Original Research Article

Jump-Preserving Estimation and Jump Detection for Nonparametric Model with Missing Data in the Covariate

Abstract

Nonparametric regression analysis has broad applications. In some cases, the regression function with jumps (i.e., the regression curve is discontinuous) seems to be more appropriate to describe the related phenomena. Existing a number of methods for estimating discontinuous curve, most of which are based on the data is complete, which is unrealistic in many practical situations. In this paper, we consider estimating discontinuous nonparametric model with missing covariate data. Based on inverse selection probability weighted and jump-preserving techniques, a jump-preserving estimation procedure is proposed. The proposed method is capable of automatically accommodating possible jumps in the nonparametric function, without the requirement of prior knowledge regarding the number and locations of jump points. The proposed estimator for the discontinuous regression function is shown to be oracally efficient in the sense that it is uniformly indistinguishable from that when the selection probabilities are known. Furthermore, it is proved that the fitted curve by this procedure is consistent in the entire design space. Numerical simulation also indicates the performance of finite sample of this method is efficient and reliable.

Keywords: *Nonparametric model; Local linear kernel smoothing; Jump-preserving estimation; Inverse probability weighted; Missing data*

1 Introduction

A basic nonparametric regression model for the dependence of the scalar response variable Y and its covariate X has the form,

$$Y = g(X) + \varepsilon, \tag{1}$$

where $g(\cdot)$ is an unknown measurable function of X to be estimated and model error ε has mean zero and finite variance σ^2. For the sake of simplicity, it is assumed that X come from a continuous distribution with density $f_X(x)$ supported on the bounded interval $[0, 1]$, it is independent of ε.

Nonparametric regression is an important branch in statistics, it has been well established as a useful data analytic tool. See, for example, the monographs Fan and Gijbels (1996) for a large variety of interesting real data examples where applications of such methods have yielded analyses essentially unobtainable by other techniques. A large body of literature exists on regression function estimation, such as kernel estimation(see Fan and Gijbels (1996) and Claeskens and Van Keilegom (2003)), local polynomial regression(see Xiao et al. (2003) and El Ghouch and Genton (2009)), spline smoothing (see Huang (2003) and Wang and Yang (2009)), and so on.

The common assumption of above works is that the function $g(\cdot)$ is continuous. In many applications, however, it may appear that a regression function is smooth except at an unknown finite number of points where jump discontinuities may occur. For instance, in image processing, the intensity function of a digital image can be regarded as a piecewise continuous regression surface, and jump regression analysis provides a natural framework for image analysis (Qiu (2005)). In quality control, a jump in a quality index of a product indicates that the production line could be out of control (Hawkins and Olwell (1998)). In oceanography, the sea-level pressure in Bombay India has been found to have experienced an abrupt change in the early 1960s (Qiu (2003)). In finance, possible jumps in the exchange rate between Korean won and U.S. dollar during December 1997 have been identified by Joo and Qiu (2009). Therefore, the research on regression model with jumps is very necessary.

In recent decades, there are several researches to fit curves with jump points. McDonald and Owen (1986) introduced a family of smoothing algorithm based on three local ordinary least-squares estimations of the regression function, including the observations on the left, right, and both sides of a given point. Then, the fitted value of a given point is obtained as a weighted average of these three estimations, with the weights depending on the goodness-of-fit values of them. Afterward, Hall and Titterington (1992) proposed an alternative procedure based on the detection of discontinuities by comparing three smooth fits at any given points. As usual, the regression curves can be fitted by conventional smoothing methods in continuous intervals separated by these detected jump points.

Besides the piecewise estimation methods above, some scholars proposed the local polynomial and kernel type methods for jump detection and jump-preserving estimation. Qiu (2003) proposed a jump-preserving curve fitting procedure based on the local linear kernel estimation. For each point, two one-sided local linear estimates are considered, and based on the comparison of the weighted residual mean squares of the two one-sided fits, the curve estimate at each point is obtained by one of the two estimates or their average. Furthermore, Gijbels et al. (2007) proposed a compromise local linear jump-preserving method. The resulting estimators preserve the jumps well and also give smooth estimates of the continuity parts of the curve. But a threshold parameter is introduced in the procedure, and the choice for it may increase the computation burden. For this reason, Qiu (2009) presented another method to distinguish smooth regions and discontinuous regions based on the fact that the variance of the two-sided estimator is about twice as that of the one-sided estimator. The method does not need to compute the threshold parameter, hence is easier to implement. Xia and Qiu (2015) suggested a jump information criterion to estimate the discontinuous curve when the number of jumps is unknown. By minimizing the criterion function which consists of a term measuring the fidelity of the estimated regression curve to the observed data and a penalty with respect to the number of jumps and the jump magnitudes, the number of jumps is obtained. Kang et al. (2021) proposed a jump-preserving backfitting procedure for jump additive model. It extends existing jump regression methods to problems where multiple predictors need to be considered. Additional works have Yang and Song (2014), Zhao et al. (2016), Han et al. (2020) and Wang et al. (2022), among others.

All the above related works for nonparametric regression are for fully observed data. However, in many applications, it is inevitable that some information maybe lost in the collection process of a large amount of data due to uncontrollable factors. For instance, in clinical trials, missing values exist for a variety of reasons, such as patient refusal to continue in the study, treatment failure or success, adverse events, patient moves, etc. In social surveys, respondents may refuse to answer questions about their income. In industrial experiments, some experimental results are not recorded completely due to various reasons. Such data is usually called by missing data, see Kim and Shao (2013) and Little and Rubin (2019) for an introduction on missing data and many examples.

In the presence of missing data, the standard inference procedures cannot be applied directly. The simplest approach to deal with missing data is to remove those observations with incomplete data, then perform a regression based or likelihood based analysis with the remaining complete data. This method is known as complete case analysis. However, it is well known that the complete case analysis can be

biased when the data is not missing completely at random (MCAR) (see Little and Rubin (2019)) and generally gives highly inefficient estimates. Thus to increase efficiency and reduce the bias, it is important to develop methods to deal with missing data.

A series of efforts have been made to deal with missing data. One method is to impute a plausible value for each missing datum and then analyze the results as if they are complete. In regression problems, commonly used imputation approaches include linear regression imputation (see Healy and Westmacott (1956)), nonparametric kernel regression imputation (see Wang and Rao (2002b)), semiparametric regression imputation (see Wang et al. (2004)), among others. There has been considerable interest in the statistical literature on analysis of missing data using the empirical likelihood method, see, for example, Wang and Rao (2002a), Liang et al. (2007) and Stute et al. (2007), among others. These approaches impute the missing data by a kernel regression function of the observed data and then use empirical likelihood to constructing confidence intervals from the observed and the imputed data. The inverse probability weighted (IPW) method (see Horvitz and Thompson (1952)) is also popular method to handle missing data, this method assigns a weight to each complete observation by the inverse probability of it being completely observed. Wooldridge (2007) discussed the inverse probability weighted estimation for general missing data problems. Wang (2009) consider the partial linear model with the covariables missing at random using the inverse probability weighted approach. Seaman and White (2013) reviewed inverse probability weighted for the missing data analysis.

Nonparametric regression method to deal with missing data was discussed relatively less. For missing covariates, Wang et al. (1998) used IPW local linear regression in generalised linear model; Liang et al. (2004) considered a nonparametric estimator in a partially linear model. For missing outcomes, Wang et al. (2010) developed a doubly robust local linear estimator and Sun et al. (2020) generalised it to the multiple robustness; Chen et al. (2006) constructed a few local quasi-likelihood estimators; estimating equations with nonparametric inputted values were developed by Zhou et al. (2008), and Wang et al. (2019) considered the case where the covariate is functional. Efromovich generalised the orthogonal series estimator in the cases of missing covariates (Efromovich (2011a)) and missing outcomes (Efromovich (2011b)).

For model (1), when the regression is discontinuous, and the data is missing at the same time, none of the individual methods described above are suitable for estimating the model. Li et al. (2023) studied the discontinuous nonparametric regression curve fitting when response is missing. Naturally, we are interested in the case of missing covariate. In order to show X is incomplete, denote δ as a missing indicator, that is, $\delta = 1$ means X is completely observed and $\delta = 0$ otherwise. To deal with missing covariate, we assume that X is missing at random (MAR) in the sense that

$$\pi(Y) := P(\delta = 1|X, Y) = P(\delta = 1|Y). \tag{2}$$

The MAR assumption implies that δ and X are conditionally independent given Y, that is, $P(\delta = 1|X, Y) = P(\delta = 1|Y)$. MAR is a common assumption for handling missing data and such assumption is also reasonable in many applications, see Little and Rubin (2019).

This paper focuses on the estimation of discontinuous regression curve with missing covariate. By the inverse probability weighted techniques and local linear kernel smoothing, we construct the jump-preserving estimation. The procedure is capable of adapting to both continuous intervals and neighbourhoods of jumps of the nonparametric function needn't for prior estimation of the the number and locations of jump points. Indeed, the resulting estimator represents a compromise between local linear smoothing and jump-preserving, which is implemented by a threshold. For our proposed estimators, it is shown to be oracally efficient in the sense that the estimator with estimated selection probabilities under a correctly specified model is uniformly as efficient as that with true selection probabilities. Besides, a brief discussion is also held regarding the detection of jump points, along with practical selection of procedure parametrics. Moreover, the asymptotical properties of proposed estimators are presented. Numerical simulation indicates the performance of finite sample of this method is efficient.

The rest of this paper is organized as follows. Section 2 first recalls the jump-preserving method with complete data, then presents the detailed procedure and main theoretical results for the proposed method. Some numerical studies are conducted to evaluate the finite sample properties of the proposed estimators in Section 3. A brief conclusion is given in Section 4. Technical proofs are presented in the Appendix.

2 Methodology and main results

In this section, we will consider the curve fitting for the nonparametric regression model (1) with unknown jumps and missing data in covariate. For jump structure, suppose that the number of jump points is J, and let $s_j \in (0,1)$ denotes the jth jump point of $g(\cdot)$ with jump magnitudes d_j, where $j = 1, 2, \ldots, J$. Without loss of generality, we assume that $g(\cdot)$ is right-continuous at each jump point. The number of jump points J, the jump locations s_j's and the jump magnitudes d_j's are all unknown.

In Subsection 2.1, we firstly review the local linear jump regression curve fitting method with complete data. Then we extend this method into the context of missing data by virtual of inverse probability weighting method.

2.1 Review the jump-preserving method

Suppose that $\{(X_i, Y_i), 1 \le i \le n\}$ is an independent and identically distributed (i.i.d.) random sample observed fully from (1). When $g(\cdot)$ is a continuous function, Fan and Gijbels (1996) proposed local linear smoothing method to estimate $g(\cdot)$. Specifically, to estimate the regression function $g(\cdot)$ at a given point $x \in [0,1]$, one can approximate

$$g(u) \approx g(x) + g'(x)(u - x) = a + b(u - x),$$

for u in a small neighbourhood of the given point x, where $a = g(x)$ and $b = g'(x)$. Then, the local parameter (a, b) is estimated by minimising the following weighted least-squares function:

$$\sum_{i=1}^{n} \{Y_i - a - b(X_i - x)\}^2 K_h(X_i - x), \tag{3}$$

where $K_h(t) = h^{-1}K(t/h)$, a rescaled kernel function of $K(t)$ with a bandwidth h. Commonly $K(\cdot)$ is chosen to be a bounded symmetric probability density function (conventional or center kernel) with support $[-\tau, \tau]$. The bandwidth $h = h(n) > 0$ is a sequence of positive constant that converge to zero with sample size n approaching infinity. We will suppress the dependence of bandwidth h on n in what follows.

The solution of (3) for a is defined as the conventional local linear kernel estimator of $g(\cdot)$. This local linear procedure has been popular in the literature due to its simplicity of computation and nice asymptotic properties. However, this method requires the continuity of curve function, and it is known that the fitted function based on conventional local linear kernel methods is not statistically consistent at jump positions where $g(\cdot)$ has jumps. To deal with this problem, some jump-preserving estimation methods were proposed. Now we briefly review techniques given by Qiu (2003) and Gijbels et al. (2007).

For fixed $x \in [0,1]$, the following three local linear estimators are defined by

$$\arg\min_{a,b} \sum_{i=1}^{n} \{Y_i - a - b(X_i - x)\}^2 K_d\left(\frac{X_i - x}{h}\right), \tag{4}$$

where $K_c(x) = K(x)$

$$K_l(x) = \begin{cases} K(x), & \text{if } x \in [-\tau, 0), \\ 0, & \text{otherwise;} \end{cases} \quad \text{and} \quad K_r(x) = \begin{cases} K(x), & \text{if } x \in [0, \tau], \\ 0, & \text{otherwise.} \end{cases}$$

The subscripts "l", "c"and "r" in notations $\{K_l, K_c, K_r\}$ represent "left", "centre" and "right", respectively, which are also used in other notation defined below.

Let $\{\hat{a}_d(x), \hat{b}_d(x), d = c, r, l\}$ denote the solutions of (4). Obviously the estimators $\hat{a}_c(x)$ are the usual local linear estimators of $g(x)$, based on data in the neighbourhood $[x - \tau h, x + \tau h]$ of x, and $\hat{a}_l(x)$ and $\hat{a}_r(x)$ are constructed from observations in the left-sided interval $[x - \tau h, x)$ and right-sided interval $[x, x + \tau h]$, respectively. From Proposition 2.3 in Gijbels et al. (2007), the three estimators are consistent in mean square sense and have the same rate of convergence in continuity regions of $g(\cdot)$. From Proposition 2.4 in Gijbels et al. (2007), however, it can be found that only $\hat{a}_l(x)$ is consistent, nevertheless, $\hat{a}_c(x)$ and $\hat{a}_r(x)$ are not consistent at any point in the neighbourhood $[s_j - \tau h, s_j)$. A similar discussion can be given for points on the right-side interval of the jump point s_j. That is, when there is no jump in $[x-\tau h, x+\tau h]$, all of them estimate $g(\cdot)$ well. In the case when x itself is not a jump point but a jump point exists in its neighborhood $[x-\tau h, x+\tau h]$, only one of $\hat{a}_l(x)$ and $\hat{a}_r(x)$ provides a good estimator of $g(\cdot)$. Therefore, it need to choose one from three estimators as an estimator of $g(\cdot)$ in such case.

Qiu (2003) suggested the following jump-preserving estimate of $g(\cdot)$

$$\hat{g}_Q(x) = \hat{a}_l(x) I^*(\mathrm{RSS}_r(x) - \mathrm{RSS}_l(x)) + \hat{a}_r(x) I^*(\mathrm{RSS}_l(x) - \mathrm{RSS}_r(x)),$$

where $I^*(t)$ is defined by $I^*(t) = 1$ if $t > 0$, $1/2$ if $t = 0$ and 0 if $t < 0$. And $\mathrm{RSS}_l(x)$ and $\mathrm{RSS}_r(x)$ are the weighted residual sums of squares (RSS) with respect to observations in $[x - \tau h, x)$ and $[x, x + \tau h]$, respectively. That is

$$\begin{aligned}
\mathrm{RSS}_l(x) &= \sum_{X_i < x} \left\{ Y_i - \hat{a}_l(x) - \hat{b}_l(x)(X_i - x) \right\}^2 K\left(\frac{X_i - x}{h} \right), \\
\mathrm{RSS}_r(x) &= \sum_{X_i \geq x} \left\{ Y_i - \hat{a}_r(x) - \hat{b}_r(x)(X_i - x) \right\}^2 K\left(\frac{X_i - x}{h} \right).
\end{aligned}$$

Basically, $\hat{g}_Q(x)$ is defined by one of $\hat{a}_l(x)$ and $\hat{a}_r(x)$ with the smaller RSS value. Qiu (2003) proved that $\hat{g}_Q(x)$ is a consistent estimator of $g(x)$ in the entire design interval.

In practice, it appears that $\hat{g}_Q(x)$ preserves jumps well, but is quite noisy in continuity regions of $g(\cdot)$, due to the fact that only one-sided (left- or right-sided) observations are used in its construction. Consequently, by combining all these considerations, Gijbels et al. (2007) introduced the conventional estimator $\hat{a}_c(x)$, and proposed the following jump-preserving estimate of $g(\cdot)$

$$\hat{g}_G(x) = \begin{cases}
\hat{a}_c(x), & \text{if } \mathrm{diff}(x) \leq \lambda, \\
\hat{a}_l(x), & \text{if } \mathrm{diff}(x) > \lambda \text{ and } \mathrm{WRMS}_l(x) < \mathrm{WRMS}_r(x), \\
\hat{a}_r(x), & \text{if } \mathrm{diff}(x) > \lambda \text{ and } \mathrm{WRMS}_l(x) > \mathrm{WRMS}_r(x), \\
(\hat{a}_l(x) + \hat{a}_r(x))/2, & \text{if } \mathrm{diff}(x) > \lambda \text{ and } \mathrm{WRMS}_l(x) = \mathrm{WRMS}_r(x),
\end{cases} \tag{5}$$

where λ is a threshold and

$$\mathrm{diff}(x) = \max\{\mathrm{WRMS}_c(x) - \mathrm{WRMS}_l(x), \mathrm{WRMS}_c(x) - \mathrm{WRMS}_r(x)\}.$$

In (5), the weighted residual mean squares (WRMSs) are defined by

$$\mathrm{WRMS}_d(x) = \frac{\sum_{i=1}^n \left[Y_i - \hat{a}_d(x) - \hat{b}_d(x)(X_i - x) \right]^2 K_d\left(\frac{X_i - x}{h} \right)}{\sum_{i=1}^n K_d\left(\frac{X_i - x}{h} \right)}.$$

to evaluate the quality of the three local linear fits. Gijbels et al. (2007) proved the (uniform) strong consistency of $\hat{g}_G(x)$.

2.2 Estimation when $\pi(Y)$ is known

Suppose that there are n i.i.d. observations $\{(X_i, Y_i, \delta_i), i = 1, 2, \ldots, n\}$, where $\delta_i = 1$ if X_i is observed and $\delta_i = 0$ otherwise. When covariates are MAR, the complete case analysis in (4) by using only fully observed (X_i, Y_i) can result in a biased estimator for $g(\cdot)$. Let $\pi_i = \pi(Y_i) = P(\delta_i = 1|Y_i)$ is the selection probability by our MAR assumption. In this Subsection, we first assume that the missing data probability $\pi(Y)$ is known, and will consider the case in which $\pi(Y)$ is unknown in next Subsection.

To estimate the latent discontinuous function $g(\cdot)$ incorporating missing data, we combine jump-preserving method in Gijbels et al. (2007) and the inverse probability weighted techniques (see Horvitz and Thompson (1952)). Specifically, for fixed $x \in [0, 1]$, the following three inverse probability weighted local linear estimators are proposed

$$\arg\min_{a,b} \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} \{Y_i - a - b(X_i - x)\}^2 K_d\left(\frac{X_i - x}{h}\right), \tag{6}$$

for $d = c, l, r$. Since the variables are subject to missingness, only fully observed cases $\delta_i = 1$ contribute to the objective function (6), and the selection bias is adjusted by inverse of the conditional probability of being a complete case.

The solutions to a_d and b_d of the minimization problem (6) are denoted as $\hat{a}_d(x)$ and $\hat{b}_d(x)$, $d = c, l, r$, respectively. By some routine algebraic manipulations, $\hat{a}_d(x)$ have nice and simple expressions:

$$\hat{a}_d(x) = \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} K_d\left(\frac{X_i - x}{h}\right) \frac{S_{2,d} - S_{1,d}(X_i - x)}{S_{0,d}S_{2,d} - S_{1,d}^2} Y_i,$$

where $S_{j,d} = \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} K_d\left(\frac{X_i - x}{h}\right)(X_i - x)^j$ for $j = 0, 1, 2$, $d = c, l, r$.

It is easy to see that $\hat{a}_l(x)$ and $\hat{a}_r(x)$ are actually local linear kernel estimators of $g(x)$ constructed from observations in the left-sided neighborhood $[x - \tau h, x)$ and the right-sided neighborhood $[x, x + \tau h]$, respectively. However, the estimators $\hat{a}_c(x)$ are the usual local linear estimator of $g(x)$, based on data in the neighbourhood $[x - \tau h, x + \tau h]$ of x. Similarly with the approach in Gijbels et al. (2007), we propose the following compromising estimator

$$\hat{g}(x) = \begin{cases} \hat{a}_c(x), & \text{if } \mathrm{diff}(x) \le \lambda, \\ \hat{a}_l(x), & \text{if } \mathrm{diff}(x) > \lambda \text{ and } \mathrm{WRMS}_l(x) < \mathrm{WRMS}_r(x), \\ \hat{a}_r(x), & \text{if } \mathrm{diff}(x) > \lambda \text{ and } \mathrm{WRMS}_l(x) > \mathrm{WRMS}_r(x), \\ (\hat{a}_l(x) + \hat{a}_r(x))/2, & \text{if } \mathrm{diff}(x) > \lambda \text{ and } \mathrm{WRMS}_l(x) = \mathrm{WRMS}_r(x). \end{cases} \tag{7}$$

In the missing data case, based on inverse probability weighting, we introduce the modified weighted residual mean squares (WRMS) $\mathrm{WRMS}_d(x)$, defined by

$$\mathrm{WRMS}_d(x) = \frac{\sum_{i=1}^{n} \frac{\delta_i}{\pi_i} \left[Y_i - \hat{a}_d(x) - \hat{b}_d(x)(X_i - x)\right]^2 K_d\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} \frac{\delta_i}{\pi_i} K_d\left(\frac{X_i - x}{h}\right)},$$

for $d = c, l, r$. In (7),

$$\mathrm{diff}(x) = \max\{\mathrm{WRMS}_c(x) - \mathrm{WRMS}_l(x), \mathrm{WRMS}_c(x) - \mathrm{WRMS}_r(x)\},$$

and λ is a suitably chosen threshold, such that away from the irregularities the two-sided estimator is chosen and the appropriate one-sided estimator is chosen close to them.

Obviously, $\mathrm{diff}(x)$ is a natural jump detection criterion. If x is a jump point, then $\mathrm{diff}(x)$ would be relatively large. By (7), thus, when x is far away from any jump points, $g(x)$ would be estimated by the conventional (or centered) kernel local linear fitting. It would still be estimated by one of the one-sided estimates around the jump points.

Next, we establish the asymptotic properties of the proposed estimators $\hat{g}(x)$. Their proofs are given in the Appendix. To proceed, we introduce some notations. Let $\mu_{k,d} = \int t^k K_d(t)dt$ for $d = c, l, r$. Furthermore, the support $[0,1]$ of $g(\cdot)$ can be divided into two regions depending on whether $g(\cdot)$ is continuous: (i) the neighborhoods of jump points $D_2 = \bigcup_{j=1}^{J}(s_j-\tau h, s_j+\tau h)$; (ii) the continuous regions $D_1 = [0,1]\backslash D_2$. The region D_2 can be further separated by two parts $D_{2,l} = \bigcup_{j=1}^{J}(s_j - \tau h, s_j)$ and $D_{2,r} = \bigcup_{j=1}^{J}(s_j, s_j + \tau h)$ to represent the left and right neighborhood of the jump points, respectively. The following technical assumptions are imposed.

(A1) The error ε has mean zero and finite variance σ^2, and $E(\varepsilon^4) < \infty$. Moreover, $\int \varepsilon^2 f_{X,\varepsilon|\delta=1}(x,\varepsilon)\,\mathrm{d}\varepsilon$ has a positive lower bound for all $x \in [0,1]$, where $f_{X,\varepsilon|\delta=1}(x,\varepsilon)$ is the joint density function of (X,ε) given $\delta = 1$.

(A2) Let $f_X(x)$ be the density function of X. $f_X(x)$ is twice differentiable for $x \in [0,1]$. We only require one-sided twice differentiability when $x = 0$ or $x = 1$. That is, for $m = 0, 1$, we assume

$$\lim_{x\to 0^+} \frac{f_X^{(m)}(x) - f_X^{(m)}(0)}{x} \quad \text{and} \quad \lim_{x\to 1^-} \frac{f_X^{(m)}(x) - f_X^{(m)}(1)}{x-1}$$

exist, $\sup |f_X^{(m)}(x)| < \infty$ for $m = 0, 1, 2$.

(A3) Suppose that $g(\cdot)$ is second-order differentiable and $g''(\cdot)$ is uniformly bounded on $[0,1]$ except the jump points $\{s_j, j = 1, \ldots, J\}$ at which $g(\cdot)$ has left and right bounded second-order derivatives.

(A4) $K(\cdot)$ is a symmetric probabiltiy density function with a bounded support, and is uniformly Lipschitz continuous.

(A5) $h \to 0$ and $nh^3 \to \infty$ as $n \to \infty$.

(A6) $\inf_{1\le i\le n}\{\pi(Y_i)\} \geqslant C > 0$ with probability one for some constant C.

The aforementioned assumptions are not the weakest possible assumptions, but they are imposed to facilitate the proofs. Conditions (A1)-(A3) ensure the rationality of local linear approximation; condition (A4) and (A5) are the conventional condition in the kernel estimation method; condition (A6) ensures the effectiveness of inverse probability weighted.

Theorem 1. *Under the regularity assumptions (A1)–(A6), the mean squared errors (MSE) of the three estimators of the function $g(\cdot)$ are as follows:*

(i) For any $x \in D_1$,

$$\mathrm{MSE}(\hat{a}_d(x)) = \left[\frac{1}{2}h^2 g''(x)B_d\right]^2 + \frac{1}{nhf_X^2(x)}P(\delta_1 = 1)V_d S(x) + o\left(h^4 + \frac{1}{nh}\right), \quad d = c, l, r,$$

(ii) For any $x \in D_{2,l}$, that is, $x = s_j + uh$ with $u \in (-\tau, 0)$, we have

$$\mathrm{MSE}(\hat{a}_l(x)) = \left[\frac{1}{2}h^2 g''(s_j-)B_l\right]^2 + \frac{1}{nhf_X^2(x)}P(\delta_1 = 1)V_l S(x) + o\left(h^4 + \frac{1}{nh}\right),$$

$$\mathrm{MSE}(\hat{a}_r(x)) = \left[d_j \int_{|u|}^{\tau} K_r(t)\frac{\mu_{2,r} - \mu_{1,r}t}{\mu_{0,r}\mu_{2,r} - \mu_{1,r}^2}\,\mathrm{d}t\right]^2 + \frac{1}{nhf_X^2(x)}P(\delta_1 = 1)V_r S(x) + o\left(\frac{1}{nh}\right),$$

$$\mathrm{MSE}(\hat{a}_c(x)) = \left[d_j \int_{|u|}^{\tau} K_c(t)\,\mathrm{d}t\right]^2 + \frac{1}{nhf_X^2(x)}P(\delta_1 = 1)V_c S(x) + o\left(\frac{1}{nh}\right).$$

(iii) For any $x \in D_{2,r}$, that is, $x = s_j + uh$ with $u \in [0, \tau)$, we have

$$\mathrm{MSE}(\hat{a}_l(x)) = \left[-d_j \int_{-\tau}^{-|u|} K_l(t) \frac{\mu_{2,l} - \mu_{1,l} t}{\mu_{0,l}\mu_{2,l} - \mu_{1,l}^2} \,\mathrm{d}t\right]^2 + \frac{1}{nhf_X^2(x)} P(\delta_1 = 1) V_l S(x) + o\left(\frac{1}{nh}\right),$$

$$\mathrm{MSE}(\hat{a}_r(x)) = \left[\frac{1}{2} h^2 g''(s_j+) B_r\right]^2 + \frac{1}{nhf_X^2(x)} P(\delta_1 = 1) V_r S(x) + o\left(h^4 + \frac{1}{nh}\right),$$

$$\mathrm{MSE}(\hat{a}_c(x)) = \left[-d_j \int_{-\tau}^{-|u|} K_c(t) \,\mathrm{d}t\right]^2 + \frac{1}{nhf_X^2(x)} P(\delta_1 = 1) V_c S(x) + o\left(\frac{1}{nh}\right),$$

where

$$B_d = \frac{\mu_{2,d}^2 - \mu_{1,d}\mu_{3,d}}{\mu_{0,d}\mu_{2,d} - \mu_{1,d}^2}, \quad V_d = \int_{-\tau}^{\tau} K_d^2(t) \left[\frac{\mu_{2,d} - \mu_{1,d} t}{\mu_{0,d}\mu_{2,d} - \mu_{1,d}^2}\right]^2 \mathrm{d}t \quad \text{and}$$

$$S(x) = \int \frac{1}{\pi^2 (g(x) + \varepsilon)} \varepsilon^2 f_{X,\varepsilon|\delta=1}(x, \varepsilon) \,\mathrm{d}\varepsilon.$$

Theorem 1 gives the asymptotic bias and asymptotic variance of the three estimators. As a matter of fact, it extends the results of the three local linear estimators from complete data to the case of missing data. Specifically, when data is observed completely, i.e., $\pi(y) = 1$, $P(\delta_1 = 1) = 1$, $S(x)$ reduces $\sigma^2 f_X(x)$. In such a case, the result degenerates to that for the local linear estimators for fully observed data, see Gijbels et al. (2007).

From Theorem 1(i), we can conclude that the three estimators are consistent in mean square sense and have the same rate of convergence in continuity regions of $g(x)$. The asymptotic expressions in (ii) reveals that $\hat{a}_r(x)$ and $\hat{a}_c(x)$ are not consistent at any point in the neighbourhood $[s_j - \tau h, s_j)$ which is uh away from s_j with $u \in [-\tau, 0)$. However, $\hat{a}_l(x)$ is consistent. A similar discussion can be given for points on the right-side interval of the jump point s_j, that is, only $\hat{a}_r(x)$ is consistent, but $\hat{a}_l(x)$ and $\hat{a}_c(x)$ are inconsistent.

Theorem 2. *Under the regularity assumptions (A1)–(A6), the asymptotic expressions of* WRMSs *are as follows:*

(i) For any $x \in D_1$,

$$\mathrm{WRMS}_d(x) = \sigma^2 + R_{d,1}(x), \quad d = c, l, r,$$

where $R_{d,1}(x)$ are random variables tending to 0 almost surely and uniformly in $x \in D_1$.

(ii) For any $x \in D_{2,l}$, that is, $x = s_j + uh$ with $u \in (-\tau, 0)$, we have

$$\begin{aligned}
\mathrm{WRMS}_l(x) &= \sigma^2 + R_{l,2}(x), \\
\mathrm{WRMS}_r(x) &= \sigma^2 + d_j^2 C_{u,r}^2 + R_{r,2}(x), \\
\mathrm{WRMS}_c(x) &= \sigma^2 + d_j^2 C_{u,c}^2 + R_{c,2}(x),
\end{aligned}$$

where $R_{d,2}(x)$ are random variables tending to 0 almost surely and uniformly in $x \in D_{2,l}$.

(iii) For any $x \in D_{2,r}$, that is, $x = s_j + uh$ with $u \in [0, \tau)$, we have

$$\begin{aligned}
\mathrm{WRMS}_r(x) &= \sigma^2 + R_{r,3}(x), \\
\mathrm{WRMS}_l(x) &= \sigma^2 + d_j^2 C_{u,l}^2 + R_{l,3}(x), \\
\mathrm{WRMS}_c(x) &= \sigma^2 + d_j^2 C_{u,c}^2 + R_{c,3}(x),
\end{aligned}$$

where $R_{d,3}(x)$ are random variables tending to 0 almost surely and uniformly in $x \in D_{2,r}$.

in which

$$C_{u,d}^2 = \int_{-u}^{\tau} \left[\int_{-\tau}^{-u} \frac{\mu_{2,d} - \mu_{1,d} t}{\mu_{0,d}\mu_{2,d} - \mu_{1,d}^2} K_d(t)\,\mathrm{d}t - z \int_{-u}^{\tau} \frac{\mu_{0,d} t - \mu_{1,d}}{\mu_{0,d}\mu_{2,d} - \mu_{1,d}^2} K_d(t)\,\mathrm{d}t \right]^2 K_d(z)\,\mathrm{d}z$$
$$+ \int_{-\tau}^{-u} \left[\int_{-u}^{\tau} \frac{\mu_{2,d} - \mu_{1,d} t}{\mu_{0,d}\mu_{2,d} - \mu_{1,d}^2} K_d(t)\,\mathrm{d}t + z \int_{-u}^{\tau} \frac{\mu_{0,r} t - \mu_{1,d}}{\mu_{0,d}\mu_{2,d} - \mu_{1,d}^2} K_d(t)\,\mathrm{d}t \right]^2 K_d(z)\,\mathrm{d}z.$$

The asymptotic expressions of WRMS lead to a similar result as Theorem 1. From Theorem 2 (i), in continuity regions of $g(x)$, the three WRMS quantities are consistent estimators of σ^2. The asymptotic expressions in (ii) reveal that only $\mathrm{WRMS}_l(u)$ is a consistent estimator for σ in the left-sided of the neighborhood of jump point, i.e., any $u \in D_{2,l}$, while $\mathrm{WRMS}_r(u)$ and $\mathrm{WRMS}_c(u)$ are affected by the jump at s_j. Similarly, in the right-sided of the neighborhood of jump points, i.e., any $u \in D_{2,r}$, only $\mathrm{WRMS}_r(u)$ is a consistent estimator, $\mathrm{WRMS}_l(u)$ and $\mathrm{WRMS}_c(u)$ are inconsistent.

Theorem 3. *Under the regularity assumptions (A1)–(A6), for any $x \in [0,1]$, as $n \to \infty$, the estimate $\hat{g}(x)$ has the following asymptotic distribution:*

$$\sqrt{nh}\left[\hat{g}(x) - g(x) - \frac{1}{2}h^2 g''(x) B_d\right] \xrightarrow{D} N\left(0, \frac{1}{f_X^2(x)} P(\delta_1 = 1) S(x) V_d\right),$$

where "$\xrightarrow{D}$" means convergence in distribution, B_d, V_d and $S(x)$ are defined in Theorem 1. Here, when $x \in D_{2,l}$ and $D_{2,r}$, $g''(x)$ is replaced by $g''(s_j-)$ and $g''(s_j+)$ respectively, the left and right limits of $g(\cdot)$ at the point s_j.

Theorem 3 reveals that the resulting estimator $\hat{g}(x)$ is asymptotically normal on the whole support of x. Specifically, $\hat{g}(x)$ is asymptotically normal when $u \in D_1$, B_c and V_c; $u \in D_{2,l}$, B_l and V_l; $u \in D_{2,r}$, B_r and V_r are used. Moreover, similar to the discussion in Theorem 1, if $\pi(y) = 1$ and $P(\delta_1 = 1) = 1$ the asymptotic distribution of $\hat{g}(x)$ reduces to that of the estimator when data is observed completely, see Li and Racine (2007).

2.3 Estimation when $\pi(Y)$ is unknown

In fact, the selection probability function $\pi(Y)$ is generally unknown but can be estimated. To estimate $\pi(Y)$, we now consider the case that it is a parametric model, denoted by $\pi(Y, \alpha)$, where α is some unknown parameter vector that needs to be estimated.

Here, $\pi(Y, \alpha)$ is assumed to be a logistic model, i.e.,

$$\pi(Y_i, \alpha) = P(\delta = 1 | Y_i) = \frac{e^{\alpha_0 + \alpha_1 Y_i}}{1 + e^{\alpha_0 + \alpha_1 Y_i}},$$

where $\alpha = (\alpha_0, \alpha_1)^\top$. By applying the maximum likelihood approach, one easily obtains a root-n-consistent estimate $\hat{\alpha}$, see Robins et al. (1994) and Wang et al. (1998) for related studies and Hosmer Jr et al. (2013) for a global statistic test for examining the pre-assumed binary regression model. Denote the resulting selection probability function estimator $\hat{\pi}_i := \pi(Y_i, \hat{\alpha})$, $i = 1, \ldots, n$. Thus, replacing π_i in (6) with $\hat{\pi}_i$, we obtain the three associated estimators $\hat{a}_d(\cdot, \hat{\pi})$ of $g(\cdot)$, they have the following expressions:

$$\hat{a}_d(x, \hat{\pi}) = \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i} K_d\left(\frac{X_i - x}{h}\right) \frac{\hat{S}_{2,d} - \hat{S}_{1,d}(X_i - x)}{\hat{S}_{0,d}\hat{S}_{2,d} - \hat{S}_{1,d}^2} Y_i,$$

where $\hat{S}_{j,d} = \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i} K_d\left(\frac{X_i - x}{h}\right)(X_i - x)^j$ for $j = 0, 1, 2$, $d = c, l, r$. Note that $\hat{a}_d(\cdot, \hat{\pi})$ is used to emphasise its dependence on the estimator $\hat{\pi}$. The same is true for the following estimators, for which the value of $\hat{\pi}$ is provided in parentheses.

Similarly, as discussed in (7), by replacing π_i with $\hat{\pi}_i$, the resulting estimator $\hat{g}(\cdot, \hat{\pi})$ of $g(\cdot)$ is derived by

$$\hat{g}(x,\hat{\pi}) = \begin{cases} \hat{a}_c(x,\hat{\pi}), & \text{if } \mathrm{diff}(x,\hat{\pi}) \le \lambda, \\ \hat{a}_l(x,\hat{\pi}), & \text{if } \mathrm{diff}(x,\hat{\pi}) > \lambda \text{ and } \mathrm{WRMS}_l(x,\hat{\pi}) < \mathrm{WRMS}_r(x,\hat{\pi}), \\ \hat{a}_r(x,\hat{\pi}), & \text{if } \mathrm{diff}(x,\hat{\pi}) > \lambda \text{ and } \mathrm{WRMS}_l(x,\hat{\pi}) > \mathrm{WRMS}_r(x,\hat{\pi}), \\ (\hat{a}_l(x,\hat{\pi}) + \hat{a}_r(x,\hat{\pi}))/2, & \text{if } \mathrm{diff}(x,\hat{\pi}) > \lambda \text{ and } \mathrm{WRMS}_l(x,\hat{\pi}) = \mathrm{WRMS}_r(x,\hat{\pi}). \end{cases} \tag{8}$$

where

$$\mathrm{WRMS}_d(x,\hat{\pi}) = \frac{\sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_i}\left[Y_i - \hat{a}_d(x,\hat{\pi}) - \hat{b}_d(x,\hat{\pi})(X_i - x)\right]^2 K_d\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_i} K_d\left(\frac{X_i - x}{h}\right)},$$

for $d = c, l, r$. In (8),

$$\mathrm{diff}(x,\hat{\pi}) = \max\{\mathrm{WRMS}_c(x,\hat{\pi}) - \mathrm{WRMS}_l(x,\hat{\pi}), \mathrm{WRMS}_c(x,\hat{\pi}) - \mathrm{WRMS}_r(x,\hat{\pi})\}.$$

Next, we establish the asymptotic properties of the above estimators. Here, it is assumed that the parametric model for π is correctly specified so that the estimator $\hat{\alpha}$ satisfies $\hat{\alpha} - \alpha = O_p(n^{-1/2})$. The following Theorem 4 compares the difference between the estimator based on the true π and that based on the estimated $\hat{\pi}$. In order to obtain this theorem, it is necessary to fulfil the following additional condition:

(A7) The selection probability function $\hat{\pi}$ follows a parametric binary model. Moreover, it has bounded second order partial derivative with respect to y and has bounded first order partial derivative with respect to α.

Theorem 4. *Under the assumptions (A1)-(A7), as $n \to \infty$,*

$$\sup_{x\in[0,1]} |\hat{g}(x,\hat{\pi}) - \hat{g}(x)| = O_p(n^{-1/2}).$$

Combining this Theorem and Theorems 1-3, when the selection probability function π is replaced by $\hat{\pi}$, it is easy to show the following results:

Theorem 5. *Under the regularity assumptions (A1)–(A7), the mean squared errors (MSE) of the three estimators of the function $g(\cdot)$ are as follows:*

(i) For any $x \in D_1$,

$$\mathrm{MSE}(\hat{a}_d(x,\hat{\pi})) = \left[\frac{1}{2}h^2 g''(x) B_d\right]^2 + \frac{1}{nhf_X^2(x)} P(\delta_1 = 1) V_d S(x) + o\left(h^4 + \frac{1}{nh}\right), \quad d = c, l, r,$$

(ii) For any $x \in D_{2,l}$, that is, $x = s_j + uh$ with $u \in (-\tau, 0)$, we have

$$\mathrm{MSE}(\hat{a}_l(x,\hat{\pi})) = \left[\frac{1}{2}h^2 g''(s_j-) B_l\right]^2 + \frac{1}{nhf_X^2(x)} P(\delta_1 = 1) V_l S(x) + o\left(h^4 + \frac{1}{nh}\right),$$

$$\mathrm{MSE}(\hat{a}_r(x,\hat{\pi})) = \left[d_j \int_{|u|}^{\tau} K_r(t) \frac{\mu_{2,r} - \mu_{1,r} t}{\mu_{0,r}\mu_{2,r} - \mu_{1,r}^2} \,\mathrm{d}t\right]^2 + \frac{1}{nhf_X^2(x)} P(\delta_1 = 1) V_r S(x) + o\left(\frac{1}{nh}\right),$$

$$\mathrm{MSE}(\hat{a}_c(x,\hat{\pi})) = \left[d_j \int_{|u|}^{\tau} K_c(t) \,\mathrm{d}t\right]^2 + \frac{1}{nhf_X^2(x)} P(\delta_1 = 1) V_c S(x) + o\left(\frac{1}{nh}\right).$$

(iii) For any $x \in D_{2,r}$, that is, $x = s_j + uh$ with $u \in [0, \tau)$, we have

$$\mathrm{MSE}(\hat{a}_l(x,\hat{\pi})) = \left[-d_j \int_{-\tau}^{-|u|} K_l(t) \frac{\mu_{2,l} - \mu_{1,l} t}{\mu_{0,l}\mu_{2,l} - \mu_{1,l}^2} \,\mathrm{d}t\right]^2 + \frac{1}{nhf_X^2(x)} P(\delta_1 = 1) V_l S(x) + o\left(\frac{1}{nh}\right),$$

$$\mathrm{MSE}(\hat{a}_r(x,\hat{\pi})) = \left[\frac{1}{2}h^2 g''(s_j+) B_r\right]^2 + \frac{1}{nhf_X^2(x)} P(\delta_1 = 1) V_r S(x) + o\left(h^4 + \frac{1}{nh}\right),$$

$$\mathrm{MSE}(\hat{a}_c(x,\hat{\pi})) = \left[-d_j \int_{-\tau}^{-|u|} K_c(t) \,\mathrm{d}t\right]^2 + \frac{1}{nhf_X^2(x)} P(\delta_1 = 1) V_c S(x) + o\left(\frac{1}{nh}\right),$$

where B_d, V_d and $S(x)$ are defined in Theorem 1.

Theorem 6. *Under the regularity assumptions (A1)–(A7), the asymptotic expressions of* WRMSs *are as follows:*

(i) For any $x \in D_1$,

$$\mathrm{WRMS}_d(x,\hat{\pi}) = \sigma^2 + R^*_{d,1}(x), \quad d = c, l, r,$$

where $R^*_{d,1}(x)$ are random variables tending to 0 almost surely and uniformly in $x \in D_1$.

(ii) For any $x \in D_{2,l}$, that is, $x = s_j + uh$ with $u \in (-\tau, 0)$, we have

$$\begin{aligned}
\mathrm{WRMS}_l(x,\hat{\pi}) &= \sigma^2 + R^*_{l,2}(x), \\
\mathrm{WRMS}_r(x,\hat{\pi}) &= \sigma^2 + d_j^2 C_{u,r}^2 + R^*_{r,2}(x), \\
\mathrm{WRMS}_c(x,\hat{\pi}) &= \sigma^2 + d_j^2 C_{u,c}^2 + R^*_{c,2}(x),
\end{aligned}$$

where $R^*_{d,2}(x)$ are random variables tending to 0 almost surely and uniformly in $x \in D_{2,l}$.

(iii) For any $x \in D_{2,r}$, that is, $x = s_j + uh$ with $u \in [0, \tau)$, we have

$$\begin{aligned}
\mathrm{WRMS}_r(x,\hat{\pi}) &= \sigma^2 + R^*_{r,3}(x), \\
\mathrm{WRMS}_l(x,\hat{\pi}) &= \sigma^2 + d_j^2 C_{u,l}^2 + R^*_{l,3}(x), \\
\mathrm{WRMS}_c(x,\hat{\pi}) &= \sigma^2 + d_j^2 C_{u,c}^2 + R^*_{c,3}(x),
\end{aligned}$$

where $R^*_{d,3}(x)$ are random variables tending to 0 almost surely and uniformly in $x \in D_{2,r}$.

The symbol $C_{u,d}$ is defined by Theorem 2.

Theorem 7. *Under the regularity assumptions (A1)–(A7), for any $x \in [0,1]$, as $n \to \infty$, the estimate $\hat{g}(x,\hat{\pi})$ has the following asymptotic distribution:*

$$\sqrt{nh}\left[\hat{g}(x,\hat{\pi}) - g(x) - \frac{1}{2}h^2 g''(x) B_d\right] \xrightarrow{D} N\left(0, \frac{1}{f_X^2(x)} P(\delta_1 = 1) S(x) V_d\right),$$

where "$\xrightarrow{D}$" means convergence in distribution, B_d, V_d and $S(x)$ are defined in Theorem 1. Here, when $x \in D_{2,l}$, $g''(x)$ is replaced by $g''(s_j-)$, when $x \in D_{2,r}$, $g''(x)$ is replaced by $g''(s_j+)$.

From Theorems 5-7, when π is replaced by its consistent estimator $\hat{\pi}$, the asymptotic property of our proposed estimators yields the same results as described in Subsection 2.2.

When x is in boundary regions $[0, \tau h)$ and $(1 - \tau h, 1]$, the estimator of $g(x)$ is not defined by (8). In such cases there are several possible approaches to estimate $g(x)$ if no jumps exist in $[0, \tau h)$ and $(1 - \tau h, 1]$. For example, $\hat{g}(x,\hat{\pi})$ could be defined by the conventional local linear kernel estimator constructed from observations in $[0, x + \tau h)$ or $(x - \tau h, 1]$ depending on whether $x \in [0, \tau h)$ or $x \in (1 - \tau h, 1]$. If there are jump points in $[0, \tau h)$ or $(1 - \tau h, 1]$. For convenience, we define $\hat{g}(x,\hat{\pi}) = \hat{a}_r(x,\hat{\pi})$ when $x \in [0, \tau h)$ and $\hat{g}(x,\hat{\pi}) = \hat{a}_l(x,\hat{\pi})$ when $x \in (1 - \tau h, 1]$.

Now we estimate the structure of the jump points. In order to achieve this goal, it follows from Theorem 6 that $\text{diff}(x, \hat{\pi})$ is an appropriate criterion for detecting jumps. If there exists a jump point around x, the jump detection criterion $\text{diff}(x, \hat{\pi})$ will be relatively large. Otherwise, it will be relatively small. Therefore, $\text{diff}(x, \hat{\pi})$ can be used to detect the jumps.

In practice, if $\text{diff}(x, \hat{\pi})$ is large enough, i.e., $\text{diff}(x, \hat{\pi}) > \lambda$, then x can be regarded as a jump point, where λ is a a threshold. However, the points in the neighbourhood of true jump points maybe wrongly regarded as jumps even if they are actually not. To delete those false jump points, inspired by Qiu (1994), we suggest the following jump detection procedure. Let $\{x_i^*, i = 1, 2, \ldots, m\}$ be the set of detected jump points satisfying

$$\text{diff}(x_i^*, \hat{\pi}) \geq \lambda, \quad \text{for } i = 1, 2, \ldots, m.$$

If there are integers $1 \leq t_1 \leq t_2 \leq m$ such that $x_j^* - x_{j-1}^* \leq \tau h$, for $j = t_1 + 1, \ldots, t_2$, $x_{t_1}^* - x_{t_1-1}^* > \tau h$, and $x_{t_2+1}^* - x_{t_2}^* > \tau h$, then the set $\left\{x_{t_1}^*, x_{t_1+1}^*, \ldots, x_{t_2}^*\right\}$ forms a tie in $\{x_i^*, i = 1, 2, \ldots, m\}$ and the entire tie set is replaced by its central point $(x_{t_1}^* + x_{t_2}^*)/2$ for estimating the jump positions. After this modification, the detected jump points and the corresponding jump magnitudes are denoted as

$$\hat{s}_j, \qquad \hat{d}_j = \hat{a}_r(\hat{s}_j, \hat{\pi}) - \hat{a}_l(\hat{s}_j, \hat{\pi}), \qquad \text{for} \quad j = 1, 2, \ldots, \hat{J}.$$

2.4 Choice of procedure parameters

In the construction of our estimator $\hat{g}(\cdot, \hat{\pi})$ in (8), the bandwidth parameter h and threshold λ need to be chosen. In an ideal scenario, the available bandwidth h should be adaptable to accommodate the unknown curve. This would require the implementation of a variable bandwidth approach. However, variable bandwidth selectors are typically complicated and computationally demanding. Nevertheless, they are not always capable of adapting to all jumps. In this paper, therefore, we use a simple procedure based on leave-one-out cross-validation (see Rice and Silverman (1991)), obtaining the global bandwidth and threshold as

$$(\hat{h}, \hat{\lambda}) = \arg\min_{h,\lambda} \sum_{i=1}^{n} \delta_i \left[Y_i - \hat{g}_{(-i)}(X_i)\right]^2, \tag{9}$$

where $\hat{g}_{(-i)}(\cdot)$ is the "leave-one-out" estimator of $g(\cdot)$ based on the sample with ith subject data deleted. Namely, the observation (X_i, Y_i) is left out in constructing $\hat{g}_{(-i)}(X_i)$, for $i = 1, 2, \ldots, n$.

To solve the minimization problem in (9), we need to specify a grid for the λ-values. A suitable range of threshold λ can be obtained by looking at the result in Theorem 6. We propose taking $\lambda_{\max} = d^2 \max_u C_{u,c}^2$ as an upper bound for a range of values for λ. The quantity d can be estimated by $\sup_x |\hat{a}_l(x, \hat{\pi}) - \hat{a}_r(x, \hat{\pi})|$. When we detect the jump points, the value of threshold λ should not be very small, in fact, it is chosen to be the 0.9th quantile of $\{\text{diff}(X_i, \hat{\pi}), i = 1, 2, \ldots, n\}$.

3 Numerical studies

In this section, we carry out some numerical simulations to investigate the finite sample performance of the procedure described in Section 2. We generated 100 Monte Carlo random samples of size $n = 200$, 500, 800 from model (1), where $X \sim U(0, 1)$ and $\varepsilon \sim N(0, \sigma^2)$. We considered three sets of $\sigma = 0.1$ and 0.2 to examine the performance for different levels of signal-to-noise ratio. The following regression is considered:

$$g(u) = \begin{cases} 0, & 0 \leq u \leq 0.3; \\ 3u^2 + 0.93, & 0.3 \leq u \leq 0.7; \\ 4u^2 + 1.24, & 0.7 \leq u \leq 1. \end{cases}$$

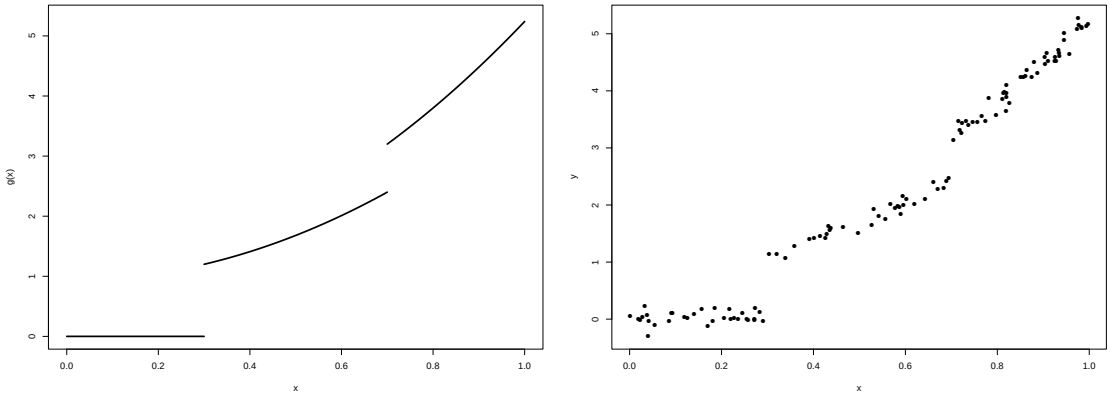

Figure 1: *The left panel shows real curve of the case for noparametric function. The right panel shows scatter plot of simulated data when* $n = 200$, $\sigma = 0.1$, *and MR*=25%.

It is clear that $g(u)$ has two jumps at $u = 0.3$ and $u = 0.7$ with jump magnitudes 1.2 and 0.8. The left panel of Figure 1 shows the real curves of this function.

The missing data X was assumed to be missing at random and the selection probabilities are specified as the logistic regression model

$$\pi(y) = \frac{e^{\alpha_0+\alpha_1 y}}{1 + e^{\alpha_0+\alpha_1 y}}.$$

Here we consider three missing rates (MR) by adjusting value α_0 and α_1. In particular, we take (α_0, α_1) as $(1, 3)$, $(1, 0.03)$ and $(-0.5, 0.5)$, respectively, leading to approximately $\pi_1 = 10\%$ (low proportion), $\pi_2 = 25\%$ and $\pi_3 = 40\%$ (high proportion of missing) of the data missing. The right panel of Figure 1 shows scatter plot of simulated data when $n = 200$, $\sigma = 0.1$, and MR=25%.

In the simulation, the kernel function K used in any estimation is chosen to be the standard Gaussian density. The bandwidth selection is introduced in Subsection 2.4. To show the estimation efficiency of the inverse probability weighting (IPW-JP) method, we compared it with the direct elimination (DE-JP) and oracle (O-JP) method. For the direct elimination method, the data with deleting missing data is used. For the oracle method, all of the data (X_i, Y_i), $i = 1, \ldots, n$ with no missing data are used. For comparison, we also list the results of above three procedures for local linear regression by directly $g(\cdot)$ is a continuous function, denoted by IPW-LL, DE-LL and O-LL. To evaluate results of curve fitting, we compute the root mean squared errors (RMSE), which is given by

$$\text{RMSE} = \left\{ \frac{1}{n} \sum_{i=1}^{n} [\hat{g}(x_i) - g(x_i)]^2 \right\}^{1/2}$$

where $\{x_i, i = 1, \ldots, n\}$ are equidistant grid points in $[0, 1]$.

For MR=10%, 40%, Figure 2 depict the true function and its estimated versions using the IPW-JP and the DE-JP estimators at noise level $= 0.1$ with sample size $n = 200$ based on $R = 100$ simulations, respectively. In each plot, the true regression function is presented by a black solid curve, a blue dotted curve and a red dashed curve depict the median of $R = 100$ replicated fits of the IPW-JP and DE-JP estimators, respectively. From the figure, it can be seen that for both the methods, the curve fitting is less effective when MR is taken as 40% as compared to when MR is taken as 10%. Moreover, the fitted curves by IPW-JP and DE-JP methods are resonably close to the true curve, which indicates that two methods performs well in this case. For further comparison, below we calculated the RMSE of the various methods.

Table 1 presents mean and standard deviation of the RMSEs based on the 100 replications with all methods described in the preceding paragraph. From the table, one can have the following conclusions.

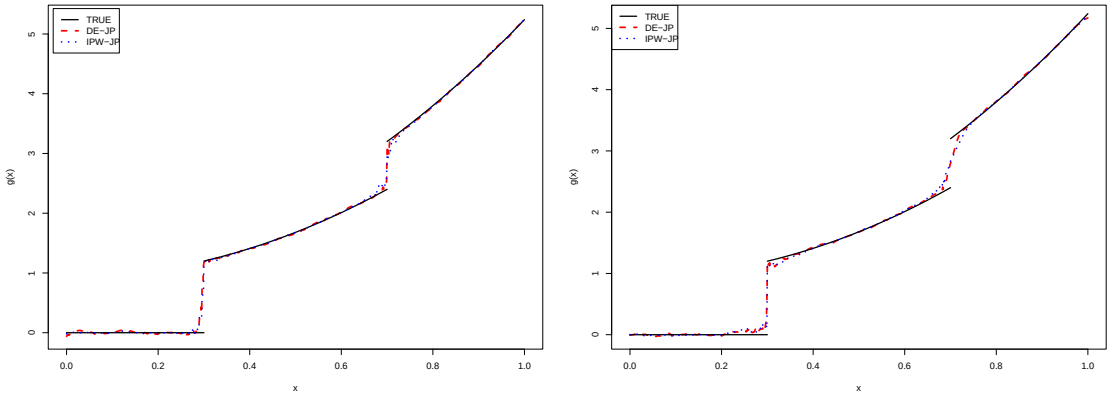

Figure 2: *Plots of fitted curves for different MR when* $n = 200$ *and* $\sigma = 0.1$ *based on* $R = 100$ *repetitions. Left panel: MR=10%; Right panel: MR=40%. The true function (black solid curve), the median of IPW-JP estimator (blue dotted curve) and the median of DE-IPW estimator (red dashed curve) are depicted.*

(i) As σ increases, the RMSE values for the six methods decreases. Meanwhile, the difference between the RMSE values of the two methods O-JP and O-LL (i.e. data fully observed) becomes smaller as σ increases, implying that σ has a greater influence on jump-preserving. (ii) When σ and missing rate π_i are fixed, the RMSE values decrease as n increase, which means consistency of our method. (iii) When the data is missing, the RMSE values increases with the increase of π_i for the four methods IPW-JP, DE-JP, IPW-LL and DE-LL. Moreover, when n is small and π_i is large, the performance of IPW-JP is a little worse than the performance of DE-JP, which may be caused by too much missing data so that less information is available. (iv) In comparison, generally, our proposed IPW-JP method is superior to other methods.

Next we consider the accuracy of the detected jump points, i.e. their number and locations. Let $\hat{S} = \{\hat{s}_1, \ldots, \hat{s}_{\hat{J}}\}$ and $S = \{s_1, \ldots, s_J\}$ denote the sets of detected jump points and true jump points, respectively. To examine how close the estimated jumps are to the true jumps, a resonable measure is the following Hausdorff distance:

$$d_H(\hat{S}, S) = \max\left(\max_{u\in\hat{S}}\min_{v\in S}|u-v|, \max_{v\in S}\min_{u\in\hat{S}}|u-v|\right).$$

The smaller the value of $d_H(\hat{S}, S)$, the closer $\hat{S}$ is to S. When $\sigma = 0.1$, the average values of detected jump points and the average Hausdorff distances for various methods, computed based on 100 replicates, are reported in Table 2. Obviously, when n is large or the missing rate is small, the number of jump points obtained by proposed method is closer to the true number 2, and the Hausdorff distance between S and $\hat{S}$ is also smaller. Moreover, as in Table 1, the DE-JP method is not as effective as the IPW-JP method in terms of he number of jump points and the Hausdorff distance.

4 Conclusion

In this paper, for the estimation of discontinuous nonparametric model with missing covariates at random, we present a weignted local linear jump-preserving estimator based on the inverse selection probability. This approach not only can provide a jump-preserving estimation for the completely unknown regression function, but also can accommodate instances of missing data on response. The proposed estimator for the discontinuous function is shown to be oracally efficient in the sense that using root-n consistent selection probability estimates is as efficient as that when the selection probabilities

Table 1: *The mean (standard deviation) of the RMSEs are report based on* 100 *replications.*

n	σ		IPW-JP	DE-JP	O-JP	IPW-LL	DE-LL	O-LL
200	0.1	π_1	0.0921 (0.0360)	0.1088 (0.0526)	0.0883 (0.0389)	0.1082 (0.0160)	0.0986 (0.0427)	0.0969 (0.0521)
		π_2	0.1217 (0.0858)	0.1141 (0.0460)	0.0840 (0.0355)	0.1144 (0.0152)	0.1005 (0.0315)	0.0892 (0.0126)
		π_3	0.1443 (0.0670)	0.1358 (0.0585)	0.0857 (0.0365)	0.1359 (0.0263)	0.1217 (0.0471)	0.1008 (0.0690)
	0.2	π_1	0.1207 (0.0295)	0.1333 (0.0310)	0.1169 (0.0333)	0.1358 (0.0127)	0.1196 (0.0097)	0.1149 (0.0099)
		π_2	0.1558 (0.0733)	0.1355 (0.0356)	0.1176 (0.0268)	0.1460 (0.0186)	0.1269 (0.0189)	0.1143 (0.0104)
		π_3	0.1624 (0.0568)	0.1462 (0.0344)	0.1128 (0.0305)	0.1574 (0.0215)	0.1375 (0.0180)	0.1166 (0.0150)
500	0.1	π_1	0.0363 (0.0177)	0.0610 (0.0183)	0.0317 (0.0084)	0.0816 (0.0092)	0.0666 (0.0073)	0.0644 (0.0067)
		π_2	0.0641 (0.0686)	0.0717 (0.0428)	0.0347 (0.0140)	0.0841 (0.0087)	0.0756 (0.0335)	0.0689 (0.0287)
		π_3	0.0889 (0.0678)	0.0909 (0.0313)	0.0332 (0.0134)	0.0957 (0.0127)	0.0816 (0.0118)	0.0650 (0.0059)
	0.2	π_1	0.0800 (0.0287)	0.0997 (0.0243)	0.0702 (0.0269)	0.1047 (0.0093)	0.0930 (0.0064)	0.0877 (0.0054)
		π_2	0.0991 (0.0570)	0.1023 (0.0228)	0.0661 (0.0219)	0.1080 (0.0101)	0.0951 (0.0078)	0.0880 (0.0067)
		π_3	0.1249 (0.0769)	0.1157 (0.0343)	0.0655 (0.0221)	0.1204 (0.0124)	0.1093 (0.0172)	0.0877 (0.0059)
800	0.1	π_1	0.0352 (0.0188)	0.0561 (0.0190)	0.0317 (0.0171)	0.0718 (0.0073)	0.0593 (0.0056)	0.0579 (0.0051)
		π_2	0.0416 (0.0431)	0.0569 (0.0192)	0.0287 (0.0141)	0.0738 (0.0063)	0.0618 (0.0050)	0.0571 (0.0048)
		π_3	0.0513 (0.0318)	0.0711 (0.0243)	0.0261 (0.0036)	0.0841 (0.0116)	0.0708 (0.0117)	0.0568 (0.0050)
	0.2	π_1	0.0616 (0.0231)	0.0910 (0.0210)	0.0526 (0.0184)	0.0940 (0.0072)	0.0817 (0.0057)	0.0775 (0.0056)
		π_2	0.0659 (0.0194)	0.1032 (0.0293)	0.0575 (0.0207)	0.0964 (0.0113)	0.0822 (0.0055)	0.0746 (0.0041)
		π_3	0.0790 (0.0229)	0.1055 (0.0255)	0.0523 (0.0192)	0.1052 (0.0096)	0.0918 (0.0087)	0.0764 (0.0053)

Table 2: *Average (standard deviation) of the number of jump detected and Hausdorff distances are report based on* 100 *replications when* $\sigma = 0.1$.

		IPW-JP		DE-JP		O-JP	
n		No	d_H	No	d_H	No	d_H
200	π_1	1.96 (0.2953)	0.0406 (0.1033)	2.53 (0.9732)	0.0983 (0.1183)	1.96 (0.2953)	0.0386 (0.1023)
	π_2	1.74 (0.6109)	0.1956 (0.2491)	2.67 (0.7581)	0.1290 (0.1205)	1.98 (0.2744)	0.0331 (0.0902)
	π_3	1.61 (0.5394)	0.2100 (0.2289)	2.43 (1.0726)	0.1107 (0.1443)	1.94 (0.2436)	0.0329 (0.0957)
500	π_1	1.97 (0.1826)	0.0150 (0.0728)	2.23 (0.6789)	0.0807 (0.1302)	1.97 (0.1826)	0.0143 (0.0729)
	π_2	1.92 (0.3959)	0.0866 (0.2143)	2.30 (0.8631)	0.0896 (0.1326)	2.00 (0.2020)	0.0132 (0.0632)
	π_3	1.89 (0.3333)	0.0822 (0.2318)	2.22 (0.4410)	0.0878 (0.0758)	1.99 (0.1738)	0.0100 (0.0580)
800	π_1	2.03 (0.1826)	0.0100 (0.0548)	2.16 (0.5834)	0.0566 (0.0973)	2.02 (0.1291)	0.0040 (0.0256)
	π_2	1.93 (0.3651)	0.0683 (0.1906)	2.08 (0.5687)	0.0890 (0.1268)	2.00 (0.1841)	0.0098 (0.0563)
	π_3	2.10 (0.5477)	0.1010 (0.1748)	1.86 (0.5899)	0.1459 (0.1695)	2.03 (0.1826)	0.0097 (0.0530)

are known as a prior. The asymptotic properties of our estimator can be established through under some mild conditions. Numerical studies indicate that the prodedure works well in applications.

However, the following issues related to this topic need further investigation. Firstly, only fully observed cases contribute to the proposed estimator, the information of partly observed cases is not used in regression (but it is used in estimating π). This leads to a loss of efficiency. Secondly, we assume that no jumps exist in $[0, \tau h)$ and $(1 - \tau h, 1]$. This condition can always be satisfied when the sample size is large. When the sample size is small, however, this condition may not be true in some cases, estimation of $g(x)$ in the boundary region is still an open problem. Finally, the selection of procedure parameters is uniform throughout the entire design interval, which may not be ideal for certain applications. Further investigation and analysis are required for selecting variable procedure parameters.

References

Chen J, Fan J, Li K H, Zhou H. Local quasi-likelihood estimation with data missing at random. *Statistica Sinica*, 2006, 16(4): 1071–1100.

Claeskens G, Van Keilegom I. Bootstrap confidence bands for regression curves and their derivatives. *The Annals of Statistics*, 2003, 31(6): 1852–1884.

El Ghouch A, Genton M G. Local polynomial quantile regression with parametric features *Journal of the American Statistical Association*, 2009, 104(488): 1416–1429.

Efromovich S. Nonparametric regression with predictors missing at random. *Journal of the American Statistical Association*, 2011, 106(493): 306–319.

Efromovich S. Nonparametric regression with responses missing at random. *Journal of Statistical Planning and Inference*, 2011, 141(12): 3744–3752.

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

Gijbels I, Lambert A, Qiu P. Jump-preserving regression and smoothing using local linear fitting: a compromise. *Annals of the Institute of Statistical Mathematics*, 2007, 59(2): 235–272.

Hall P, Titterington D M. Edge-preserving and peak-preserving smoothing. *Technometrics*, 1992, 34(4): 429–440.

Han Z, Lin J, Zhao Y. Adaptive semiparametric estimation for single index models with jumps. *Computational Statistics & Data Analysis*, 2020, 151: 107013.

Hawkins D M, Olwell D H. *Cumulative Sum Charts and Charting for Quality Improvement*. Springer, New York, 1998.

Healy M, Westmacott M. Missing values in experiments analysed on automatic computers. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 1956, 5(3): 203–206.

Horvitz D G, Thompson D J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 1952, 47(260): 663–685.

Hosmer Jr D W, Lemeshow S, Sturdivant R X. *Applied Logistic Regression*. John Wiley & Sons, New York, 2013.

Huang J Z. Local asymptotics for polynomial spline regression. *The Annals of Statistics*, 2003, 31(5): 1600–1635.

Joo J H, Qiu P. Jump detection in a regression curve and its derivative. *Technometrics*, 2009, 51(3): 289–305.

Kang Y, Shi Y, Jiao Y, Li W, Xiang D. Fitting jump additive models. *Computational Statistics & Data Analysis*, 2021, 162: 107266.

Kim J K, Shao, J. *Statistical Methods for Handling Incomplete Data, 2rd ed.* Chapman and Hall/CRC, New York, 2013.

Li C, Gu L, Wang Q, Wang S. Simultaneous confidence bands for nonparametric regression with missing covariate data. *Annals of the Institute of Statistical Mathematics*, 2021, 73(6): 1249–1279.

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

Li Q, Li J, Cheng Y, Zhang R. Curve fitting and jump detection on nonparametric regression with missing data. *Journal of Applied Statistics*, 2023, 50(4): 963–983.

Liang H, Wang S, Carroll R J. Partially linear models with missing response variables and error-prone covariates. *Biometrika*, 2007, 94(1): 185–198.

Liang H, Wang S, Robins J M, Carroll R J. Estimation in partially linear models with missing covariates. *Journal of the American Statistical Association*, 2004, 99(466): 357–367.

Little R J A, Rubin D B. *Statistical Analysis with Missing Data, 3rd ed.* John Wiley & Sons, New York, 2019.

McDonald J A, Owen A B. Smoothing with split linear fits. *Technometrics*, 1986, 28(3): 195–208.

Qiu P. Estimation of the number of jumps of the jump regression functions *Communications in Statistics-Theory and Methods*, 1994, 23(8): 2141–2155.

Qiu P. A jump-preserving curve fitting procedure based on local piecewise-linear kernel estimation. *Journal of Nonparametric Statistics*, 2003, 15(4-5): 437–453.

Qiu P. *Image Processing and Jump Regression Analysis*. John Wiley & Sons, New York, 2005.

Qiu P. Jump-preserving surface reconstruction from noisy data. *Annals of the Institute of Statistical Mathematics*, 2009, 61(3): 715–751.

Rice J A, Silverman B W. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 1991, 53(1): 233–243.

Robins J M, Rotnitzky A, Zhao L P. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 1994, 89(427): 846–866.

Seaman S R, White I R. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 2013, 22(3): 278–295.

Stute W, Xue L, Zhu L. Empirical likelihood inference in nonlinear errors-in-covariables models with validation data. *Journal of the American Statistical Association*, 2007, 102(477): 332–346.

Sun Y, Wang L, Han P. Multiply robust estimation in nonparametric regression with missing data. *Journal of Nonparametric Statistics*, 2020, 32(1): 73–92.

Wang C, Wang S, Gutierrez R G, Carroll R J. Local linear regression for generalized linear models with missing data. *The Annals of Statistics*, 1998, 26(3): 1028–1050.

Wang G, Zou C, Qiu P. Data-driven determination of the number of jumps in regression curves. *Technometrics*, 2022, 64(3): 312–322.

Wang J, Yang L. Polynomial spline confidence bands for regression curves. *Statistica Sinica*, 2009, 19(1): 325–342.

Wang L, Cao R, Du J, Zhang Z. A nonparametric inverse probability weighted estimation for functional data with missing response data at random. *Journal of the Korean Statistical Society*, 2019, 48(4): 537–546.

Wang L, Rotnitzky A, Lin X. Nonparametric regression with missing outcomes using weighted kernel estimating equations. *Journal of the American Statistical Association*, 2010, 105(491): 1135–1146.

Wang Q. Statistical estimation in partial linear models with covariate data missing at random. *Annals of the Institute of Statistical Mathematics*, 2009, 61(1): 47–84.

Wang Q, Rao J N K. Empirical likelihood-based inference in linear errors-in-covariables models with validation data. *Biometrika*, 2002, 89(2): 345–358.

Wang Q, Rao J N K. Empirical likelihood-based inference under imputation for missing response data *The Annals of Statistics*, 2002, 30(3): 896–924.

Wang Q, Linton O, Härdle W. Semiparametric regression analysis with missing response at random. *Journal of the American Statistical Association*, 2004, 99(466): 334–345.

Wooldridge J M. Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, 2007, 141(2): 1281–1301.

Xia Z, Qiu P. Jump information criterion for statistical inference in estimating discontinuous curves. *Biometrika*, 2015, 102(2): 397–408.

Xiao Z, Linton O B, Carrol r j, Mammen E. More Efficient Local Polynomial Estimation in Nonparametric Regression With Autocorrelated Errors. *Journal of the American Statistical Association*, 2003, 98(464): 980–992.

Yang Y, Song Q. Jump detection in time series nonparametric regression models: a polynomial spline approach. *Annals of the Institute of Statistical Mathematics*, 2014, 66(2): 325–344.

Zhao Y, Lin J, Huang X, Wang H. Adaptive jump-preserving estimates in varying-coefficient models. *Journal of Multivariate Analysis*, 2016, 149: 65–80.

Zhou Y, Wan A, Wang X. Estimating equations inference with missing data. *Journal of the American Statistical Association*, 2008, 103(483): 1187–1199.

Appendix

In this section, we provide proofs for Theorems 1 and 2. Firstly, a lemma is introduced, it will be used in the proofs of the Theorems.

A.1 Proof of Theorem 1

Proof. (i). Suppose $x \in D_1$, by Taylor's expansion, it followws that

$$g(X_i) = g(x) + g'(x)(X_i - x) + \frac{1}{2}g''(x)(X_i - x)^2 + o(h^2),$$

where $X_i \in [x - \tau h, x)$. Therefore, $\hat{a}_d(x)$ can be writen as

$$\begin{aligned}
\hat{a}_d(x) &= \sum_{i=1}^{n} \frac{\delta_i}{\pi_i}(g(X_i) + \varepsilon_i)K_d\left(\frac{X_i - x}{h}\right)\frac{S_{2,d} - S_{1,d}(X_i - x)}{S_{0,d}S_{2,d} - S_{1,d}^2} \\
&= g(x) + \frac{1}{2}g''(x)\frac{S_{2,d}^2 - S_{1,d}S_{3,d}}{S_{0,d}S_{2,d} - S_{1,d}^2} + \sum_{i=1}^{n} \frac{\delta_i}{\pi_i}\varepsilon_i K_d\left(\frac{X_i - x}{h}\right)\frac{S_{2,d} - S_{1,d}(X_i - x)}{S_{0,d}S_{2,d} - S_{1,d}^2} + o(h^2) \\
&\triangleq g(x) + A_1 + A_2 + o(h^2).
\end{aligned} \tag{10}$$

Furthermore, for $S_{j,d}$, from lemma 3 of Li et al. (2021), it can be deduced that

$$\frac{1}{nh^{j+1}}S_{j,d} = f_X(x)\mu_{j,d} + o_p(1). \tag{11}$$

which imply that

$$A_1 = \frac{1}{2}h^2 g''(x)\frac{\mu_{2,d}^2 - \mu_{1,d}\mu_{3,d}}{\mu_{0,d}\mu_{2,d} - \mu_{1,d}^2} + o_p(1). \tag{12}$$

For A_2, notice that

$$\begin{aligned}
E(A_2) &= nE\left[\frac{\delta_i}{\pi_i}\varepsilon_i K_d\left(\frac{X_i - x}{h}\right)\frac{S_{2,d} - S_{1,d}(X_i - x)}{S_{0,d}S_{2,d} - S_{1,d}^2}\right] \\
&= nE\left[E\left\{\frac{\delta_i}{\pi_i}\varepsilon_i K_d\left(\frac{X_i - x}{h}\right)\frac{S_{2,d} - S_{1,d}(X_i - x)}{S_{0,d}S_{2,d} - S_{1,d}^2}\middle|\delta_i\right\}\right] \\
&= nE\left[\frac{1}{\pi_i}\varepsilon_i K_d\left(\frac{X_i - x}{h}\right)\frac{S_{2,d} - S_{1,d}(X_i - x)}{S_{0,d}S_{2,d} - S_{1,d}^2}\middle|\delta_i = 1\right]P(\delta_i = 1) \\
&= 0,
\end{aligned} \tag{13}$$

which imply that

$$E(\hat{a}_d(x)) = g(x) + \frac{1}{2}h^2 g''(x)\frac{\mu_{2,d}^2 - \mu_{1,d}\mu_{3,d}}{\mu_{0,d}\mu_{2,d} - \mu_{1,d}^2} + o(h^2).$$

Clearly, the bias of $\hat{a}_d(x)$ is

$$\text{bias}(\hat{a}_d(x)) = \frac{1}{2}h^2 g''(x)\frac{\mu_{2,d}^2 - \mu_{1,d}\mu_{3,d}}{\mu_{0,d}\mu_{2,d} - \mu_{1,d}^2} + o(h^2). \tag{14}$$

Next, we calculate the asymptotic variance of the estimator. According to (12) and (13), one has

$$\begin{aligned}\text{Var}(\hat{a}_d(x)) &= \text{Var}[g(x) + A_1 + A_2 + o(h^2)] \\ &= \text{Var}(A_1) + \text{Var}(A_2) + \text{Cov}(A_1, A_2) + o(h^2).\end{aligned}$$

It is easy to see that $\text{Var}(A_1) = o(1)$ and $\text{Cov}(A_1, A_2) = o(1)$. For $\text{Var}(A_2)$, we have the following expression

$$\begin{aligned}\text{Var}(A_2) &= \text{Var}\left[\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}\varepsilon_i K_d\left(\frac{X_i - x}{h}\right)\frac{S_{2,d} - S_{1,d}(X_i - x)}{S_{0,d}S_{2,d} - S_{1,d}^2}\right] \\ &= E\left[\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}\varepsilon_i K_d\left(\frac{X_i - x}{h}\right)\frac{S_{2,d} - S_{1,d}(X_i - x)}{S_{0,d}S_{2,d} - S_{1,d}^2}\right]^2 \\ &= nE\left[\frac{\delta_i^2}{\pi_i^2}\varepsilon_i^2 K_d^2\left(\frac{X_i - x}{h}\right)\left(\frac{S_{2,d} - S_{1,d}(X_i - x)}{S_{0,d}S_{2,d} - S_{1,d}^2}\right)^2\right] \\ &= nE\left[E\left\{\frac{\delta_i^2}{\pi_i^2}\varepsilon_i^2 K_d^2\left(\frac{X_i - x}{h}\right)\left(\frac{S_{2,d} - S_{1,d}(X_i - x)}{S_{0,d}S_{2,d} - S_{1,d}^2}\right)^2\Big|\delta_i\right\}\right] \\ &= nE\left[\frac{1}{\pi_1^2}\varepsilon_1^2 K_d^2\left(\frac{X_1 - x}{h}\right)\left(\frac{S_{2,d} - S_{1,d}(X_1 - x)}{S_{0,d}S_{2,d} - S_{1,d}^2}\right)^2\Big|\delta_1 = 1\right]P(\delta_1 = 1) \\ &= nP(\delta_1 = 1)\iint\frac{1}{\pi^2(g(z) + \varepsilon)}\varepsilon^2 K_d^2\left(\frac{z - x}{h}\right)\left(\frac{S_{2,d} - S_{1,d}(z - x)}{S_{0,d}S_{2,d} - S_{1,d}^2}\right)^2 f_{X,\varepsilon|\delta=1}(z,\varepsilon)\,\mathrm{d}z\,\mathrm{d}\varepsilon \\ &= nhP(\delta_1 = 1)\iint\frac{1}{\pi^2(g(x + th) + \varepsilon)}\varepsilon^2 K_d^2(t)\left(\frac{S_{2,d} - S_{1,d}th}{S_{0,d}S_{2,d} - S_{1,d}^2}\right)^2 f_{X,\varepsilon|\delta=1}(x + th,\varepsilon)\,\mathrm{d}t\,\mathrm{d}\varepsilon \\ &= \frac{1}{nhf_X^2(x)}P(\delta_1 = 1)\int K_d^2(t)\left(\frac{\mu_{2,d} - \mu_{1,d}t}{\mu_{0,d}\mu_{2,d} - \mu_{1,d}^2}\right)^2\mathrm{d}t\int\frac{1}{\pi^2(g(x) + \varepsilon)}\varepsilon^2 f_{X,\varepsilon|\delta=1}(x,\varepsilon)\,\mathrm{d}\varepsilon(1 + o(1)) \\ &= \frac{1}{nhf_X^2(x)}P(\delta_1 = 1)V_d S(x)(1 + o(1)).\end{aligned}$$

Therefore, one can obtain

$$\text{Var}(\hat{a}_d(x)) = \frac{1}{nhf_X^2(x)}P(\delta_1 = 1)V_d S(x)(1 + o(1)).$$

Hence, it together with (14), we can obtain the result (i) of Theorem 1

(ii). Suppose $x \in D_{2,l}$, let $x = s_j + uh, u \in (-\tau, 0)$. The left estimator of $g(\cdot)$ has the same bias

and variances shown before at any point $x \in D_{2,l}$. Meanwhile, the right estimator $\hat{a}_r(x)$ is given by

$$
\begin{aligned}
\hat{a}_r(x) =& \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} Y_i K_r\left(\frac{X_i - x}{h}\right) \frac{S_{2,r} - S_{1,r}(X_i - x)}{S_{0,r}S_{2,r} - S_{1,r}^2} \\
=& \sum_{X_i < s_j} \frac{\delta_i}{\pi_i} (g(X_i)) K_r\left(\frac{X_i - x}{h}\right) \frac{S_{2,r} - S_{1,r}(X_i - x)}{S_{0,r}S_{2,r} - S_{1,r}^2} \\
&+ \sum_{X_i \geq s_j} \frac{\delta_i}{\pi_i} (g(X_i) + d_j) K_r\left(\frac{X_i - x}{h}\right) \frac{S_{2,r} - S_{1,r}(X_i - x)}{S_{0,r}S_{2,r} - S_{1,r}^2} \\
&+ \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} \varepsilon_i K_r\left(\frac{X_i - x}{h}\right) \frac{S_{2,r} - S_{1,r}(X_i - x)}{S_{0,r}S_{2,r} - S_{1,r}^2} \\
=& \sum_{X_i < s_j} \frac{\delta_i}{\pi_i} (g(s_j-) + o_p(1)) K_r\left(\frac{X_i - x}{h}\right) \frac{S_{2,r} - S_{1,r}(X_i - x)}{S_{0,r}S_{2,r} - S_{1,r}^2} \\
&+ \sum_{X_i \geq s_j} \frac{\delta_i}{\pi_i} (g(s_j-) + d_j + o_p(1)) K_r\left(\frac{X_i - x}{h}\right) \frac{S_{2,r} - S_{1,r}(X_i - x)}{S_{0,r}S_{2,r} - S_{1,r}^2} + o_p(1) \\
=& (g(s_j-) + o_p(1)) \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} K_r\left(\frac{X_i - x}{h}\right) \frac{S_{2,r} - S_{1,r}(X_i - x)}{S_{0,r}S_{2,r} - S_{1,r}^2} \\
&+ \sum_{X_i \geq s_j} \frac{\delta_i}{\pi_i} d_j K_r\left(\frac{X_i - x}{h}\right) \frac{S_{2,r} - S_{1,r}(X_i - x)}{S_{0,r}S_{2,r} - S_{1,r}^2} + o_p(1) \\
=& g(s_j-) + d_j \int_{|u|}^{\tau} K_r(t) \frac{\mu_{2,r} - \mu_{1,r} t}{\mu_{0,r}\mu_{2,r} - \mu_{1,r}^2} \, \mathrm{d}t + o_p(1).
\end{aligned} \tag{15}
$$

So the bias of $\hat{a}_r(x)$ is

$$
\operatorname{bias}(\hat{a}_r(x)) = d_j \int_{|u|}^{\tau} K_r(t) \frac{\mu_{2,r} - \mu_{1,r} t}{\mu_{0,r}\mu_{2,r} - \mu_{1,r}^2} \, \mathrm{d}t + o(1).
$$

Similarly, the expression of the centered estimator $(\hat{a}_c(x))$ can be obtained by using the centered kernel and applying that $\mu_{0,c} = 1$ and $\mu_{1,c} = 0$.

(iii). The third part of Theorem 1 can be proved in the same way as (ii).

□

A.2 Proof of Theorem 2

Proof. From the definition of WRMS, it follows that

$$
\mathrm{WRMS}_d(x) = \frac{\frac{1}{nh}\sum_{i=1}^{n} \frac{\delta_i}{\pi_i} [Y_i - \hat{a}_d(x) - \hat{b}_d(x)(X_i - x)]^2 K_d\left(\frac{X_i - X}{h}\right)}{\frac{1}{nh}\sum_{i=1}^{n} \frac{\delta_i}{\pi_i} K_d\left(\frac{X_i - X}{h}\right)}. \tag{16}
$$

According to (11), the denominator of (16) can be written as

$$
\frac{1}{nh}\sum_{i=1}^{n} \frac{\delta_i}{\pi_i} K_d\left(\frac{X_i - x}{h}\right) = \mu_{0,d} f_X(x) + o_p(1). \tag{17}
$$

For the numerator of (16), one has

$$\begin{aligned}
&\frac{1}{nh}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}[Y_i-\hat{a}_d(x)-\hat{b}_d(x)(X_i-x)]^2K_d\left(\frac{X_i-x}{h}\right)\\
&=\frac{1}{nh}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}[g(X_i)+\varepsilon_i-\hat{a}_d(x)-\hat{b}_d(x)(X_i-x)]^2K_d\left(\frac{X_i-x}{h}\right)\\
&=\frac{1}{nh}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}[g(X_i)-\hat{a}_d(x)-\hat{b}_d(x)(X_i-x)]^2K_d\left(\frac{X_i-x}{h}\right)+\frac{1}{nh}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}\varepsilon_i^2K_d\left(\frac{X_i-x}{h}\right)\\
&\quad+\frac{2}{nh}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}[g(X_i)-\hat{a}_d(x)-\hat{b}_d(x)(X_i-x)]\varepsilon_iK_d\left(\frac{X_i-x}{h}\right)\\
&\triangleq I_1+I_2+I_3.
\end{aligned}\tag{18}$$

(i) Suppose $u\in D_1$, using the similar derivations to those in the proof of Theorem 1, it can be obtained that

$$I_2=\sigma^2\mu_{0,d}f_X(x)+o_p(1).$$

For I_3, applying the Taylor's expansion, it is clear

$$I_3=I_{31}+I_{32}+I_{33},$$

where

$$\begin{aligned}
I_{31}&=\frac{2}{nh}(g(x)-\hat{a}_d(x))\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}\varepsilon_iK_d\left(\frac{X_i-x}{h}\right),\\
I_{32}&=\frac{2}{nh}(g'(x)-\hat{b}_d(x))\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}\varepsilon_iK_d\left(\frac{X_i-x}{h}\right)(X_i-x),\\
I_{33}&=\frac{2}{nh}o(h)\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}\varepsilon_iK_d\left(\frac{X_i-x}{h}\right).
\end{aligned}$$

Furthermore, from Theorem 2.8 and Theorem 2.9 of Li and Racine (2007), one can get $\hat{a}_d(x)-g(x)=O_p\left(h^2+n(nh)^{-1/2}\right)=o_p(1)$ and $\hat{b}_d(x)-g'(x)=O_p\left(h^2+(nh^3)^{-1/2}\right)=o_p(h^{-1})$, we get $I_{31}=o_p(1)$, $I_{32}=o_p(1)$ and $I_{33}=o_p(1)$. It imply that

$$I_3=o_p(1).$$

Similarly, it can be proved

$$I_1=o_p(1).$$

It is easily seen from (17) and (18) that

$$\mathrm{WRMS}_d(x)=\sigma^2+o_p(1).$$

(ii) Suppose $x\in D_{2,l}$, let $x=s_j+uh, u\in(-\tau,0)$. $\mathrm{WRMS}_l(x)$ can be proved in the same way as

(i). For $\mathrm{WRMS}_r(u)$, the right sided estimator of the first-order derivation of $g(\cdot)$ is given by

$$\begin{aligned}
\hat{b}_r(x) =& \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} Y_i K_r\left(\frac{X_i - X}{h}\right) \frac{S_{0,r}(X_i - x) - S_{1,r}}{S_{0,r}S_{2,r} - S_{1,r}^2} \\
=& \sum_{X_i < s_j} \frac{\delta_i}{\pi_i} (g(X_i)) K_r\left(\frac{X_i - x}{h}\right) \frac{S_{0,r}(X_i - x) - S_{1,r}}{S_{0,r}S_{2,r} - S_{1,r}^2} \\
&+ \sum_{X_i \geq s_j} \frac{\delta_i}{\pi_i} (g(X_i)) K_r\left(\frac{X_i - x}{h}\right) \frac{S_{0,r}(X_i - x) - S_{1,r}}{S_{0,r}S_{2,r} - S_{1,r}^2} \\
&+ \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} \varepsilon_i K_r\left(\frac{X_i - x}{h}\right) \frac{S_{0,r}(X_i - x) - S_{1,r}}{S_{0,r}S_{2,r} - S_{1,r}^2} \\
=& \sum_{X_i < s_j} \frac{\delta_i}{\pi_i} (g(s_j-) + o_p(1)) K_r\left(\frac{X_i - x}{h}\right) \frac{S_{0,r}(X_i - x) - S_{1,r}}{S_{0,r}S_{2,r} - S_{1,r}^2} \\
&+ \sum_{X_i \geq s_j} \frac{\delta_i}{\pi_i} (g(s_j-) + d_j + o_p(1)) K_r\left(\frac{X_i - x}{h}\right) \frac{S_{0,r}(X_i - x) - S_{1,r}}{S_{0,r}S_{2,r} - S_{1,r}^2} + o_p(h^{-1}) \\
=& (g(s_j-) + o_p(1) \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} K_r\left(\frac{X_i - x}{h}\right) \frac{S_{0,r}(X_i - x) - S_{1,r}}{S_{0,r}S_{2,r} - S_{1,r}^2} \\
&+ \sum_{X_i \geq s_j} \frac{\delta_i}{\pi_i} d_j K_r\left(\frac{X_i - x}{h}\right) \frac{S_{0,r}(X_i - x) - S_{1,r}}{S_{0,r}S_{2,r} - S_{1,r}^2} + o_p(h^{-1}) \\
=& \frac{1}{h} d_j \int_{|u|}^{\tau} K_r(t) \frac{\mu_{0,r}t - \mu_{1,r}}{\mu_{0,r}\mu_{2,r} - \mu_{1,r}^2} \,\mathrm{d}t + o_p(h^{-1}).
\end{aligned} \tag{19}$$

Therefore, using (15) and (19), the expression of I_1 for the right-sided estimator is

$$\begin{aligned}
I_1 =& \frac{1}{nh} \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} [g(X_i) - \hat{a}_r(x) - \hat{b}_r(x)(X_i - x)]^2 K_r\left(\frac{X_i - x}{h}\right) \\
=& \frac{1}{nh} \sum_{X_i \geq s_j} \frac{\delta_i}{\pi_i} \Bigg[g(X_i) - g(s_j-) - d_j \int_{-u}^{\tau} K_r(t) \frac{\mu_{2,r} - \mu_{1,r}t}{\mu_{0,r}\mu_{2,r} - \mu_{1,r}^2} \,\mathrm{d}t \\
& - \frac{1}{h} d_j \int_{-u}^{\tau} \frac{\mu_{0,r}t - \mu_{1,r}}{\mu_{0,r}\mu_{2,r} - \mu_{1,r}^2} K_r(t) \,\mathrm{d}t (X_i - x) + o(1) \Bigg]^2 K_r\left(\frac{X_i - x}{h}\right) \\
& + \frac{1}{nh} \sum_{X_i < s_j} \frac{\delta_i}{\pi_i} \Bigg[g(X_i) - g(s_j-) - d_j \int_{-u}^{\tau} K_r(t) \frac{\mu_{2,r} - \mu_{1,r}t}{\mu_{0,r}\mu_{2,r} - \mu_{1,r}^2} \,\mathrm{d}t \\
& - \frac{1}{h} d_q \int_{-u}^{\tau} \frac{\mu_{0,r}t - \mu_{1,r}}{\mu_{0,r}\mu_{2,r} - \mu_{1,r}^2} K_r(t) \,\mathrm{d}t (X_i - x) + o(1) \Bigg]^2 K_r\left(\frac{X_i - x}{h}\right) + o_p(1) \\
=& f_X(x) \int_{-u}^{\tau} \left[d_j \int_{-\tau}^{-u} \frac{\mu_{2,r} - \mu_{1,r}t}{\mu_{0,r}\mu_{2,r} - \mu_{1,r}^2} K_r(t) \,\mathrm{d}t - z d_j \int_{-u}^{\tau} \frac{\mu_{0,r}t - \mu_{1,r}}{\mu_{0,r}\mu_{2,r} - \mu_{1,r}^2} K_r(t) \,\mathrm{d}t \right]^2 K_r(z) \,\mathrm{d}z \\
& + f_X(x) \int_{-\tau}^{-u} \left[d_j \int_{-u}^{\tau} \frac{\mu_{2,r} - \mu_{1,r}t}{\mu_{0,r}\mu_{2,r} - \mu_{1,r}^2} K_r(t) \,\mathrm{d}t + z d_j \int_{-u}^{\tau} \frac{\mu_{0,r}t - \mu_{1,r}}{\mu_{0,r}\mu_{2,r} - \mu_{1,r}^2} K_r(t) \,\mathrm{d}t \right]^2 K_r(z) \,\mathrm{d}z (1 + o_p(1)) \\
=& f_X(x) d_j C_{u,r}^2 + o_p(1).
\end{aligned}$$

Furthermore, similar to the proof of (i), $I_2 = \sigma^2 \mu_{0,r} f(x) + o_p(1)$ and $I_3 = o_p(1)$. Thus (ii) of Theorem 2 is proved.

(iii) Suppose $x \in D_{2,r}$, the third part can be obtained in the same way. □

A.3 Proof of Theorem 3

Proof. For $x \in D_1$, similar to the proof of Theorem 1, by the central limit theorem, we have

$$\sqrt{nh}\left[\hat{a}_d(x) - g(x) - \frac{1}{2}h^2 g''(x)B_d\right] \xrightarrow{D} N\left(0, f_X^{-2}(x)P(\delta_1 = 1)S(x)V_d\right). \tag{20}$$

In addition, it follows from Theorem 1 that for the left side estimator $\hat{a}_l(x)$, (20) holds for $x \in D_1 \cup D_{2,l}$ and for the right side estimator $\hat{a}_r(x)$, (20) holds for $x \in D_1 \cup D_{2,r}$.

For any $x \in [0,1]$, the resulting estimator of $g(\cdot)$ can be rewritten as

$$\hat{g}(x) = \hat{a}_c(x)I(D_1(x)) + \hat{a}_l(x)I(D_{2,l}(x)) + \hat{a}_r(x)I(D_{2,r}(x)).$$

Note that D_1, $D_{2,l}$ and $D_{2,r}$ are mutually exclusive and $I(D_1(x)) + I(D_{2,l}(x)) + I(D_{2,r}(x)) = 1$.

For any $x \in D_1$, it can be seen that $\text{diff}(x) \to 0$ and $\lambda \to 0$ as $n \to \infty$ from Theorem 2. It means that when $x \in D_1$, $\hat{g}(x) = \hat{a}_c(x)$ a.s..

For any $x \in D_{2,l}$, that is, $u = s_j + uh$, $u \in (-\tau, 0)$. From the second part of Theorem 2,

$$\text{diff}(x) = \max\left\{d_j^2 C_{u,c}^2 + R_{c,2}(x) - R_{l,2}(x), d_j^2(C_{u,c}^2 - C_{u,r}^2) + R_{c,2}(x) - R_{r,2}(x)\right\}.$$

Since

$$\lim_{n\to\infty}[d_j^2 C_{u,c}^2 + R_{c,2}(x) - R_{l,2}(x)] = d_j^2 C_{u,c}^2,$$

and

$$\lim_{n\to\infty}[d_j^2(C_{u,c}^2 - C_{u,r}^2) + R_{c,2}(x) - R_{r,2}(x)] = d_j^2(C_{u,c}^2 - C_{u,r}^2),$$

which implies that

$$\lim_{n\to\infty}\text{diff}(x) = \max\left\{d_j^2 C_{u,c}^2, d_j^2(C_{u,c}^2 - C_{u,r}^2)\right\} = d_j^2 C_{u,c}^2,$$

and by $0 < \lambda < d_j^2 C_{u,c}^2$, so $I(D_1(x)) = 0$ as $n \to \infty$. Note that

$$\text{WRMS}_r(x) - \text{WRMS}_l(x) = d_j^2 C_{u,r}^2 + R_{r,2}(x) - R_{l,2}(x) \to d_j^2 C_{u,r}^2 > 0,$$

so $I(D_{2,r}(x)) = 0$ and $I(D_{2,l}(x)) = 1$ a.s., i.e. $\hat{g}(x) = \hat{a}_l(x)$.

For $x \in D_{2,r}$, similarly, we have $\hat{g}(x) = \hat{a}_r(x)$.

It is clear that $\hat{a}_c(x)$, $\hat{a}_l(x)$, $\hat{a}_r(x)$ are asymptotically normal in D_1, $D_{2,l}$ and $D_{2,r}$ respectively, thus Theorem 3 is proved.

□

A.4 Proof of Theorem 4

Proof. To prove this theorem, we first prove prove the following equations

$$\sup_{x\in[0,1]}|\hat{a}_d(x) - \hat{a}_d(x,\hat{\pi})| = O_p(n^{-1/2}). \tag{21}$$

Similar to the proof of Theorem 3, when $x \in D_1$, $\hat{g}(x) = \hat{a}_c(x)$ and $\hat{g}(x,\hat{\pi}) = \hat{a}_c(x,\hat{\pi})$. Thus (21) holds for two center estimators $\hat{a}_c(x)$ and $\hat{a}_c(x,\hat{\pi})$. The proof is presented below.

Since $\pi(y)$ is assumed to follow a parametric model $\pi(y,\alpha)$ and has bounded first order partial derivative with respect to α, it is easy to show that $\sup_{y\in R}|\pi(y) - \hat{\pi}(y)| = \sup_{y\in R}|\pi(y,\alpha) - \hat{\pi}(y,\hat{\alpha})| = O_p(n^{-1/2})$, where $\hat{\alpha}$ is a root-n consistent estimator of α. This together with (11) implies that there exists some constant $M > 0$ such that

$$\begin{aligned}\sup_{x\in[0,1]}\left|\frac{1}{nh}(S_{j,c} - \hat{S}_{j,c})\right| &= \sup_{x\in[0,1]}\left|\frac{1}{nh}\sum_{i=1}^{n}\left(\frac{\delta_i}{\pi_i} - \frac{\delta_i}{\hat{\pi}_i}\right)K_c\left(\frac{X_i - x}{h}\right)(X_i - x)^j\right| \\ &\le dh^j \sup_{1\le i\le n}|\pi_i - \hat{\pi}_i| \sup_{x\in[0,1]}\left|\frac{1}{nh}S_{0,c}\right| \\ &= O_p(n^{-1/2}).\end{aligned} \tag{22}$$

Next, from lemma 5 of Li et al. (2021), one has

$$\sup_{x\in[0,1]}\left|M_{l,c}(x)-\hat{M}_{l,c}(x)\right|=O_p(n^{-1/2}),\quad l=1,2, \tag{23}$$

where

$$M_{l,c}(x)=\frac{1}{nh}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}K_c\left(\frac{X_i-x}{h}\right)(X_i-x)^l\left(\frac{1}{2}g''(x)(X_i-x)^2+\varepsilon_i+o(h^2)\right),$$
$$\hat{M}_{l,c}(x)=\frac{1}{nh}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i}K_c\left(\frac{X_i-x}{h}\right)(X_i-x)^l\left(\frac{1}{2}g''(x)(X_i-x)^2+\varepsilon_i+o(h^2)\right).$$

Meanwhile, from (10), one can obtain that

$$\hat{a}_c(x)-g(x)=\boldsymbol{e}_0^\top\begin{pmatrix}(nh)^{-1}S_{0,c} & (nh)^{-1}S_{1,c}\\ (nh)^{-1}S_{1,c} & (nh)^{-1}S_{2,c}\end{pmatrix}^{-1}\begin{pmatrix}M_{0,c}\\ M_{1,c}\end{pmatrix},$$

where $\boldsymbol{e}_0=(1,0)^\top$. Similarly, for $\hat{a}_d(x,\hat{\pi})$,

$$\hat{a}_c(x,\hat{\pi})-g(x)=\boldsymbol{e}_0^\top\begin{pmatrix}(nh)^{-1}\hat{S}_{0,c} & (nh)^{-1}\hat{S}_{1,c}\\ (nh)^{-1}\hat{S}_{1,c} & (nh)^{-1}\hat{S}_{2,c}\end{pmatrix}^{-1}\begin{pmatrix}\hat{M}_{0,c}\\ \hat{M}_{1,c}\end{pmatrix}$$

Therefore, one has

$$\begin{aligned}\hat{a}_c(x)-\hat{a}_c(x,\hat{\pi})=&\boldsymbol{e}_0^\top\begin{pmatrix}(nh)^{-1}S_{0,c} & (nh)^{-1}S_{1,c}\\ (nh)^{-1}S_{1,c} & (nh)^{-1}S_{2,c}\end{pmatrix}^{-1}\begin{pmatrix}M_{0,c}\\ M_{1,c}\end{pmatrix}\\ &-\boldsymbol{e}_0^\top\begin{pmatrix}(nh)^{-1}\hat{S}_{0,c} & (nh)^{-1}\hat{S}_{1,c}\\ (nh)^{-1}\hat{S}_{1,c} & (nh)^{-1}\hat{S}_{2,c}\end{pmatrix}^{-1}\begin{pmatrix}\hat{M}_{0,c}\\ \hat{M}_{1,c}\end{pmatrix}\end{aligned}$$

By (22) and (23), it can be seen that

$$\sup_{x\in[0,1]}|\hat{a}_c(x)-\hat{a}_c(x,\hat{\pi})|=O_p(n^{-1/2}).$$

Similarly, when $x\in D_{2,l}$, $\hat{g}(x)=\hat{a}_l(x)$ and $\hat{g}(x,\hat{\pi})=\hat{a}_l(x,\hat{\pi})$. Thus (21) holds for two left estimators $\hat{a}_l(x)$ and $\hat{a}_l(x,\hat{\pi})$. When $x\in D_{2,r}$, (21) holds for two left estimators $\hat{a}_r(x)$ and $\hat{a}_r(x,\hat{\pi})$.

When $d=c,l,r$, it is clear that (21) holds in D_1, $D_{2,l}$ and $D_{2,r}$ respectively, thus Theorem 4 is proved. □