarantees that there is at lease on data point in the $h$ neighborhood of any $x_i$, $i = [n_{low}(1-\tau)], \cdots, n_{low} + [n_{up}\tau]$. $h_{up}$ makes sure that at most $O(n\tau)$ data points are used in $\widehat{m}_{-i}\left(x_i|\overrightarrow{\mathbf{y}}\right)$. In step 2, the specification of $D_h$ exhausts almost all possible estimates of $\widehat{m}_{-i}\left(x_i|\overrightarrow{\mathbf{y}}\right)$. This is because the average distance between the contiguous $x_i$'s is roughly *step*, and thus increasing $h$ by one *step* is roughly equivalent to adding one more data point in the kernel smoothing of $\widehat{m}_{-i}\left(x_i|\overrightarrow{\mathbf{y}}\right)$. Such a specification of $R_h$ and $D_h$ is not rigid. In practice, we must make sure the minimizer is not obtained at the boundary points, $h_{low}$ and $h_{up}$, of $R_h$.

It is an interesting theoretical problem to derive the asymptotic distribution of $\widehat{h}$ as in Hardle et al (1988) and of $\widehat{\alpha}\left(\widehat{h}\right)$ as in Li and Racine (2004). Also, it is admirable to derive some optimality properties of $\widehat{h}$ as in Hardle and Marron (1985).

### 3.1.2 Other Issues

We discuss three other issues related to the application of the PPE in practice. The first issue is about the set estimation of the treatment effect. The asymptotic biases in Theorem 1 involve complicated functionals of the kernel and the derivatives of $f(x)$ and $m(x)$ at $\pi$, but their estimation is straightforward and is a byproduct of the partially polynomial estimation.[5] The estimation of the common variance can be obtained

[4] $\sigma^2(\pi+)$ and $\sigma^2(\pi-)$ also affect the bandwidth on each side of $\pi$.

[5] Another possibility is undersmoothing. As mentioned in Pagan and Ullah (1999), a tuning parameter that is good for estimation purposes is not necessarily good for testing purposes.

by Theorem 4 of Porter (2003) since the estimation procedure there only requires that $\widehat{\alpha}$ is consistent. Besides the Wald-type CI, there are two other alternatives. The wild bootstrap can be used to construct the CI for $\alpha$ to get a better finite-sample approximation of the distribution of $\widehat{\alpha}$. Ludwig and Miller (2007) use the paired-bootstrap to conduct inference of $\alpha$, but we suspect the paired-bootstrap is not valid here since the more stringent conditional moment restrictions instead of the orthogonal conditions are imposed in the RD design. Another method for the CI construction of $\alpha$ is based on empirical likelihood which carries out the studentizing internally and adopts Bartlett correction; see, e.g., Chen and Qin (2000). The second issue is about the choice of the order of the PPE. This problem is considered for the LPE in Sun (2005). His procedure is ready to be adapted to the PPE case since only an estimator of $\alpha$ is needed there. We can combine the bandwidth selection and the order choice in one algorithm to robustify the bandwidth selection; see also Fan and Gijbels (1995) for more discussions on this issue in standard cases with interior points. The third issue is about the kernel selection. Since the Epanechnikov kernel is optimal in minimizing MSE and MISE at interior points and is nearly optimal at the most boundary point, we recommended to use this kernel function. Of course, the kernel that minimizes the MSE of the PPE is still unknown.

### 3.2 Extensions

We discuss four extensions in this subsection. The first extension is to adapt the procedure of the PPE in the sharp design to the fuzzy design. A similar procedure can be used to estimate $E[T|x=\pi+]-E[T|x=\pi-]$ by changing $\{y_i\}_{i=1}^n$ to $\{T_i\}_{i=1}^n$. The asymptotic distribution can be derived in a similar way as in Section 3.6 of Porter (2003). For example, the covariance there is

$$C_{\alpha\vartheta}=\frac{e_1'N_p^{-1}\left[\sigma_{\varepsilon\eta}(\pi+)\,\Omega_p^{+}+\sigma_{\varepsilon\eta}(\pi-)\,\Omega_p^{-}\right]N_p^{-1}e_1}{f(\pi)},$$

where $\eta$ is the error term in the expression $T=t(x)+\vartheta d+\eta$ with $E[\eta|x,d]=0$, and $\sigma_{\varepsilon\eta}(x)=E[\varepsilon\eta|x]$. As to the bandwidth selection, it is appropriate to choose a different bandwidth for the treatment rule from the conditional mean of the outcome, so (13) can be easily extended; see Section 5.2 of Imbens and Lemieux (2008) for a discussion about this issue. The second extension is to consider the estimation of the treatment effect when there are additional covariates $z$. This problem is also discussed in Frölich (2007). Suppose

$$y=m(x,z)+\alpha(z)d+\varepsilon,\text{ where }E[\varepsilon|x,z,d]=0,\ d=1(x\geq\pi),$$

and we are interested in $\alpha=\int\alpha(z)dF(z)$. For each $z_i$, estimate $\alpha(z_i)$ using the procedure in Section 2 except that the kernel $k(\cdot)$ is put on the $(x,z)$ space instead of the $x$ space only and in the definition of $\overrightarrow{y}_i$, the expansion at $(\pi,z_i)$ instead of $\pi$ is included. Then $\widehat{\alpha}=\frac{1}{n}\sum_{i=1}^n\alpha(z_i)$. The third extension is to cover the case where the cut-off point is unknown. From Porter and Yu (2010), we expect the estimation of $\pi$ will not affect the asymptotic distribution of $\widehat{\alpha}$. The fourth extension is related to the relative efficiency between the LPE and PPE. It is hard to compare the MSE of the PPE and the LPE, so a natural question is what is the efficiency constant for $\alpha$ as discussed in Donoho and Liu (1991).

## 4 Conclusion

In this paper, we propose a new estimator, the partially polynomial estimator, of the treatment effect in regression discontinuity design. Such an estimator can be treated as an extension of the partially linear estimator in Porter (2003) by also incorporating the derivative differences in the left and right neighborhoods

of the cut-off point. We show unlike the partially linear estimator, the partially polynomial estimator can achieve the optimal rate of convergence even under broader conditions of the data generating process. Moreover, we reveal the speciality of the partially linear estimator by noting that the form of its bias can not be extended to the partially polynomial estimator.

# References


Angrist, J.D. and V. Lavy, 1999, Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement, *Quarterly Journal of Economics*, 114, 533–575.

Battistin, E. and E. Rettore, 2002, Testing for Programme Effects in a Regression Discontinuity Design with Imperfect Compliance, *Journal of the Royal Statistical Society, Series A*, 165, 39–57.

Black, S., 1999, Do Better Schools Matter? Parental Valuation of Elementary Education, *Quarterly Journal of Economics*, 114, 577–599.

Card, D. et al, 2008, Tipping and the Dynamics of Segregation in Neighborhoods and Schools, *Quarterly Journal of Economics*, 123, 177-218.

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

Chay, K. and M. Greenstone, 2005, Does Air Quality Matter? Evidence From the Housing Market, *Journal of Political Economy*, 113, 376–424.

Chay, K. et al, 2005, The Central Role of Noise in Evaluating Interventions that Use Test Scores to Rank Schools, *American Economic Review*, 95, 1237–1258.

Chen, S.X. and Y.S. Qin, 2000, Empirical Likelihood Confience Intervals for Local Linear Smoothers, *Biometrika*, 87, 946-953.

DesJardins, S.L., and B.P. McCall, 2008, The Impact of the Gates Millennium Scholars Program on the Retention, College Finance- and Work-Related Choices, and Future Educational Aspirations of Low-Income Minority Students, unpublished manuscript, Department of Economics, University of Michigan.

DiNardo, J. and D.S. Lee, 2004, Economic Impacts of New Unionization on Private Sector Employers: 1984–2001, *Quarterly Journal of Economics*, 119, 1383–1441.

Donoho, D.L. and R.C. Liu, 1991, Geometrizing Rates of Convergence, II, *Annals of Statistics*, 19, 633-667.

Fan, J. et al, 1997, Local Polynomial Regression: Optimal Kernels and Asymptotic Minimax Efficiency, *Annals of the Institute of Statistical Mathematics*, 49, 79-99.

Fan, J. and I. Gijbels, 1995, Adaptive Order Polynomial Fitting: Bandwidth Robustification and Bias Reduction, *Journal of Computational and Graphical Statistics*, 4, 213-227.

Fan, J. and I. Gijbels, 1996, *Local Polynomial Modelling and its Applications*, Chapman & Hall, London.

Frölich, M., 2007, Regression Discontinuity Design with Covariates, IZA Discussion Paper 3024 Bonn.

Gasser, T. et al, 1985, Kernels for Nonparametric Curve Estimation, *Journal of the Royal Statistical Society: Series B*, 47, 238-252.

Hahn, J. et al, 2001, Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design, *Econometrica*, 69, 201–209.

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

Hardle, W. and J.S. Marron, 1985, Optimal Bandwidth Selection in Nonparametric Regression Function Estimation, *Annals of Statistics*, 13, 1465-1481.

Hardle, W. et al, 1988, How Far are Automatically Chosen Regression Smoothing Parameters from Their Optimum?, *Journal of the American Statistical Association*, 83, 86-95.

Imbens, G.W. and T. Lemieux, 2008, Regression Discontinuity Designs: A Guide to Practice, *Journal of Econometrics*, 142, 615-635.

Imbens, G.W. and K. Kalyanaraman, 2009, Optimal Bandwidth Choice for the Regression Discontinuity Estimator, cemmap working paper CWP05/10.

Jacob, B.A. and L. Lefgren, 2004, Remedial Education and Student Achievement: A Regression-discontinuity Analysis, *The Review of Economics and Statistics*, 86, 226–244.

Lee, D.S., 2008, Randomized Experiments From Non-random Selection in U.S. House Elections, *Journal of Econometrics*, 142, 675–697.

Lee, D.S., and T. Lemieux, 2009, Regression Discontinuity Designs in Economics, Working Paper, Department of Economics, Princeton University.

Li, Q., and J. Racine, 2004, Cross-validated Local Linear Nonparametric Regression, *Statistica Sinica*, 14, 485-512.

Li, Q., and J. Racine, 2007, *Nonparametric Econometrics: Theory and Practice*, Princeton, N.J.: Princeton University Press.

Ludwig, J. and D. Miller, 2005, Does Head Start Improve Children's Life Chances? Evidence From a Regression Discontinuity Disign, NBER Working Paper 11702.

Ludwig, J. and D. Miller, 2007, Does Head Start Improve Children's Life Chances? Evidence From a Regression Discontinuity Disign, *Quarterly Journal of Economics*, 122, 159-208.

Masry, E., 1996, Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates, *Journal of Time Series Analysis,* 17, 571-599.

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

Newey, W.K. and D.L. McFadden, 1994, Large Sample Estimation and Hypothesis Testing, *Handbook of Econometrics*, Vol. 4, R.F. Eagle and D.L. McFadden, eds., Elsevier Science B.V., Ch. 36, 2113-2245.

Pagan, A. and A. Ullah, 1999, *Nonparametric Econometrics*, New York: Cambridge University Press.

Pence, K., 2006, Foreclosing on Opportunity: State Laws and Mortgage Credit, *The Review of Economics and Statistics*, 88, 177-182.

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

Porter, J. and P. Yu, 2010, Regression Discontinuity with Unknown Discontinuity Points: Testing and Estimation, Mimeo, Department of Economics, University of Wisconsin at Madison.

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

Ruppert, D. and M.P. Wand, 1994, Multivariate Locally Weighted Least Squares Regression, *Annals of Statistics*, 22, 1346-1370.

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

Sun, Y., 2005, Adaptive Estimation of the Regression Discontinuity Model, Unpublished Manuscript, Department of Economics, University of California at San Diego.

Thistlewaite, D. and D. Campbell, 1960, Regression-discontinuity Analysis: an Alternative to the Ex-post facto Experiment, *Journal of Educational Psychology*, 51, 309–317.

Trochim, W., 1984, *Research Design for Program Evaluation: The Regression Discontinuity Approach*, Beverly Hills: Sage Publications.

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

Van der Klaauw, W., 2008, Regression–Discontinuity Analysis: A Survey of Recent Developments in Economics, *Labour*, 22, 219-245.

Yu, P., 2007, Likelihood-based Estimation and Inference in Threshold Regression, unpublished manuscript, Department of Economics, University of Wisconsin at Madison.

# Appendix A: Proof of Theorem 1

From (5) and (6),

$$\sqrt{nh}\left(\widehat{\alpha}-\alpha\right)=e_1'\left(\frac{1}{nh}\sum_{l=1}^{n}\underline{Z}_l^d\left(\pi\right)\underline{Z}_l^{d'}\left(\pi\right)\right)^{-1}\left(\frac{1}{\sqrt{nh}}\sum_{l=1}^{n}\underline{Z}_l^d\left(\pi\right)\left(m(x_l)-\overline{m}(x_l)+\varepsilon_l-\overline{\varepsilon}_l\right)\right).$$

We first analyze the numerator, then the denominator. For $1\leq i\leq q+1$, the $i$th term of $\underline{Z}_l^d\left(\pi\right)$ is

$$\begin{aligned}
&\left(\frac{x_l-\pi}{h}\right)^{i-1}1\left(x_l\geq\pi\right)-\frac{1}{h^{i-1}}P_{x_l}^n(X^{i-1d}\left(\pi\right))\\
=\;& e_1'S_n^{-1}(x_l)(S_n^+(x_l)+S_n^-(x_l))e_1\left(\frac{x_l-\pi}{h}\right)^{i-1}d_l-e_1'S_n^{-1}(x_l)\frac{1}{n}\sum_{j=1}^{n}Z_j\left(x_l\right)k_h\left(x_j-x_l\right)\left(\frac{x_j-\pi}{h}\right)^{i-1}d_j\\
=\;& e_1'S_n^{-1}(x_l)\frac{1}{n}\sum_{j=1}^{n}Z_j\left(x_l\right)k_h\left(x_j-x_l\right)\left(Z_l^i(\pi)d_l-Z_j^i(\pi)\right)d_j\\
&+e_1'S_n^{-1}(x_l)\frac{1}{n}\sum_{j=1}^{n}Z_j\left(x_l\right)k_h\left(x_j-x_l\right)Z_l^i(\pi)d_ld_j^c\\
\equiv\;& e_1'S_n^{-1}(x_l)\delta_{ni-1}^+(x_l)+e_1'S_n^{-1}(x_l)\delta_{ni-1}^-(x_l)\equiv e_1'S_n^{-1}(x_l)\delta_{ni-1}(x_l).
\end{aligned}$$

Here, $\delta^{+}_{ni-1}(x_l)$ plays the role of $-\widehat{f}_{+}(x_l)(1-d_l)$, $\delta^{-}_{ni-1}(x_l)$ plays the role of $\widehat{f}_{-}(x_l)d_l$, and $S_n(x_l)$ plays the role of $\widehat{f}(x_l)$ in Porter (2003).

## Numerator

Concentrate on the $i$th term and take an expansion to linearize. We need different linearizations under Assumption 2(a) and 2(b). We first discuss the case under Assumption 2(a), then under Assumption 2(b).

### Under Assumption 2(a)

The $i$th term of the numerator is

$$\begin{aligned}
& \frac{1}{\sqrt{nh}}\sum_{l=1}^{n} e_1' S_n^{-1}(x_l)\delta_{ni-1}(x_l)\left(m(x_l)-\overline{m}(x_l)+\varepsilon_l-\overline{\varepsilon}_l\right) \\
= & \frac{1}{\sqrt{nh}}\sum_{l=1}^{n} e_1' \overline{S}(x_l)^{-1}\overline{\delta}_{i-1}(x_l)\left(-\overline{L}(m(x_l))+\varepsilon_l-P_{x_l}(\mathbf{e})\right) \\
& +\frac{1}{\sqrt{nh}}\sum_{l=1}^{n} e_1' \overline{S}(x_l)^{-1}\overline{\delta}_{i-1}(x_l)\left(\overline{L}(m(x_l))-L(m(x_l))\right) \\
& +\frac{1}{\sqrt{nh}}\sum_{l=1}^{n} L_{i-1}(x_l)\varepsilon_l+R_n \\
\equiv & \ \text{Term1}+\text{Term2}+\text{Term3}+R_n,
\end{aligned}$$

where

$$\begin{aligned}
L(m(x)) &= e_1' S^{-1}(x)\,\overline{r}(m(x))-e_1' S^{-1}(x)\left(S_n(x)-S(x)\right)S^{-1}(x)\,\overline{r}(m(x))+e_1' S^{-1}(x)\left(\widetilde{r}(m(x))-\overline{r}(m(x))\right), \\
\overline{L}(m(x)) &= e_1' S^{-1}(x)\,\overline{r}(m(x))-e_1' S^{-1}(x)\left(\overline{S}(x)-S(x)\right)S^{-1}(x)\,\overline{r}(m(x)), \\
L_{i-1}(x) &= e_1'\overline{S}^{-1}(x)\left(\delta_{ni-1}(x)-\overline{\delta}_{i-1}(x)\right)-e_1'\overline{S}^{-1}(x)\left(S_n(x)-\overline{S}(x)\right)\overline{S}^{-1}(x)\overline{\delta}_{i-1}(x), \\
P_x(\boldsymbol{\varepsilon}) &= e_1' S^{-1}(x)\,\widetilde{r}(\varepsilon(x)), \\
\overline{\delta}_{i-1}(x) &= \overline{\delta}^{+}_{i-1}(x)+\overline{\delta}^{-}_{i-1}(x),
\end{aligned}$$

with

$$\begin{aligned}
\widetilde{r}(m(x)) &= \frac{1}{n}\sum_{j=1}^{n} Z_j(x)k_h(x_j-x)\left\{m(x_j)-m(x)-\sum_{k=1}^{q}\frac{m^{(k)}(x)}{k!}(x_j-x)^k\right\} \\
\widetilde{r}(\varepsilon(x)) &= \frac{1}{n}\sum_{j=1}^{n} Z_j(x)k_h(x_j-x)\,\varepsilon_j, \\
\overline{r}(m(x)) &= \int \delta(u)f(x+uh)\left\{m(x+uh)-m(x)-\sum_{k=1}^{q}\frac{m^{(k)}(x)}{k!}(uh)^k\right\}du, \\
\overline{S}(x) &= E\left[Z_j(x)\,Z_j'(x)\,k_h(x_j-x)\right]=\left(\int u^{i+j-2}k(u)\,f(x+uh)du\right)_{(p+1)\times(p+1)},
\end{aligned}$$

$$
\begin{aligned}
\overline{\delta}_{i-1}^{+}(x) &= E\left[Z_j(x)\,k_h(x_j-x)\left(\left(\frac{x-\pi}{h}\right)^{i-1}1(x\geq\pi)-\left(\frac{x_j-\pi}{h}\right)^{i-1}\right)d_j\right] \\
&= \left(\int_{-M}^{M}\delta(u)\left(\left(\frac{x-\pi}{h}\right)^{i-1}1(x\geq\pi)-\left(\frac{x-\pi}{h}+u\right)^{i-1}\right)f(x+uh)1\left(x+uh\geq\pi\right)du\right)_{(p+1)\times 1}, \\
\overline{\delta}_{i-1}^{-}(x) &= E\left[Z_j(x)\,k_h(x_j-x)\left(\frac{x-\pi}{h}\right)^{i-1}1(x\geq\pi)d_j^c\right] \\
&= \left(\int_{-M}^{M}\delta(u)\left(\frac{x-\pi}{h}\right)^{i-1}1(x\geq\pi)f(x+uh)1\left(x+uh<\pi\right)du\right)_{(p+1)\times 1},
\end{aligned}
$$

and $R_n$ is the remainder term including the quadratic terms in the expansion:

$$
\begin{aligned}
R_n &= -\frac{1}{\sqrt{nh}}\sum_{l=1}^{n}e_1'\overline{S}^{-1}(x_l)\overline{\delta}_{i-1}(x_l)R(\widetilde{y}(x_l)) \\
&\quad+\frac{1}{\sqrt{nh}}\sum_{l=1}^{n}R_{i-1}(x_l)\left(m(x_l)-\widetilde{m}(x_l)+\varepsilon_l\right) \\
&\quad+\frac{1}{\sqrt{nh}}\sum_{l=1}^{n}L_{i-1}(x_l)\left(m(x_l)-\widetilde{m}(x_l)\right),
\end{aligned}
$$

with

$$
\begin{aligned}
R(\widetilde{y}(x)) &= e_1'S^{-1}(x)\left(S_n(x)-S(x)\right)S^{-1}(x)\left(S_n(x)-S(x)\right)S_n^{-1}(x)\overline{r}(m(x)) \\
&\quad -e_1'S^{-1}(x)\left(S_n(x)-S(x)\right)S_n^{-1}(x)\left(\widetilde{r}(\widetilde{y}(x))-\overline{r}(m(x))\right), \\
\widetilde{r}(\widetilde{y}(x)) &= \widetilde{r}(m(x))+\widetilde{r}(\varepsilon(x)), \\
R_{i-1}(x) &= e_1'\overline{S}^{-1}(x)\left(S_n(x)-\overline{S}(x)\right)\overline{S}^{-1}(x)\left(S_n(x)-\overline{S}(x)\right)S_n^{-1}(x)\overline{\delta}_{i-1}(x) \\
&\quad -e_1'\overline{S}^{-1}(x)\left(S_n(x)-\overline{S}(x)\right)S_n^{-1}(x)\left(\delta_{ni-1}(x)-\overline{\delta}_{i-1}(x)\right).
\end{aligned}
$$

$L(m(x))$ is the linear expansion of $P_x^n(m(\mathbf{x}))-m(x)$ as shown in Lemma 2, and $\overline{L}(m(x))$ is its mean. $L_{i-1}(x)$ is the linear expansion of $e_1'S_n^{-1}(x)\delta_{ni-1}(x)$ at $e_1'\overline{S}^{-1}(x)\overline{\delta}_{i-1}(x)$. Note that $e_1'S_n^{-1}(x)\delta_{ni-1}(x)$ is linearized at $\overline{S}^{-1}(x)$ and $\overline{\delta}_{i-1}(x)$ instead of their limits which are $S^{-1}(x)$ and 0 respectively.[6] This is mainly because $\overline{\delta}_{i-1}(x)$ is not a smooth function of $x$ when $x$ is in a neighborhood of $\pi$. As a result, $S_n^{-1}(x)$ can not be linearized at $S^{-1}(x)$, or $R_{i-1}(x)$ can not be a higher-order term.

Our analysis includes three steps. In step1, we show $R_n=o_p(1)$. In step 2, we show Term3 $=o_p(1)$ and Term2 $=o_p(1)$. In step 3, we show $-\overline{L}(m(x_l))$ in Term1 contributes to the bias, and $\varepsilon_l-\overline{\varepsilon}_l$ contributes to the variance. Although there is randomness in Term 2, it does not contribute to the asymptotic distribution. With the three steps in hand, the Liapunov central limit theorem is applied to find the asymptotic distribution.

**Step 1** First, some basic results. $\sup_{x\in N_0}S_n^{-1}(x)=\sup_{x\in N_0}\overline{S}^{-1}(x)+o_p(1)=O_p(1)$ from Lemma B5 of Porter (2003), Lemma 3 and 4, $\sup_{x\in N_0}\overline{\delta}_{i-1}(x)=O(1)$, $\sup_{x\in N_0}e_1'\overline{S}^{-1}(x)\overline{\delta}_{i-1}(x)=O(1)$, $\sup_{x\in N_0}\overline{r}(m(x))=O(h^{q+1})$,

[6] In Porter (2003), $\overline{f}_{+}(x_l)(1-d_l)$ and $\overline{f}_{-}(x_l)d_l$ converges to 0 for a fixed $x_l$ when $h$ converges to zero. This result can be applied to $\overline{\delta}_{i-1}^{-}(x_l)$ and $\overline{\delta}_{0}^{+}(x_l)$. For $i>1$, it is still true for $h^{i-1}\overline{\delta}_{i-1}^{+}(x_l)$.

$\sup_{x\in N_0}\frac{1}{f(x)} = O(1)$, $\sup_{x\in N_0}|\widetilde{m}(x)-m(x)| = O_p\left(\sqrt{\frac{\ln n}{nh}}+h^{q+1}\right)$, $\frac{1}{nh}\sum_{l=1}^{n}|\varepsilon_l|\,1\left(\pi-Mh\le x_l\le\pi+Mh\right) = O_p(1)$, where $N_0=[\pi-Mh,\pi+Mh]$.

(i)

$$\begin{aligned}
&\frac{1}{\sqrt{nh}}\sum_{l=1}^{n}e_1'\overline{S}^{-1}(x_l)\overline{\delta}_{i-1}(x_l)\cdot e_1'S^{-1}\left(x_l\right)\left(S_n\left(x_l\right)-S\left(x_l\right)\right)S^{-1}(x_l)\left(S_n\left(x_l\right)-S\left(x_l\right)\right)S_n^{-1}(x_l)\overline{r}(m(x_l))\\
\approx\;&\frac{1}{\sqrt{nh}}\sum_{l=1}^{n}e_1'\Gamma^{-1}\overline{\delta}_{i-1}(x_l)\cdot e_1'\Gamma^{-1}\left(S_n\left(x_l\right)-S\left(x_l\right)\right)\Gamma^{-1}\left(S_n\left(x_l\right)-S\left(x_l\right)\right)\Gamma^{-1}\overline{r}(m(x_l))\\
\approx\;&\frac{1}{\sqrt{nh}}\sum_{l=1}^{n}O(1)\left(O_p\left(\sqrt{\frac{\ln n}{nh}}\right)+h\right)\left(O_p\left(\sqrt{\frac{\ln n}{nh}}\right)+h\right)O(h^{q+1})\\
=\;&\sqrt{nh}O_p\left(\sqrt{\frac{\ln n}{nh}}+h\right)O_p\left(\sqrt{\frac{\ln n}{nh}}+h\right)O(h^{q+1})\\
=\;&O_p\left(\left(\frac{\ln n}{\sqrt{nh}}+h\sqrt{\ln n}+h^2\sqrt{nh}\right)h^{q+1}\right).
\end{aligned}$$

$$\begin{aligned}
&\frac{1}{\sqrt{nh}}\sum_{l=1}^{n}R_{i-1}(x_l)(m(x_l)-\widetilde{m}(x_l))\\
\approx\;&\sqrt{nh}\left[O_p\left(\sqrt{\frac{\ln n}{nh}}\right)O_p\left(\sqrt{\frac{\ln n}{nh}}\right)+O_p\left(\sqrt{\frac{\ln n}{nh}}\right)O_p\left(\sqrt{\frac{\ln n}{nh}}\right)\right]\left(O_p\left(\sqrt{\frac{\ln n}{nh}}\right)+h^{q+1}\right)\\
=\;&O_p\left(\frac{\ln n\sqrt{\ln n}}{nh}+\frac{\ln n}{\sqrt{nh}}h^{q+1}\right).
\end{aligned}$$

(ii)

$$\begin{aligned}
&\frac{1}{\sqrt{nh}}\sum_{l=1}^{n}R_{i-1}(x_l)\varepsilon_l\\
\approx\;&\sqrt{nh}O_p\left(\sqrt{\frac{\ln n}{nh}}\right)O_p\left(\sqrt{\frac{\ln n}{nh}}\right)\left(\frac{1}{nh}\sum_{l=1}^{n}|\varepsilon_l|\,1\left(\pi-Mh\le x_l\le\pi+Mh\right)\right)\\
=\;&O_p\left(\frac{\ln n}{\sqrt{nh}}\right)=o_p\left(1\right).
\end{aligned}$$

(iii)

$$\begin{aligned}
&\frac{1}{\sqrt{nh}}\sum_{l=1}^{n}e_1'\overline{S}^{-1}(x_l)\overline{\delta}_{i-1}(x_l)\cdot e_1'S^{-1}\left(x_l\right)\left(S_n\left(x_l\right)-S\left(x_l\right)\right)S_n^{-1}\left(x_l\right)\left(\widetilde{r}(\widetilde{y}(x_l))-\overline{r}(m(x_l))\right)\\
\approx\;&\sqrt{nh}O_p\left(\sqrt{\frac{\ln n}{nh}}+h\right)O_p\left(\sqrt{\frac{\ln n}{nh}}\right)=O_p\left(\frac{\ln n}{\sqrt{nh}}+h\sqrt{\ln n}\right).
\end{aligned}$$

(iv)

$$
\begin{aligned}
& \frac{1}{\sqrt{nh}}\sum_{l=1}^{n} L_{i-1}(x_l)\left(m(x_l)-\widetilde{m}(x_l)\right) \\
\approx\ & \sqrt{nh}O_p\left(\sqrt{\frac{\ln n}{nh}}\right)\left(O_p\left(\sqrt{\frac{\ln n}{nh}}\right)+h^{q+1}\right) \\
=\ & O_p\left(\frac{\ln n}{\sqrt{nh}}+h^{q+1}\sqrt{\ln n}\right).
\end{aligned}
$$

From Assumption B(a) and (i)-(iv) above, $R_n = o_p(1)$.

**Step 2** To prove Term3 $= o_p(1)$, we will use the U and V-statistic projection. First, note that

$$
\begin{aligned}
& \frac{1}{\sqrt{nh}}\sum_{l=1}^{n} L_{i-1}(x_l)\varepsilon_l \\
=\ & \frac{1}{\sqrt{nh}}\sum_{l=1}^{n} e_1'\overline{S}^{-1}(x_l)\left(\delta_{ni-1}^{+}(x_l)-\overline{\delta}_{i-1}^{+}(x_l)\right)\varepsilon_l+\frac{1}{\sqrt{nh}}\sum_{l=1}^{n} e_1'\overline{S}^{-1}(x_l)\left(\delta_{ni-1}^{-}(x_l)-\overline{\delta}_{i-1}^{-}(x_l)\right)\varepsilon_l \\
& -\frac{1}{\sqrt{nh}}\sum_{l=1}^{n} e_1'\overline{S}^{-1}(x_l)\left(S_n(x_l)-\overline{S}(x_l)\right)\overline{S}^{-1}(x_l)\overline{\delta}_{i-1}^{+}(x_l)\varepsilon_l-\frac{1}{\sqrt{nh}}\sum_{l=1}^{n} e_1'\overline{S}^{-1}(x_l)\left(S_n(x_l)-\overline{S}(x_l)\right)\overline{S}^{-1}(x_l)\overline{\delta}_{i-1}^{-}(x_l)\varepsilon_l \\
\equiv\ & T1+T2+T3+T4.
\end{aligned}
$$

Let $z_l = (x_l, \varepsilon_l)$. For $T1$,

$$
\frac{1}{\sqrt{nh}}\sum_{l=1}^{n} e_1'\overline{S}^{-1}(x_l)\delta_{ni-1}^{+}(x_l)\varepsilon_l = \sqrt{\frac{n}{h}}\frac{1}{n^2}\sum_{l=1}^{n}\sum_{j=1}^{n} b_n(z_l,z_j),
$$

where

$$
b_n(z_l,z_j) = e_1'\overline{S}^{-1}(x_l)Z_j(x_l)k_h(x_j-x_l)\left(\left(\frac{x_l-\pi}{h}\right)^{i-1}d_l-(\frac{x_j-\pi}{h})^{i-1}\right)d_j\varepsilon_l.
$$

Note that $b_n(z_l, z_l) = 0$ so that this term is a U-statistic. Under the Assumptions in Section 2.3, it is easy, although tedious in notations, to show that $E\left[b_n(z_l,z_j)^2\right] = O(1)$. Then by standard U-statistic projection results,

$$
T1 = \sqrt{\frac{n}{h}}O_p\left(\frac{\left(E\left[b_n(z_l,z_j)^2\right]\right)^{1/2}}{n}\right) = O_p\left(\frac{1}{\sqrt{nh}}\right) = o_p(1).
$$

$T2$ follows similarly.

For $T3$, let

$$
b_n(z_l,z_j) = e_1'\overline{S}^{-1}(x_l)\left(Z_j(x_l)Z_j'(x_l)k_h(x_j-x_l)\right)\overline{S}^{-1}(x_l)\overline{\delta}_{i-1}^{+}(x_l)\varepsilon_l,
$$

then

$$
\frac{1}{\sqrt{nh}}\sum_{l=1}^{n} e_1'\overline{S}^{-1}(x_l)S_n(x_l)\overline{S}^{-1}(x_l)\overline{\delta}_{i-1}^{+}(x_l)\varepsilon_l = \sqrt{\frac{n}{h}}\frac{1}{n^2}\sum_{l=1}^{n}\sum_{j=1}^{n} b_n(z_l,z_j).
$$

As above, $E\left[b_n(z_l,z_j)^2\right] = O(1)$. Also, it is easy to show that $E\left[|b_n(z_l,z_l)|\right] = O(1)$ for $n$ large enough. By

a V-statistic projection theorem; see, e.g., Lemma 8.4 of Newey and McFadden (1994),

$$\begin{aligned}
&\frac{1}{\sqrt{nh}}\sum_{l=1}^{n} e_1'\overline{S}^{-1}(x_l)\left(S_n(x_l)-\overline{S}(x_l)\right)\overline{S}^{-1}(x_l)\overline{\delta}_{i-1}^{+}(x_l)\varepsilon_l \\
= &\sqrt{\frac{n}{h}}O_p\left(\frac{\left(E\left[b_n(z_l,z_j)^2\right]\right)^{1/2}}{n}+\frac{E\left[|b_n(z_l,z_l)|\right]}{n}\right) \\
= &O_p\left(\frac{1}{\sqrt{nh}}\right).
\end{aligned}$$

$T4$ follows similarly.

To prove Term2 $= o_p(1)$, we will use the V-statistic projection again. First, note that

$$\begin{aligned}
&\frac{1}{\sqrt{nh}}\sum_{l=1}^{n} e_1'\overline{S}(x_l)^{-1}\overline{\delta}_{i-1}(x_l)\left(\overline{L}(m(x_l))-L(m(x_l))\right) \\
= &\frac{1}{\sqrt{nh}}\sum_{l=1}^{n} e_1'\overline{S}(x_l)^{-1}\overline{\delta}_{i-1}(x_l)\left(e_1'S^{-1}(x_l)\left(S_n(x_l)-\overline{S}(x_l)\right)S^{-1}(x_l)\overline{r}(m(x_l))-e_1'S^{-1}(x_l)\left(\widetilde{r}(m(x_l))-\overline{r}(m(x_l))\right)\right) \\
= &\left(\begin{array}{c}
\frac{1}{\sqrt{nh}}\sum_{l=1}^{n} e_1'\overline{S}(x_l)^{-1}\overline{\delta}_{i-1}(x_l)e_1'S^{-1}(x_l)\left(S_n(x_l)-\overline{S}(x_l)\right)S^{-1}(x_l)\overline{r}(m(x_l)) \\
-\frac{1}{\sqrt{nh}}\sum_{l=1}^{n} e_1'\overline{S}(x_l)^{-1}\overline{\delta}_{i-1}(x_l)e_1'S^{-1}(x_l)\left(\widetilde{r}(m(x_l))-\overline{r}(m(x_l))\right)
\end{array}\right) \\
\equiv &T5-T6.
\end{aligned}$$

For $T5$, let

$$b_n(x_l,x_j)=e_1'\overline{S}(x_l)^{-1}\overline{\delta}_{i-1}(x_l)e_1'S^{-1}(x_l)\left(Z_j(x_l)Z_j'(x_l)k_h(x_j-x_l)\right)S^{-1}(x_l)\overline{r}(m(x_l)),$$

then

$$T5=\sqrt{\frac{n}{h}}\frac{1}{n^2}\sum_{l=1}^{n}\sum_{j=1}^{n} b_n(x_l,x_j).$$

It is easy to show that $E\left[b_n(x_l,x_j)^2\right]=O(h^{2(q+1)})$ and $E\left[|b_n(x_l,x_j)|\right]=O(h^{q+1})$, so

$$\begin{aligned}
&\frac{1}{\sqrt{nh}}\sum_{l=1}^{n} e_1'S^{-1}(x_l)\left(S_n(x_l)-\overline{S}(x_l)\right)S^{-1}(x_l)\overline{r}(x_l) \\
= &\sqrt{\frac{n}{h}}O_p\left(\frac{\left(E\left[b_n(x_l,x_j)^2\right]\right)^{1/2}}{n}+\frac{E\left[|b_n(x_l,x_j)|\right]}{n}\right) \\
= &O_p\left(\frac{h^{q+1}}{\sqrt{nh}}\right)=o_p(1).
\end{aligned}$$

A similar proof can be applied to $T6$ except now

$$b_n(x_l,x_j)=e_1'\overline{S}(x_l)^{-1}\overline{\delta}_{i-1}(x_l)e_1'S^{-1}(x_l)Z_j(x_l)k_h(x_j-x_l)\left\{m(x_j)-m(x_l)-\sum_{k=1}^{q}\frac{m^{(k)}(x_l)}{k!}(x_j-x_l)^k\right\}.$$

**Step 3** First, analyze the bias term $\frac{1}{\sqrt{nh}}\sum_{l=1}^{n} e_1'\overline{S}^{-1}(x_l)\overline{\delta}_{i-1}(x_l)\left(-\overline{L}(m(x_l))\right)$.

$$
\begin{aligned}
& E\left[\frac{1}{\sqrt{nh}}\sum\nolimits_{l=1}^{n} e_1'\overline{S}^{-1}(x_l)\overline{\delta}_{i-1}(x_l)\overline{L}(m(x_l))\right] \\
\approx\ & \sqrt{\frac{n}{h}}\int\left[\int_{\frac{\pi-x}{h}}^{M} K_p^*(u)\left(\left(\frac{x-\pi}{h}\right)^{i-1}1(x\geq\pi)-(\frac{x-\pi}{h}+u)^{i-1}\right)f(x+uh)du\right. \\
& \left.+\int_{-M}^{\frac{\pi-x}{h}} K_p^*(u)\left(\frac{x-\pi}{h}\right)^{i-1}1(x\geq\pi)f(x+uh)du\right]\cdot \\
& \frac{e_1'\Gamma^{-1}}{f(x)}\left(\int\delta(u)f(x+uh)\left\{m(x+uh)-m(x)-\sum_{k=1}^{q}\frac{m^{(k)}(x)}{k!}(uh)^k\right\}du\right)dx \\
=\ & \sqrt{nh}\int\left[\int_{-M}^{M} K_p^*(u)\,w^{i-1}1(w\geq 0)\frac{f(\pi+wh+uh)}{f(\pi+wh)}du-\int_{-w}^{M} K_p^*(u)(w+u)^{i-1}\frac{f(\pi+wh+uh)}{f(\pi+wh)}du\right] \\
& e_1'\Gamma^{-1}\left(\int\delta(u)f(\pi+wh+uh)\left\{m(\pi+wh+uh)-m(\pi+wh)-\sum_{k=1}^{q}\frac{m^{(k)}(\pi+wh)}{k!}(uh)^k\right\}du\right)dw \\
\approx\ & \sqrt{nh}f(\pi)\int_0^M\left[w^{i-1}-\int_{-w}^{M} K_p^*(u)(w+u)^{i-1}du\right]e_1'\Gamma^{-1} \\
& \left(\left(\int_{-w}^{M}\delta(u)\frac{m^{(q+1)}(\pi+)}{(q+1)!}\left(((w+u)h)^{q+1}-(wh)^{q+1}\right)du\right)\right. \\
& \left.+\int_{-M}^{-w}\delta(u)\frac{m^{(q+1)}(\pi-)((w+u)h)^{q+1}-m^{(q+1)}(\pi+)(wh)^{q+1}}{(q+1)!}du\right)dw \\
& -\sqrt{nh}f(\pi)\int_{-M}^{0}\left(\int_{-w}^{M} K_p^*(u)(w+u)^{i-1}du\right)e_1'\Gamma^{-1} \\
& \left(\int_{-w}^{M}\delta(u)\frac{m^{(q+1)}(\pi+)((w+u)h)^{q+1}-m^{(q+1)}(\pi-)(wh)^{q+1}}{(q+1)!}du\right. \\
& \left.\left|+\int_{-M}^{-w}\delta(u)\frac{m^{(q+1)}(\pi-)}{(q+1)!}\left(((w+u)h)^{q+1}-(wh)^{q+1}\right)du\right)dw\right. \\
\equiv\ & \sqrt{nh}h^{q+1}\frac{f(\pi)}{(q+1)!}\left[m^{(q+1)}(\pi+)Q_{pq}^{+}(i)+m^{(q+1)}(\pi-)Q_{pq}^{-}(i)\right].
\end{aligned}
$$

where the third equality is from Taylor expanding both $m(\pi+wh+uh)$ and $m(\pi+wh)$ at $m(\pi)$.

In summary, under Assumption 2(a), the order of the bias is determined by $q$: the rate is $\sqrt{nh}h^{q+1}$, and the constant is

$$-\frac{1}{(q+1)!}\left[m^{(q+1)}(\pi+)\cdot e_1'N_p^{-1}Q_{pq}^{+}+m^{(q+1)}(\pi-)\cdot e_1'N_p^{-1}Q_{pq}^{-}\right]$$

Note that $Q_{pq}^{+}\neq Q_{pq}^{-}$, so even when $m^{(q+1)}(\pi+)=m^{(q+1)}(\pi-)$, the bias of order $\sqrt{nh}h^{q+1}$ does not disappear. This is because the form of $\widetilde{r}(m(x))$ determines the bias, while $\widetilde{r}(m(x))$ in the discussion above critically depends on the smoothness of $m(x)$.

Second, analyze the variance term $\frac{1}{\sqrt{nh}}\sum_{l=1}^{n} e_1'\overline{S}(x_l)^{-1}\overline{\delta}_{i-1}(x_l)\left(\varepsilon_l-P_{x_l}(\mathbf{e})\right)$. By the V-statistic projection,

this term is statistically equivalent to

$$\frac{1}{\sqrt{nh}}\sum_{j=1}^{n} e_1'\overline{S}(x_j)^{-1}\overline{\delta}_{i-1}(x_j)\varepsilon_j - \frac{1}{\sqrt{nh}}\sum_{j=1}^{n} E_{x_l}\left[e_1'\overline{S}(x_l)^{-1}\overline{\delta}_{i-1}(x_l)\frac{K_p^*\left(\frac{x_j-x_l}{h}\right)}{hf(x_l)}\right]\varepsilon_j.$$

The $(i,k)$ term of the variance matrix is

$$\begin{aligned}
&\frac{1}{nh}E\left\{\left[\sum_{j=1}^{n}\left(e_1'\overline{S}(x_j)^{-1}\overline{\delta}_{i-1}(x_j) - E_{x_l}\left[e_1'\overline{S}(x_l)^{-1}\overline{\delta}_{i-1}(x_l)\frac{K_p^*\left(\frac{x_j-x_l}{h}\right)}{hf(x_l)}\right]\right)\varepsilon_j\right]\right.\\
&\left.\cdot\left[\sum_{j=1}^{n}\left(e_1'\overline{S}(x_j)^{-1}\overline{\delta}_{k-1}(x_j) - E_{x_l}\left[e_1'\overline{S}(x_l)^{-1}\overline{\delta}_{k-1}(x_l)\frac{K_p^*\left(\frac{x_j-x_l}{h}\right)}{hf(x_l)}\right]\right)\varepsilon_j\right]\right\}\\
=\;& \sigma^2(\pi+)\int_0^M\left(e_1'\overline{S}(\pi+wh)^{-1}\overline{\delta}_{i-1}(\pi+wh) - E_{x_l}\left[e_1'\overline{S}(x_l)^{-1}\overline{\delta}_{i-1}(x_l)\frac{K_p^*\left(\frac{\pi+wh-x_l}{h}\right)}{hf(x_l)}\right]\right)\\
&\left(e_1'\overline{S}(\pi+wh)^{-1}\overline{\delta}_{k-1}(\pi+wh) - E_{x_l}\left[e_1'\overline{S}(x_l)^{-1}\overline{\delta}_{k-1}(x_l)\frac{K_p^*\left(\frac{\pi+wh-x_l}{h}\right)}{hf(x_l)}\right]\right)f(\pi+wh)dw\\
&+\sigma^2(\pi-)\int_{-M}^0\left(e_1'\overline{S}(\pi+wh)^{-1}\overline{\delta}_{i-1}(\pi+wh) - E_{x_l}\left[e_1'\overline{S}(x_l)^{-1}\overline{\delta}_{i-1}(x_l)\frac{K_p^*\left(\frac{\pi+wh-x_l}{h}\right)}{hf(x_l)}\right]\right)\\
&\left(e_1'\overline{S}(\pi+wh)^{-1}\overline{\delta}_{k-1}(\pi+wh) - E_{x_l}\left[e_1'\overline{S}(x_l)^{-1}\overline{\delta}_{k-1}(x_l)\frac{K_p^*\left(\frac{\pi+wh-x_l}{h}\right)}{hf(x_l)}\right]\right)f(\pi+wh)dw\\
\approx\;& f(\pi)\left[\sigma^2(\pi+)\cdot\Omega_p^+(i,k) + \sigma^2(\pi-)\cdot\Omega_p^-(i,k)\right].
\end{aligned}$$

To apply the Liapunov central limit theorem, it suffices that for some $\zeta > 0$,

$$\sum_{j=1}^{n} E\left|\frac{1}{\sqrt{nh}}\left[e_1'\overline{S}(x_j)^{-1}\overline{\delta}_{i-1}(x_j) - E_{x_l}\left[e_1'\overline{S}(x_l)^{-1}\overline{\delta}_{i-1}(x_l)\frac{K_p^*\left(\frac{x_j-x_l}{h}\right)}{hf(x_l)}\right]\right]\varepsilon_j\right|^{2+\zeta} = o(h^{(i-1)(2+\zeta)}),$$

which is bounded by $C\sum_{j=1}^{n}\left[E\left|\frac{1}{\sqrt{nh}}e_1'\overline{S}(x_j)^{-1}\overline{\delta}_{i-1}(x_j)\varepsilon_j\right|^{2+\zeta} + E\left|\frac{1}{\sqrt{nh}}E_{x_l}\left[e_1'\overline{S}(x_l)^{-1}\overline{\delta}_{i-1}(x_l)\frac{K_p^*\left(\frac{x_j-x_l}{h}\right)}{hf(x_l)}\right]\varepsilon_j\right|^{2+\zeta}\right]$ for some $C > 0$.

$$\begin{aligned}
&\sum_{j=1}^{n} E\left|\frac{1}{\sqrt{nh}}e_1'\overline{S}(x_j)^{-1}\overline{\delta}_{i-1}(x_j)\varepsilon_j\right|^{2+\zeta}\\
\le\;& \frac{1}{(nh)^{\zeta/2}}\sup_{x\in N_0}E\left[|\varepsilon|^{2+\zeta}|x\right]\sup_{x\in N_0}\left|e_1'\overline{S}(x)^{-1}\overline{\delta}_{i-1}(x)\right|^{2+\zeta}\frac{1}{h}E\left[1\left(\pi - Mh \le x \le \pi + Mh\right)\right]\\
\le\;& O\left(\frac{1}{(nh)^{\zeta/2}}\right) = o(1).
\end{aligned}$$

Another term can be bounded similarly, so the Liapunov condition is satisfied.

**Under Assumption 2(b)**

Under Assumption 2(b), redefine

$$\begin{aligned}\widetilde{r}(m(x)) &= \frac{1}{n}\sum_{j=1}^{n} Z_j(x)k_h\left(x_j - x\right)\left\{m(x_j) - m(x) - \sum_{k=1}^{p}\frac{m^{(k)}(x)}{k!}\left(x_j - x\right)^k\right\},\\ \overline{r}(m(x)) &= \int \delta(u)f(x+uh)\left\{m(x+uh) - m(x) - \sum_{k=1}^{p}\frac{m^{(k)}(x)}{k!}\left(uh\right)^k\right\}du.\end{aligned}$$

When $p$ is odd, there is no essential change in the proof above except that $q$ is replaced by $p$ in a few places. The asymptotic variance remains the same, but the form of the bias changes.

$$\begin{aligned}& E\left[\frac{1}{\sqrt{nh}}\sum\nolimits_{l=1}^{n} e_1'\overline{S}^{-1}(x_l)\overline{\delta}_{i-1}(x_l)\overline{L}(m(x_l))\right]\\ \approx\ & \sqrt{\frac{n}{h}}\int\left[\int_{\frac{\pi-x}{h}}^{M} K_p^*\left(u\right)\left(\left(\frac{x-\pi}{h}\right)^{i-1}1(x\geq\pi) - (\frac{x-\pi}{h}+u)^{i-1}\right)f(x+uh)du\right.\\ & \left.+\int_{-M}^{\frac{\pi-x}{h}} K_p^*\left(u\right)\left(\frac{x-\pi}{h}\right)^{i-1}1(x\geq\pi)f(x+uh)du\right]\cdot\\ & \frac{e_1'\Gamma^{-1}}{f(x)}\left(\int\delta(u)f(x+uh)\left\{m(x+uh) - m(x) - \sum_{k=1}^{p}\frac{m^{(k)}(x)}{k!}\left(uh\right)^k\right\}du\right)dx\\ =\ & \sqrt{nh}\int\left[\int_{-M}^{M} K_p^*\left(u\right)w^{i-1}1(w\geq 0)\frac{f(\pi+wh+uh)}{f(\pi+wh)}du - \int_{-w}^{M} K_p^*\left(u\right)(w+u)^{i-1}\frac{f(\pi+wh+uh)}{f(\pi+wh)}du\right]\\ & e_1'\Gamma^{-1}\left(\int\delta(u)f(\pi+wh+uh)\left\{m(\pi+wh+uh) - m(\pi+wh) - \sum_{k=1}^{p}\frac{m^{(k)}(\pi+wh)}{k!}\left(uh\right)^k\right\}du\right)dw\\ \approx\ & \sqrt{nh}f(\pi)\int_0^M\left[w^{i-1} - \int_{-w}^{M} K_p^*\left(u\right)(w+u)^{i-1}du\right]\left(\int_{-M}^{M} K_p^*\left(u\right)\frac{m^{(p+1)}(\pi)}{(p+1)!}\left(uh\right)^{p+1}du\right)dw\\ & -\sqrt{nh}f(\pi)\int_{-M}^{0}\left(\int_{-w}^{M} K_p^*\left(u\right)(w+u)^{i-1}du\right)\left(\int_{-M}^{M} K_p^*\left(u\right)\frac{m^{(p+1)}(\pi)}{(p+1)!}\left(uh\right)^{p+1}du\right)dw\\ =\ & \sqrt{nh}h^{p+1}\frac{f(\pi)m^{(p+1)}(\pi)}{(p+1)!}\int_{-M}^{M} K_p^*\left(u\right)u^{p+1}du\left[\int_0^M K_p^*(\overline{\delta}_{i-1}(w+))dw + \int_{-M}^{0} K_p^*(\overline{\delta}_{i-1}(w-))dw\right]\\ =\ & \sqrt{nh}h^{p+1}\frac{f(\pi)m^{(p+1)}(\pi)}{(p+1)!}\int_{-M}^{M} K_p^*\left(u\right)u^{p+1}duQ_p(i).\end{aligned}$$

When $p$ is even, there are some changes in Steps 1 and 3. In Step 1,

(i)

$$
\begin{aligned}
& \frac{1}{\sqrt{nh}}\sum_{l=1}^{n} e_1'\overline{S}^{-1}(x_l)\overline{\delta}_{i-1}(x_l)\cdot e_1'S^{-1}(x_l)\left(S_n(x_l)-S(x_l)\right)S^{-1}(x_l)\left(S_n(x_l)-S(x_l)\right)S_n^{-1}(x_l)\overline{r}(m(x_l)) \\
\approx\ & \frac{1}{\sqrt{nh}}\sum_{l=1}^{n} e_1'\Gamma^{-1}\left(S_n(x_l)-S(x_l)\right)\Gamma^{-1}\left(S_n(x_l)-S(x_l)\right)\Gamma^{-1}\overline{r}(m(x_l)) \\
\approx\ & \frac{1}{\sqrt{nh}}\sum_{l=1}^{n} e_1'\Gamma^{-1}\left(O_p\left(\sqrt{\frac{\ln n}{nh}}\right)+O(h)\Gamma_{(+1)}+O\left(h^2\right)\Gamma_{(+2)}\right)\Gamma^{-1} \\
& \left(O_p\left(\sqrt{\frac{\ln n}{nh}}\right)+O(h)\Gamma_{(+1)}+O\left(h^2\right)\Gamma_{(+2)}\right)\Gamma^{-1}\int\delta(u)u^{p+1}duO(h^{p+1}) \\
=\ & O_p\left(\left(\frac{\ln n}{\sqrt{nh}}+h\sqrt{\ln n}+h^2\sqrt{nh}\right)h^{p+1}\right),
\end{aligned}
$$

where $\Gamma_{(+k)}=\left(\gamma_{i+j-2+k}\right)_{1\le i,j\le p+1}$, $k=1,2$. Note that the dominating terms in the last equality are $h\sqrt{\ln n}+h^2\sqrt{nh}$ instead of $h^2\sqrt{\ln n}+h^3\sqrt{nh}$. This is because although $e_1'\Gamma^{-1}\Gamma_{(+1)}=0$ by the arguments in Ruppert and Wand (1994), $\Gamma^{-1}\Gamma_{(+1)}\Gamma^{-1}\int\delta(u)h^{p+1}du\neq 0$. All other terms are the same as in the case when $p$ is odd. (Note that we need only use $\sup_{x\in N_0}\left(|\widetilde{m}(x)-m(x)|=O_p\left(\sqrt{\frac{\ln n}{nh}}+h^{p+1}\right)\right.$ instead of the stronger result that $\sup_{x\in N_0}\left(|\widetilde{m}(x)-m(x)|=O_p\left(\sqrt{\frac{\ln n}{nh}}+h^{p+2}\right)\right.$. Such a stronger result seems not be proved in the literature although Li and Racine (2007) conjecture it to be true.)

In Step 3, the bias changes.

$$
\begin{aligned}
& E\left[\frac{1}{\sqrt{nh}}\sum\nolimits_{l=1}^{n} e_1'\overline{S}^{-1}(x_l)\overline{\delta}_{i-1}(x_l)\overline{L}(m(x_l))\right] \\
\approx\ & \sqrt{nh}\int\left[\int_{-M}^{M}K_p^*(u)\,w^{i-1}1(w\ge 0)du-\int_{-w}^{M}K_p^*(u)\,(w+u)^{i-1}du\right] \\
& e_1'\Gamma^{-1}\left(\int\delta(u)\left\{f(\pi)+f'(\pi)(w+u)h\right\}\left\{\frac{m^{(p+1)}(\pi)}{(p+1)!}(uh)^{p+1}+\frac{m^{(p+2)}(\pi)}{(p+2)!}(uh)^{p+2}\right\}du\right)dw \\
=\ & \sqrt{nh}h^{p+2}\int\left[\int_{-M}^{M}K_p^*(u)\,w^{i-1}1(w\ge 0)du-\int_{-w}^{M}K_p^*(u)\,(w+u)^{i-1}du\right] \\
& \left(\int_{-M}^{M}K_p^*(u)\left\{f(\pi)\frac{m^{(p+2)}(\pi)}{(p+2)!}u^{p+2}+f'(\pi)\frac{m^{(p+1)}(\pi)}{(p+1)!}(w+u)\,u^{p+1}\right\}du\right)dw \\
=\ & \sqrt{nh}h^{p+2}\left[\int_{0}^{M}K_p^*(\overline{\delta}_{i-1}(w+))dw+\int_{-M}^{0}K_p^*(\overline{\delta}_{i-1}(w-))dw\right] \\
& \left(\int_{-M}^{M}K_p^*(u)\,u^{p+2}du\right)\left[f(\pi)\frac{m^{(p+2)}(\pi)}{(p+2)!}+f'(\pi)\frac{m^{(p+1)}(\pi)}{(p+1)!}\right].
\end{aligned}
$$

Note that $e_1'S^{-1}(x)\left(\overline{S}(x)-S(x)\right)S^{-1}(x)\,\overline{r}(m(x))$ in $\overline{L}(m(x))$ does not contribute to the bias regardless of whether $p$ is odd or even since it only contributes a higher-order term in both cases.

### Denominator

We get the asymptotic limit of $\frac{1}{nh}\underline{Z}^{d'}(\pi)\underline{Z}^{d}(\pi)$ here. Note that the $(i,j)$ term of $\frac{1}{nh}\underline{Z}^{d'}(\pi)\underline{Z}^{d}(\pi)$ is

$$\frac{1}{nh}\sum\nolimits_{l=1}^{n} e_1' S_n^{-1}(x_l)\delta_{ni-1}(x_l) e_1' S_n^{-1}(x_l)\delta_{nj-1}(x_l),$$

which, by a similar argument as in the numerator, is asymptotically equivalent to

$$\frac{1}{nh}\sum\nolimits_{l=1}^{n} e_1'\overline{S}^{-1}(x_l)\overline{\delta}_{i-1}(x_l)\cdot e_1'\overline{S}^{-1}\overline{\delta}_{j-1}(x_l). \tag{14}$$

It is easy, although tedious, to show that its variance converges to zero. By Markov's inequality, (14) converges in probability to

$$\begin{aligned}
& \frac{1}{h}E\left[e_1'\overline{S}^{-1}(x_l)\overline{\delta}_{i-1}(x_l)\cdot e_1'\overline{S}^{-1}(x_l)\overline{\delta}_{j-1}(x_l)\right] \\
\approx\ & f(\pi)\int\left[w^{i-1}1(w\geq 0)-\int_{-w}^{M}K_p^*(u)(w+u)^{i-1}du\right]\left[\int_{-M}^{M}K_p^*(u)w^{j-1}1(w\geq 0)du-\int_{-w}^{M}K_p^*(u)(w+u)^{j-1}du\right]dw \\
=\ & f(\pi)\left[\int_0^M K_p^*(\overline{\delta}_{i-1}(w+))K_p^*(\overline{\delta}_{j-1}(w+))dw+\int_{-M}^{0}K_p^*(\overline{\delta}_{i-1}(w-))K_p^*(\overline{\delta}_{j-1}(w-))dw\right] \\
=\ & f(\pi)N_p(i,j).
\end{aligned}$$

By continuity of the matrix inversion,

$$e_1'\left(\frac{1}{nh}\underline{Z}^{d'}(\pi)\underline{Z}^{d}(\pi)\right)^{-1}\xrightarrow{p} f(\pi)^{-1}e_1'N_p^{-1}.$$

Based on the analysis on the numerator and denominator, the results in Theorem 1 follow.

## Appendix B: Lemmas

**Lemma 1** *$\underline{X}_l^d(\pi)=0$ for $|x_l-\pi|>Mh$, $l=1,\cdots,n$.*

**Proof.** From (2.4) of Fan et al (1997),

$$\sum\nolimits_{j=1}^{n}(x_j-x)^{\nu}W_j^n(x)=\delta_{0,\nu},\ 0\leq\nu\leq p,$$

where $\delta_{0,\nu}$ is equal to 1 if $\nu=0$, and equal to 0 otherwise. Based on this result, for any $x_l$ such that $|x_l-\pi|>Mh$,

$$(x_l-x)^{i-1}1(x_l>\pi)-P_{x_l}^n\left(X^{i-1d}(\pi)\right)=0,\ 1\leq i\leq q+1\leq p+1.$$

For example, if $x-\pi>Mh$, for $i=1$,

$$(x-\pi)^{i-1}1(x>\pi)-P_x^n\left(X^{i-1d}(\pi)\right)=1-\sum\nolimits_{j=1}^{n}W_j^n(x)=0.$$

Note that the indicator function $1(x_j > \pi)$ in $X^{i-1d}(\pi)$ does not play any role here. For $i = 2$,

$$\begin{aligned} & (x-\pi) - \sum_{j=1}^{n} W_j^n(x)(x_j - \pi) \\ = \; & (x-\pi) - \sum_{j=1}^{n} W_j^n(x)(x_j - x + x - \pi) \\ = \; & (x-\pi) - (x-\pi)\sum_{j=1}^{n} W_j^n(x) \\ = \; & 0. \end{aligned}$$

By induction, we can show all other terms are zero as long as $q \leq p$. ■

**Lemma 2** *Suppose $m(x) = E[\omega_i | x_i = x]$ is $q$ times continuously differentiable with $q \leq p$, then*

$$P_x^n(\omega) - m(x) = e_1' S^{-1}(x)\overline{r}(x) + P_x^L(\omega) + P_x^Q(\omega),$$

*where $S(x) = \Gamma f(x)$,*

$$\overline{r}(x) = \int \delta(u) f(x+uh)\left(m(x+uh) - m(x) - \sum_{k=1}^{q} \frac{m^{(k)}(x)}{k!}(uh)^k\right) du,$$

*and $P_x^L(\omega)$ and $P_x^Q(\omega)$ are defined in (15). If $q > p$, then the $q$ in $\overline{r}(x)$ is changed to $p$, and $P_x^L(\omega)$ and $P_x^Q(\omega)$ are adjusted correspondingly.*

**Proof.** Define $\omega_i = m(x_i) + e_i$, then

$$\begin{aligned} & P_x^n(\omega) - m(x) \\ = \; & e_1'\left(Z(x)' K_h(x) Z(x)\right)^{-1} Z(x)' K_h(x)(\boldsymbol{\omega} - m(x)), \\ = \; & e_1'\left(\frac{1}{n}\sum_{j=1}^{n} Z_j(x) Z_j'(x) k_h(x_j - x)\right)^{-1} \frac{1}{nh}\sum_{j=1}^{n} Z_j(x) k_h(x_j - x)(E[\omega_j | x_j] - m(x) + e_j) \\ = \; & e_1'\left(\frac{1}{n}\sum_{j=1}^{n} Z_j(x) Z_j'(x) k_h(x_j - x)\right)^{-1} \frac{1}{n}\sum_{j=1}^{n} Z_j(x) k_h(x_j - x)\left\{E[\omega_j | x_j] - m(x) - \sum_{k=1}^{q} \frac{m^{(k)}(x)}{k!}(x_j - x)^k + e_j\right\} \\ \equiv \; & e_1' S_n^{-1}(x)\widetilde{r}(x). \end{aligned}$$

Linearize the denominator at its limit $S(x)$ and the numerator at its mean $\overline{r}(x)$. Note that $\overline{r}(x)$ converges to 0 when $h$ goes to zero, so we can not linearize at the limit of the numerator.

$$\begin{aligned} & e_1' S_n^{-1}(x)\widetilde{r}(x) - e_1' S^{-1}(x)\overline{r}(x) \qquad (15) \\ = \; & -e_1' S^{-1}(x)(S_n(x) - S(x)) S^{-1}(x)\overline{r}(x) + e_1' S^{-1}(x)(\widetilde{r}(x) - \overline{r}(x)) \text{ (linear terms)} \\ & +e_1' S^{-1}(x)(S_n(x) - S(x)) S^{-1}(x)(S_n(x) - S(x)) S_n^{-1}(x)\overline{r}(x) \\ & -e_1' S^{-1}(x)(S_n(x) - S(x)) S_n^{-1}(x)(\widetilde{r}(x) - \overline{r}(x)) \text{ (quadratic terms)} \\ \equiv \; & P_x^L(\omega) + P_x^Q(\omega). \end{aligned}$$

■

**Lemma 3** *If $\sup_{x \in N} E\left[|\varepsilon|^{2+\zeta} | x\right] < \infty$ for some $\zeta > 0$, $n^{\zeta/(2+\zeta)} h / \ln n \to \infty$, and $m(x) \in C^{(q+1)}(N)$, then for $N_0 = [\pi - Mh, \pi + Mh]$,*

**(i)** $\sup_{x\in N_0}\left|S_n(x)-\overline{S}(x)\right| = O_p\left(\sqrt{\frac{\ln n}{nh}}\right)$, $\sup_{x\in N_0}\left|S_n^{-1}(x)-\overline{S}^{-1}(x)\right| = O_p\left(\sqrt{\frac{\ln n}{nh}}\right)$, $\sup_{x\in N_0}\left|\overline{S}(x)-S(x)\right| = O(h)$.

**(ii)** $\sup_{x\in N_0}|\widetilde{r}(m(x))-\overline{r}(m(x))| = O_p\left(\sqrt{\frac{\ln n}{nh}}\right)$, $\sup_{x\in N_0}|\widetilde{r}(\varepsilon(x))| = O_p\left(\sqrt{\frac{\ln n}{nh}}\right)$.

**(iii)** $\sup_{x\in N_0}|\widetilde{m}(x)-m(x)| = O_p\left(\sqrt{\frac{\ln n}{nh}}+h^{q+1}\right)$,.

**(iv)** $\sup_{x\in N_0}\left|\delta_{ni-1}(x)-\overline{\delta}_{i-1}(x)\right| = O_p\left(\sqrt{\frac{\ln n}{nh}}\right)$.

*Here, the norm $|\cdot|$ for a vector or matrix is the maximum absolute value among all elements.*

**Proof.** The proof follows from Lemma B.1 and B.2 of Newey (1994). The basic proof techniques are truncation and Bernstein's inequality. Since the proof is very standard, omitted here for simplicity. See also Masry (1996) for more details. We only discuss a little about $\sup_{x\in N_0}\left|S_n^{-1}(x)-\overline{S}^{-1}(x)\right|$. Note that

$$\begin{aligned}\sup_{x\in N_0}\left|S_n^{-1}(x)-\overline{S}^{-1}(x)\right| &\leq \sup_{x\in N_0}\left|\overline{S}^{-1}(x)\right|\sup_{x\in N_0}\left|S_n(x)-\overline{S}(x)\right|\sup_{x\in N_0}\left|S_n^{-1}(x)\right| \\ &= O(1)O_p\left(\sqrt{\frac{\ln n}{nh}}\right)O_p(1) = O_p\left(\sqrt{\frac{\ln n}{nh}}\right).\end{aligned}$$

■

**Lemma 4** $\frac{1}{nh}\sum_{l=1}^{n}|\varepsilon_l|\,1(\pi-Mh\leq x_l\leq\pi+Mh) = O_p(1)$, $\frac{1}{nh}\sum_{l=1}^{n}1(\pi-Mh\leq x_l\leq\pi+Mh) = O_p(1)$.

**Proof.** These are intermediate results in Porter (2003), and can be proved by Markov's inequality. ■