ESAIM: PS 30 (2026) 242–284
https://doi.org/10.1051/ps/2026002

ESAIM: Probability and Statistics
www.esaim-ps.org

JUMP-PRESERVING ESTIMATION AND STRUCTURAL BREAK DETECTION IN NONPARAMETRIC REGRESSION MODELS WITH MISSING COVARIATES

Abbes Rabhi[1,*], Anis Allal[2] and Nadia Kadiri[3]

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. A number of methods exist for estimating discontinuous curve, most of which are based on complete data, which is unrealistic in many practical situations. In this paper, we consider estimating discontinuous nonparametric model with covariate with missing values. 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 finite sample performance of this method is efficient and reliable.

Mathematics Subject Classification. 62G08, 60J75, 62F12.

Received January 10, 2025. Accepted January 18, 2026.

1. Introduction

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

$$Y = g(X) + \varepsilon, \tag{1.1}$$

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 comes 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

Keywords and phrases: Nonparametric model, local linear kernel smoothing, jump-preserving estimation, inverse probability weighted, missing data.

[1] University Djillali LIABES of Sidi Bel Abbes, Algeria.
[2] EEDIS Laboratory, Computer Science Departement, University Djillali LIABES of Sidi Bel Abbes, Algeria.
[3] National Higher School of Telecommunications and Information and Communication Technologies (ENSTTIC), Oran, Algeria.

* Corresponding author: rabhi_abbes@yahoo.fr

important branch in statistics, it has been well established as a useful data analytical tool. See, for example, the monographs Fan and Gijbels [1] for a large variety of interesting real data examples where applications of such methods have yielded analyzes essentially unobtainable by other techniques. A large body of literature exists on regression function estimation, such as kernel estimation (see Fan and Gijbels [1] and Claeskens and Van Keilegom [2]), local polynomial regression(see Xiao *et al.* [3] and El Ghouch and Genton [4]), spline smoothing (see Huang [5] and Wang and Yang [6]), 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 [7]). 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 [8]). In oceanography, the sea-level pressure in Bombay, India, has been found to have experienced an abrupt change in the early 1960s (Qiu [9]). 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 [10]. Therefore, the research on regression model with jumps is very necessary.

Nonparametric regression methods have become essential tools in analyzing complex data sets without relying on strict model assumptions. However, these techniques often assume that the regression function is smooth over the entire domain. This assumption may not be valid in applications such as economics, biostatistics, or climate modeling where abrupt changes (jumps) in the response variable occur.

Additional early contributions to jump regression include Qiu *et al.* [11], Müller [12], and Wu and Chu [13]. These works laid the groundwork for modern jump regression analysis and motivate our own methodology. They share a common structure in which jump points are first estimated and then the regression function is estimated on subintervals separated by the detected jumps and which extends this framework to account for covariates with missing values a setting not directly addressed in these earlier models. Their approaches are conceptually related to the framework adopted in this paper.

Several early contributions have addressed the estimation of jump regression functions. Notably, Qiu *et al.* [11], Müller [12], and Wu and Chu [13] proposed foundational methods that follow a similar general strategy: first estimating the jump locations, and then estimating the regression function separately on sub-intervals defined by these estimated jumps. Specifically, Qiu *et al.* [11] developed estimators tailored to piecewise-smooth regression functions, while Müller [12] addressed change-point detection within nonparametric regression. Wu and Chu [13] introduced kernel-type estimators for both the jump points and the function values at those points.

The jump-preserving estimator developed by Gijbels *et al.* [14] offers a locally adaptive method for capturing discontinuities. While the jump-preserving estimator itself was previously proposed by Gijbels *et al.* [14], the present contribution extends it to the context of covariates subject to missingness under the MAR assumption. The key novelty lies in adapting the selection weighted framework to appropriately handle incomplete covariate data, which significantly enhances the practical applicability of this class of estimators.

Moreover, the asymptotical properties of proposed estimators are presented. We emphasize that the selection probability $\pi(Y)$ is modeled using a logistic function. This allows for consistent estimation under a standard parametric assumption.

In recent decades, there are several research studies to fit curves with jump points. McDonald and Owen [15] 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 [16] 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 [9] 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.* [14] 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 [17] 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 [18] 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.* [19] 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 [20], Zhao *et al.* [21], Han *et al.* [22] and Wang *et al.* [23], 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 [24] and Little and Rubin [25] 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 [25]) 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 [26]), nonparametric kernel regression imputation (see Wang and Rao [27]), semi-parametric regression imputation (see Wang *et al.* [28]), 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 [29], Liang *et al.* [30] and Stute *et al.* [31], 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 [32]) 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 [33] discussed the inverse probability weighted estimation for general missing data problems. Wang [34] consider the partial linear model with the covariables missing at random using the inverse probability weighted approach. Seaman and White [35] reviewed inverse probability weighted for the missing data analysis.

Nonparametric regression method to deal with missing data was discussed relatively less. For covariates with missing values, Wang *et al.* [36] used IPW local linear regression in generalized linear model; Liang *et al.* [37] considered a nonparametric estimator in a partially linear model. For missing outcomes, Wang *et al.* [38] developed a doubly robust local linear estimator and Sun *et al.* [39] generalized it to the multiple robustness; Chen *et al.* [40] constructed a few local quasi-likelihood estimators; estimating equations with nonparametric imputed values were developed by Zhou *et al.* [41], and Wang *et al.* [42] considered the case where the covariate is

functional. Efromovich generalised the orthogonal series estimator in the cases of missing covariates (Efromovich [43]) and missing outcomes (Efromovich [44]).

For model (1.1), 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.* [45] studied the discontinuous nonparametric regression curve fitting when response is missing. Naturally, we are interested in the case of covariate with missing values. 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 covariate with missing values, we assume that X is missing at random (MAR) in the sense that

$$\pi(Y) := \mathbb{P}(\delta = 1|X, Y) = \mathbb{P}(\delta = 1|Y). \tag{1.2}$$

The MAR assumption implies that δ and X are conditionally independent given Y, that is, $\mathbb{P}(\delta = 1|X, Y) = \mathbb{P}(\delta = 1|Y)$. MAR is a common assumption for handling missing data and such assumption is also reasonable in many applications, see Little and Rubin [25].

This paper focuses on the estimation of discontinuous regression curve with covariate with missing values. 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 neighborhoods of jumps of the nonparametric function does not require prior estimation of 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 parametric. Moreover, the asymptotical properties of proposed estimators are presented. Numerical simulation indicates the performance of finite sample of this method is efficient.

To address this, we adopt an inverse probability weighting (IPW) approach, where the probability of observing the covariate given the response is modeled by a logistic regression model:

$$\pi(Y) := \mathbb{P}(\delta = 1 \mid Y) = \frac{\exp(\alpha_0 + \alpha_1 Y)}{1 + \exp(\alpha_0 + \alpha_1 Y)}.$$

This parametric specification of the selection probability enables stable and practical estimation while maintaining flexibility. Under the assumption that the covariate is missing at random (MAR), we develop a jump-preserving nonparametric regression estimator using local linear kernel smoothing. Our procedure effectively adapts to both smooth and discontinuous regions of the function, without requiring prior knowledge of the number or locations of jumps. Moreover, the estimator achieves oracally efficient properties and maintains consistency across the entire design space. Although the core jump-preserving estimation method was introduced by Gijbels *et al.* [14], our contribution lies in extending this framework to settings where the covariates are subject to missingness at random (MAR). To our knowledge, such an extension-along with the analysis of oracle efficiency under estimated selection probabilities-has not been previously addressed in the literature.

While the jump-preserving estimator itself was previously proposed by Gijbels *et al.* [14], the present contribution extends it to the context of covariates subject to missingness under the MAR assumption. The key novelty lies in adapting the selection-weighted framework to appropriately handle incomplete covariate data, which significantly enhances the practical applicability of this class of estimators.

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 A.

2. Methodology and main results

In this section, we will consider the curve fitting for the nonparametric regression model (1.1) with unknown jumps and missing data in the 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 Sction 2.1, we first review the local linear jump regression curve fitting method with complete data. Then we extend this method into the context of missing data by virtue of inverse probability weighting method.

2.1. Review the jump-preserving method

Suppose that $\{(X_i, Y_i), 1 \leq i \leq n\}$ is an independent and identically distributed (i.i.d.) random sample observed fully from (1.1). When $g(\cdot)$ is a continuous function, Fan and Gijbels [1] 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 neighborhood 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:

$$\sum_{i=1}^{n} \left(Y_i - a - b(X_i - x)\right)^2 K_h\left(X_i - x\right), \tag{2.1}$$

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 (2.1) for a is defined as the conventional local linear kernel estimator of $g(\cdot)$ where $g(\cdot)$ has jumps. 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 the curve function (regression function), and it is known that the fitted function based on conventional local linear kernel methods is not statistically consistent at jump positions. To deal with this problem, some jump-preserving estimation methods were proposed. Now we briefly review techniques given by Qiu [9] and Gijbels *et al.* [14]. For fixed $x \in [0,1]$, the following three local linear estimators are defined by for $d = c, l, r$;

$$\arg\min_{a,b} \sum_{i=1}^{n} \left(Y_i - a - b(X_i - x)\right)^2 K_d\left(\frac{X_i - x}{h}\right), \tag{2.2}$$

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

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

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

Let $\{\widehat{a}_d(x), \widehat{b}_d(x),\ d = c, r, l\}$ denote the solutions of (2.2). Obviously the estimators $\widehat{a}_c(x)$ are the usual local linear estimators of $g(x)$, based on data in the neighborhood $[x - \tau h, x + \tau h]$ of x, and $\widehat{a}_l(x)$ and $\widehat{a}_\tau(x)$ 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.* [14], 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.* [14], however, it can be found that only $\widehat{a}_l(x)$ is consistent, nevertheless, $\widehat{a}_c(x)$ and $\widehat{a}_r(x)$ are not consistent at any point in the neighborhood $[s_j - \tau h, s_j)$. 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 $\widehat{a}_l(x)$ and $\widehat{a}_r(x)$ provides a good estimator of $g(\cdot)$. Therefore, it needs to choose one from three estimators as an estimator of $g(\cdot)$ in such case. Qiu [9] suggested the following jump-preserving estimate of $g(\cdot)$ in such case.

Qiu [9] suggested the following jump-preserving estimate of $g(\cdot)$

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

where $I^*(t)$ is defined by $I^*(t) = 1$ if $t > 0$, $I^*(t) = 1/2$ if $t = 0$ and $I^*(t) = 0$ 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

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

Basically, $\widehat{g}_Q(x)$ is defined by one of $\widehat{a}_l(x)$ and $\widehat{a}_r(x)$ with the smaller RSS value. Qiu [9] proved that $\widehat{g}_Q(x)$ is a consistent estimator of $g(x)$ in the entire design interval.

In practice, it appears that $\widehat{g}_Q(x)$ 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.* [14] introduced the conventional estimator $\widehat{a}_c(x)$, and proposed the following jump-preserving estimate of $g(\cdot)$

$$\widehat{g}_G(x) = \begin{cases} \widehat{a}_c(x), & \text{if } \mathrm{diff}(x) \leq \lambda \\ \widehat{a}_l(x), & \text{if } \mathrm{diff}(x) > \lambda \text{ and } \mathrm{WRMS}_l(x) < \mathrm{WRMS}_r(x) \\ \widehat{a}_r(x), & \text{if } \mathrm{diff}(x) > \lambda \text{ and } \mathrm{WRMS}_l(x) > \mathrm{WRMS}_r(x) \\ (\widehat{a}_l(x) + \widehat{a}_r(x))/2, & \text{if } \mathrm{diff}(x) > \lambda \text{ and } \mathrm{WRMS}_l(x) = \mathrm{WRMS}_r(x), \end{cases} \tag{2.3}$$

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

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

to evaluate the quality of the three local linear fits. Gijbels *et al.* [14] proved the (uniform) strong consistency of $\widehat{g}_G(x)$.

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 (2.2) by using only fully observed (X_i, Y_i) can result in a biased estimator for $g(\cdot)$. Let $\pi_i = \pi(Yi) = \mathbb{P}(\delta_i = 1|Y_i)$ is the selection probability by our MAR assumption. In this section, 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 section.

To estimate the latent discontinuous function $g(\cdot)$ incorporating missing data, we combine jump-preserving method in Gijbels *et al.* [14] and the inverse probability weighted techniques (see Horvitz and Thompson [32]). Specifically, for fixed $x \in [0, 1]$, the following three inverse probability weighted local linear estimators are proposed

$$\arg\min_{a,b} \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} \left(Y_i - a - b(X_i - x)\right)^2 K_d\left(\frac{X_i - x}{h}\right), \tag{2.4}$$

for $d = c, l, r$. Since the variables are subject to missingness, only fully observed cases $\delta_i = 1$ contribute to the objective function (2.4), 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 (2.4) are denoted as $\widehat{a}_d(x)$ and $\widehat{b}_d(x)$, $d = c, l, r$, respectively. By some routine algebraic manipulations, $\widehat{a}_d(x)$ have nice and simple expressions:

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

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 $\widehat{a}_l(x)$ and $\widehat{a}_r(x)$ 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 $\widehat{a}_c(x)$ are the usual local linear estimator of $g(x)$, based on data in the neighborhood $[x - \tau h, x + \tau h]$ of x. Similarly with the approach in Gijbels *et al.* [14], we propose the following compromising estimator

$$\widehat{g}(x) = \begin{cases} \widehat{a}_c(x), & \text{if } \mathrm{diff}(x) \le \lambda \\ \widehat{a}_l(x), & \text{if } \mathrm{diff}(x) > \lambda \text{ and } \mathrm{WRMS}_l(x) < \mathrm{WRMS}_r(x) \\ \widehat{a}_r(x), & \text{if } \mathrm{diff}(x) > \lambda \text{ and } \mathrm{WRMS}_l(x) > \mathrm{WRMS}_r(x) \\ (\widehat{a}_l(x) + \widehat{a}_r(x))/2, & \text{if } \mathrm{diff}(x) > \lambda \text{ and } \mathrm{WRMS}_l(x) = \mathrm{WRMS}_r(x). \end{cases} \tag{2.5}$$

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

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

for $d = c, l, r$. In (2.5),

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

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, diff(x) is a natural jump detection criterion. If x is a jump point, then diff(x) would be relatively large. By (2.5), 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 $\widehat{g}(x)$. Their proofs are given in the Appendix B. To proceed, we introduce some notations. Let $\mu_{k,d} = \int t^k K_d(t)\mathrm{d}t$ dor $d = c, l, r$. 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 $\mathbb{E}(\varepsilon^4) > \infty$. Moreover, $\int \varepsilon^2 f_{X,\varepsilon|\delta=1}(x,\varepsilon)d\varepsilon$ 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 at 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 probability 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) $\inf_{1\le i\le n}\{\pi(Y_i)\} \ge C > 0$ with probability one for some constant C.

Remark 2.1. 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 conditions in kernel estimation method; condition (A6) ensures the effectiveness of inverse probability weighted.

(i) Although Assumption (A4) requires a kernel with bounded support, a Gaussian kernel was used in our simulations for its smoothness and practical performance. In practice, the Gaussian kernel provides good empirical results, even if its support is unbounded. Nonetheless, alternative bounded kernels such as the Epanechnikov kernel could be used to fully comply with this assumption.
(ii) Assumption (A5), which requires $nh^3 \to 0$, implies a bandwidth of order $n^{-1/3}$, which may not be optimal in a standard nonparametric setting. However, this condition is imposed here to ensure bias control and asymptotic validity under jump-preserving estimation and MAR weighting. Whether this condition can be relaxed to achieve a more optimal convergence rate is an open question that deserves further investigation.
(iii) Assumption (A6) ensures that the selection probabilities remain bounded away from zero, which is necessary for the stability of inverse probability weighting. While this may appear restrictive, it is a standard condition in the missing data literature (*e.g.*, Seaman and White, 2013) and is reasonable in many real-world scenarios where missingness is not extreme. In cases where the probability of observing data becomes very small, regularization techniques or trimming strategies may be required to maintain estimator performance.

Theorem 2.2. *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 $x \in D_1$,

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

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

$$\begin{aligned}
\mathrm{MSE}(\widehat{a}_l(x)) &= \left(\frac{1}{2}h^2 g''(s_j-)B_l\right)^2 + \frac{1}{nhf_X^2(x)}\mathbb{P}(\delta=1)V_l S(x) + o\left(h^4 + \frac{1}{nh}\right),\\
\mathrm{MSE}(\widehat{a}_r(x)) &= \left(d_j \int_{|u|}^{\tau} K_r(t)\frac{\mu_{2,r} - \mu_{1,r}t}{\mu_{0,r}\mu_{2,r} - \mu_{1,r}^2}\mathrm{d}t\right)^2 + \frac{1}{nhf_X^2(x)}\mathbb{P}(\delta=1)V_r S(x) + o\left(\frac{1}{nh}\right),\\
\mathrm{MSE}(\widehat{a}_c(x)) &= \left(d_j \int_{|u|}^{\tau} K_c(t)\mathrm{d}t\right)^2 + \frac{1}{nhf_X^2(x)}\mathbb{P}(\delta=1)V_c S(x) + o\left(\frac{1}{nh}\right).
\end{aligned}$$

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

$$\begin{aligned}
\mathrm{MSE}(\widehat{a}_l(x)) &= \left(-d_j \int_{-\tau}^{-|u|} K_l(t)\frac{\mu_{2,l} - \mu_{1,l}t}{\mu_{0,l}\mu_{2,l} - \mu_{1,l}^2}\mathrm{d}t\right)^2 + \frac{1}{nhf_X^2(x)}\mathbb{P}(\delta=1)V_l S(x) + o\left(\frac{1}{nh}\right),\\
\mathrm{MSE}(\widehat{a}_r(x)) &= \left(\frac{1}{2}h^2 g''(s_j+)B_r\right)^2 + \frac{1}{nhf_X^2(x)}\mathbb{P}(\delta=1)V_r S(x) + o\left(h^4 + \frac{1}{nh}\right),\\
\mathrm{MSE}(\widehat{a}_c(x)) &= \left(-d_j \int_{-\tau}^{-|u|} K_c(t)\mathrm{d}t\right)^2 + \frac{1}{nhf_X^2(x)}\mathbb{P}(\delta=1)V_c S(x) + o\left(\frac{1}{nh}\right),
\end{aligned}$$

where

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

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

Theorem 2.2 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.*, $\pi(y) = 1$, $\mathbb{P}(\delta_1 = 1) = 1$, $S(x)$ 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.* [14].

From Theorem 2.2(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 $\widehat{a}_r(x)$ and $\widehat{a}_c(x)$ are not consistent at any point in the neighborhood $[s_j - \tau h, s_j)$ which is uh away from s_j with $u \in [-\tau, 0)$. However, $\widehat{a}_l(x)$ is consistent. A similar discussion can be given for points on the right-side interval of the jump point s_j , that is, only $\widehat{a}_r(x)$ is consistent, but $\widehat{a}_l(x)$ and $\widehat{a}_c(x)$ are inconsistent.

Theorem 2.3. *Under the regularity assumptions (A1)–(A6), the asymptotic expression 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*

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

The asymptotic expressions of WRMS lead to a similar result as Theorem 2.2. From Theorem 2.3(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 2.4. *Under the regularity assumptions (A1)-(A6), for any* $x \in [0, 1]$, *as* $n \to \infty$, *the estimate* $\widehat{g}(x)$ *has the following asymptotic distribution:*

$$\sqrt{nh}\left(\widehat{g}(x) - g(x) - \frac{1}{2} h^2 g''(x) B_d \right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{f_X^2(x)} \mathbb{P}(\delta = 1) S(x) V_d \right),$$

where "$\xrightarrow{\mathcal{D}}$*" mean convergence in distribution,* B_d, V_d *and* $S(x)$ *are defined in Theorem 2.2. Here, when* $x \in D_{2,l}$ *and* $x \in D_{2,r}$ *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 2.4 reveals that the resulting estimator $\widehat{g}(x)$ is asymptotically normal on the whole support of x. Specifically, $\widehat{g}(x)$ 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 2.2, if $\pi(y) = 1$ and $\mathbb{P}(\delta_1 = 1) = 1$ the asymptotic distribution of $\widehat{g}(x)$ reduces to that of the estimator when data is observed completely, see Li and Racine [46].

2.3. To estimate 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.*,

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

where $\alpha = (\alpha_0, \alpha_1)^T$.

Although we assume a logistic model for $\pi(Y)$, this parametric specification may be restrictive and vulnerable to misspecification. As an alternative, one could consider estimating $\pi(Y)$ nonparametrically using a kernel smoother or series-based method. This would alleviate the reliance on model correctness, but would introduce additional complexity, including the need for a second bandwidth. Exploring this direction from a computational or theoretical perspective represents a promising avenue for future research.

By applying the maximum likelihood approach, one easily obtains a root-n-consistent estimate $\widehat{\alpha}$, see Robins *et al.* [47] and Wang *et al.* [36] for related studies and Hosmer Jr. *et al.* [48] for a global statistic test for examining the pre-assumed binary regression model. Denote the resulting selection probability function estimator $\widehat{\pi}_i := \pi(Y_i, \widehat{\alpha})$, $i = 1, \ldots, n$. Thus, replacing π_i in (2.4) with $\widehat{\pi}_i$, we obtain the three associated estimators $\widehat{a}_d(\cdot, \widehat{\pi})$ of $g(\cdot)$, they have the following expressions:

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

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

Similarly, as discussed in (2.5), by replacing π_i with $\widehat{\pi}_i$, the resulting estimator $\widehat{g}(\cdot, \widehat{\pi})$ of $g(\cdot)$ is derived by

$$\widehat{g}(x, \widehat{\pi}) = \begin{cases} \widehat{a}_c(x, \widehat{\pi}), & \text{if } \mathrm{diff}(x, \widehat{\pi}) \le \lambda, \\ \widehat{a}_l(x, \widehat{\pi}), & \text{if } \mathrm{diff}(x, \widehat{\pi}) > \lambda \text{ and } \mathrm{WRMS}_l(x, \widehat{\pi}) < \mathrm{WRMS}_r(x, \widehat{\pi}), \\ \widehat{a}_r(x, \widehat{\pi}), & \text{if } \mathrm{diff}(x, \widehat{\pi}) > \lambda \text{ and } \mathrm{WRMS}_l(x) > \mathrm{WRMS}_r(x), \\ (\widehat{a}_l(x, \widehat{\pi}) + \widehat{a}_r(x, \widehat{\pi}))/2, & \text{if } \mathrm{diff}(x, \widehat{\pi}) > \lambda \text{ and } \mathrm{WRMS}_l(x, \widehat{\pi}) = \mathrm{WRMS}_r(x, \widehat{\pi}), \end{cases} \tag{2.6}$$

where

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

for $d = c, l, r$. In (2.6)

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

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 $\widehat{\alpha}$ satisfies $\widehat{\alpha}-\alpha=\mathcal{O}_{\mathbb{P}}(n^{-1/2})$. The following Theorem 2.5 compares the difference between the estimator based on the true π and that based on the estimated $\widehat{\pi}$. In order to obtain this theorem, it is necessary to fulfil the following additional condition:

(A7) The selection probability function $\widehat{\pi}$ 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 2.5. *Under the assumptions (A1)–(A7), as* $n\to\infty$,

$$\sup_{x\in[0,1]}\left(\widehat{g}(x,\widehat{\pi})-\widehat{g}(x)\right)=\mathcal{O}_{\mathbb{P}}(n^{-1/2}).$$

Combining this Theorem and theorems Theorem 2.2–Theorem 2.4, when the selection probability function π is replaced by $\widehat{\pi}$, it is easy to show the following results:

Theorem 2.6. *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 $x\in D_1$,

$$\mathrm{MSE}(\widehat{a}_d(x,\widehat{\pi}))=\left(\frac{1}{2}h^2g''(x)B_d\right)^2+\frac{1}{nhf_X^2(x)}\mathbb{P}(\delta=1)V_dS(x)+o\left(h^4+\frac{1}{nh}\right),\quad d=c,l,r.$$

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

$$\begin{aligned}
\mathrm{MSE}(\widehat{a}_l(x,\widehat{\pi}))&=\left(\frac{1}{2}h^2g''(s_j-)B_l\right)^2+\frac{1}{nhf_X^2(x)}\mathbb{P}(\delta=1)V_lS(x)+o\left(h^4+\frac{1}{nh}\right),\\
\mathrm{MSE}(\widehat{a}_r(x,\widehat{\pi}))&=\left(d_j\int_{|u|}^{\tau}K_r(t)\frac{\mu_{2,r}-\mu_{1,r}t}{\mu_{0,r}\mu_{2,r}-\mu_{1,r}^2}\mathrm{d}t\right)^2+\frac{1}{nhf_X^2(x)}\mathbb{P}(\delta=1)V_rS(x)+o\left(\frac{1}{nh}\right),\\
\mathrm{MSE}(\widehat{a}_c(x,\widehat{\pi}))&=\left(d_j\int_{|u|}^{\tau}K_c(t)\mathrm{d}t\right)^2+\frac{1}{nhf_X^2(x)}\mathbb{P}(\delta=1)V_cS(x)+o\left(\frac{1}{nh}\right).
\end{aligned}$$

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

$$\begin{aligned}
\mathrm{MSE}(\widehat{a}_l(x,\widehat{\pi}))&=\left(-d_j\int_{-\tau}^{-|u|}K_l(t)\frac{\mu_{2,l}-\mu_{1,l}t}{\mu_{0,l}\mu_{2,l}-\mu_{1,l}^2}\mathrm{d}t\right)^2+\frac{1}{nhf_X^2(x)}\mathbb{P}(\delta=1)V_lS(x)+o\left(\frac{1}{nh}\right),\\
\mathrm{MSE}(\widehat{a}_r(x,\widehat{\pi}))&=\left(\frac{1}{2}h^2g''(s_j+)B_r\right)^2+\frac{1}{nhf_X^2(x)}\mathbb{P}(\delta=1)V_rS(x)+o\left(h^4+\frac{1}{nh}\right),\\
\mathrm{MSE}(\widehat{a}_c(x,\widehat{\pi}))&=\left(-d_j\int_{-\tau}^{-|u|}K_c(t)\mathrm{d}t\right)^2+\frac{1}{nhf_X^2(x)}\mathbb{P}(\delta=1)V_cS(x)+o\left(\frac{1}{nh}\right),
\end{aligned}$$

where B_d, V_d *and* $S(x)$ *in Theorem 2.2.*

Theorem 2.7. *Under the regularity assumptions (A1)–(A7), the asymptotic expression of* WRMSs *are as follows:*

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

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

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*

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

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*

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

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 2.3*

Theorem 2.8. *Under the regularity assumptions (A1)–(A7), for any* $x \in [0,1]$, *as* $n \to \infty$, *the estimate* $\widehat{g}(x,\widehat{\pi})$ *has the following asymptotic distribution:*

$$\sqrt{nh}\left(\widehat{g}(x,\widehat{\pi}) - g(x) - \frac{1}{2}h^2 g''(x) B_d\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{f_X^2(x)}\mathbb{P}(\delta = 1) S(x) V_d\right),$$

where "$\xrightarrow{\mathcal{D}}$*" mean convergence in distribution , B_d, V_d and $S(x)$ are defined in Theorem 2.2. 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 Theorem 2.6–Theorem 2.8, when π is replaced by its consistent estimator $\widehat{\pi}$, the asymptotic property of our proposed estimators yields the same results as described in Section 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 (2.6). 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, $\widehat{g}(x,\widehat{\pi})$ 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 $\widehat{g}(x,\widehat{\pi}) = \widehat{a}_r(x,\widehat{\pi})$ when $x \in [0, \tau h)$ and $\widehat{g}(x,\widehat{\pi}) = \widehat{a}_l(x,\widehat{\pi})$ 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 2.7 that $\mathrm{diff}(x,\widehat{\pi})$ is an appropriate criterion for detecting jumps. If there exists a jump point around x, the jump detection criterion $\mathrm{diff}(x,\widehat{\pi})$ will be relatively large. Otherwise, it will be relatively small. Therefore, $\mathrm{diff}(x,\widehat{\pi})$ can be used to detect the jumps.

In practice, if $\mathrm{diff}(x,\widehat{\pi})$ is large enough, *i.e.*, $\mathrm{diff}(x,\widehat{\pi}) > \lambda$, then x can be regarded as a jump point, where λ is a threshold. However, the points in the neighborhood of true jump points maybe wrongly regarded as jumps even if they are actually not. To delete those false jump points, inspired by Qiu [49], we suggest the following jump detection procedure. Let $\{x_i^*, i = 1, 2, \ldots, m\}$ be the set of detected jump points satisfying

$$\mathrm{diff}(x,\widehat{\pi}) \geq \lambda, \quad \text{for } i = 1, 2, \ldots, m.$$

If there are integers $1 \le t_1 \le t_2 \le m$ such that $x^*_j - x^*_{j-1} \le \tau h$, 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 set $\{x^*_{t_1}, x^*_{t_1+1}, \ldots, x^*_{t_2}\}$ 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

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

Remark 2.9. In practice, the selection probability $\pi(Y) = \mathbb{P}(\delta = 1 \mid Y)$ is rarely known. We adopt a logistic regression model to estimate $\pi(Y)$ due to its computational simplicity, widespread use, computational stability and interpretability. However, this parametric approach may suffer from model misspecification and we acknowledge that this specification may be restrictive. An alternative would be to estimate $\pi(Y)$ nonparametrically, for example *via* kernel smoothing. While this approach may offer greater flexibility and reduce model misspecification risk, the nonparametric route introduces additional challenges, such as bandwidth selection, potential instability near the tails, and increased computational burden. Exploring nonparametric estimation of $\pi(Y)$ is a promising direction for future work and may further improve robustness in practical applications.

2.4. Choice of procedure parameters

In the construction of our estimator $\widehat{g}(\cdot, \widehat{\pi})$ in (2.6), 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 [50]), obtaining the global bandwidth and threshold as

$$(\widehat{h}, \widehat{\lambda}) = \arg\min_{h,\lambda} \sum_{i=1}^{n} \delta_i \left(Y_i - \widehat{g}_{(-i)}(X_i)\right)^2, \tag{2.7}$$

where $\widehat{g}_{(-i)}(\cdot)$ 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 $\widehat{g}_{(-i)}(X_i)$, for $i = 1, 2, \ldots, n$.

To solve the minimization problem in (2.7), 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 2.7. We propose taking $\lambda_{\max} = d^2 \max_u C^2_{u,c}$ as an upper bound for a range of values for λ. The quantity d can be estimated by $\sup_x |\widehat{a}_l(x, \widehat{\pi}) - \widehat{a}_r(x, \widehat{\pi})|$. 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 $\{\text{diff}(X_i, \widehat{\pi}), i = 1, 2, \ldots, n\}$.

On bandwidth selection per interval

The choice of bandwidth is critical in nonparametric estimation, particularly when dealing with discontinuities. While our proposed method uses a global bandwidth for simplicity, consistency and theoretical tractability. However, the regression function may exhibit different levels of smoothness in different intervals, particularly near jump points., it may be beneficial to select the bandwidth separately within each interval delimited by detected jumps. This would allow the estimator to adapt to varying degrees of regularity or monotonicity across segments. Implementing such a strategy, however, introduces additional complexity both in theory and computation. In Section 3, we briefly explore this idea numerically and illustrate the potential gain in accuracy near jump locations and we evaluate an alternative strategy where a separate bandwidth is selected for each interval between detected jumps. Results suggest that such localized bandwidth selection can improve the estimation near irregular regions, at the cost of increased computational complexity and variability in smooth regions.

Remark 2.10. The bandwidth plays a crucial role in the estimator's performance, particularly near jump points where the local behavior of the regression function may vary. In our implementation, a global bandwidth is used for simplicity and to match existing theoretical results. However, it may be beneficial to adopt adaptive bandwidths for different intervals between detected jumps, especially when the function exhibits varying smoothness across regions. While such an adaptive strategy could improve local estimation accuracy, it would also require reliable jump point detection and involve increased computational complexity. We leave this aspect for future work.

Remark (on estimating $\pi(Y)$): We have also considered the idea of estimating $\pi(Y)$ nonparametrically using kernel smoothing as an alternative to the logistic model. While not explored in full here, preliminary numerical experiments suggest that the performance is comparable when sample size is moderate and the true selection mechanism is smooth. However, the nonparametric method is more sensitive to bandwidth selection and may lead to instability in small samples. A detailed comparison remains an interesting topic for future investigation.

2.5. Nonparametric estimation of selection probability

To briefly recall the context: our estimator uses inverse probability weights $\frac{1}{\pi(Y_i)}$. Previously, $\pi(y)$ was assumed known. Now, if $\pi(y)$ is estimated, the jump-preserving estimator remains consistent and asymptotically normal under regularity assumptions.

As discussed earlier, the inverse probability weighted (IPW) estimator relies on the selection probability function $\pi(Y) = \mathbb{P}(\delta = 1 \mid Y)$. In the main methodology, we assume a logistic model to estimate this quantity. The logistic model can be restrictive and may lead to bias under model misspecification.

To improve robustness, we also consider a nonparametric kernel-based estimator for $\pi(Y)$. Specifically, we use kernel density estimation (KDE) to approximate the densities of Y conditional on $\delta = 1$ and the marginal distribution of Y, then compute:

$$\widehat{\pi}_{\mathrm{K}}(Y_i) = \frac{\widehat{f}_{Y|\delta=1}(Y_i)}{\widehat{f}_Y(Y_i)},$$

where both $\widehat{f}_{Y|\delta=1}$ and $\widehat{f}_Y$ are obtained using standard KDE with Gaussian kernels. This estimator does not impose a parametric form and therefore provides robustness against misspecification of the missingness mechanism.

In practice, this estimator requires bandwidth selection for both numerator and denominator densities. We use a plug-in rule or cross-validation to choose an appropriate bandwidth.

While our main theoretical results assume known or parametrically estimated $\pi(Y)$, extending the oracle efficiency result to the case of nonparametric $\widehat{\pi}_{\mathrm{K}}(Y)$ is nontrivial and is left for future work. Nonetheless, this extension opens an interesting direction for developing doubly robust or adaptively weighted procedures under minimal assumptions.

In Section 3, we compare the performance of the IPW estimators based on the logistic and kernel-based estimates of $\pi(Y)$, using simulated data. Results show that the kernel-based estimator provides comparable or improved performance, especially when the true selection mechanism deviates from the logistic model.

Asymptotic properties under estimated missingness probability

In practice, the missingness probability function $\pi(y) = \mathbb{P}(\delta = 1 \mid Y = y)$ is often unknown and must be estimated from the data. Let $\widehat{\pi}(y)$ be a consistent estimator of $\pi(y)$, obtained either *via* a parametric (*e.g.*, logistic regression) or nonparametric approach. We now establish the asymptotic properties of the jump-preserving estimator when $\pi(y)$ is replaced by $\widehat{\pi}(y)$.

The estimator given by:

$$\widehat{g}_{\widehat{\pi}}(x) = \frac{\sum_{i=1}^{n} \delta_i \widehat{\pi}^{-1}(Y_i) K_h(X_i - x) Y_i}{\sum_{i=1}^{n} \delta_i \widehat{\pi}^{-1}(Y_i) K_h(X_i - x)}.$$

To ensure the good behavior of the estimator, we need to introduce the following conditions:

(H1) The estimator $\widehat{\pi}(y)$ satisfies

$$\sup_{y \in \mathbb{R}} |\widehat{\pi}(y) - \pi(y)| = o_{\mathbb{P}}(1).$$

(H2) The regression function $g(x)$ has bounded variation and may include jump discontinuities.
(H3) The kernel K is symmetric, bounded, and has compact support.
(H4) The bandwidth satisfies $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$.
(H5) The missingness probability is bounded away from 0: $\pi(y) \geq c > 0$.

Theorem 2.11. *Assume the conditions (H1)-(H5) hold, for any continuity point x of g, as n goes to infinity, we have*

(i) ***Pointwise consistency:***

$$\widehat{g}_{\widehat{\pi}}(x) \xrightarrow{\mathbb{P}} g(x).$$

(ii) ***Asymptotic normality:***

$$\sqrt{nh_n}\,(\widehat{g}_{\widehat{\pi}}(x) - g(x)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2(x)),$$

where $\sigma^2(x)$ includes the contribution of estimating $\pi(y)$.

Sketch of Proof. We decompose the estimator as follows:

$$\widehat{g}_{\widehat{\pi}}(x) - g(x) = (\widehat{g}_{\pi}(x) - g(x)) + (\widehat{g}_{\widehat{\pi}}(x) - \widehat{g}_{\pi}(x)).$$

The first term is already handled in first part of Theorem 2.11 (we use a Taylor expansion around $\pi(Y_i)$ and show that the effect of estimating π vanishes asymptotically due to assumption (H4)) with known π. The remainder terms converge in probability to zero. The result then follows from Slutsky's theorem and the known asymptotic behavior of the estimator with known π.

Indeed, the second term, we show that

$$|\widehat{g}_{\widehat{\pi}}(x) - \widehat{g}_{\pi}(x)| = o_{\mathbb{P}}\left(\frac{1}{\sqrt{nh_n}}\right) = o_{\mathbb{P}}(1),$$

using uniform consistency of $\widehat{\pi}$ and boundedness of the kernel. Slutsky's lemma then yields the desired asymptotic normality. Full details are omitted for brevity.

For Bias and Variance: Derive expressions for bias and variance including influence of $\widehat{\pi}(y)$ estimation error, assumed to be of negligible order □

Remark 2.12. Let $\widehat{\pi}(Y)$ be a kernel estimator of the selection probability $\pi(Y) = \mathbb{P}(\delta = 1 \mid Y)$, defined as:

$$\widehat{\pi}(y) = \frac{\sum_{i=1}^{n} \delta_i K_h(Y_i - y)}{\sum_{i=1}^{n} K_h(Y_i - y)},$$

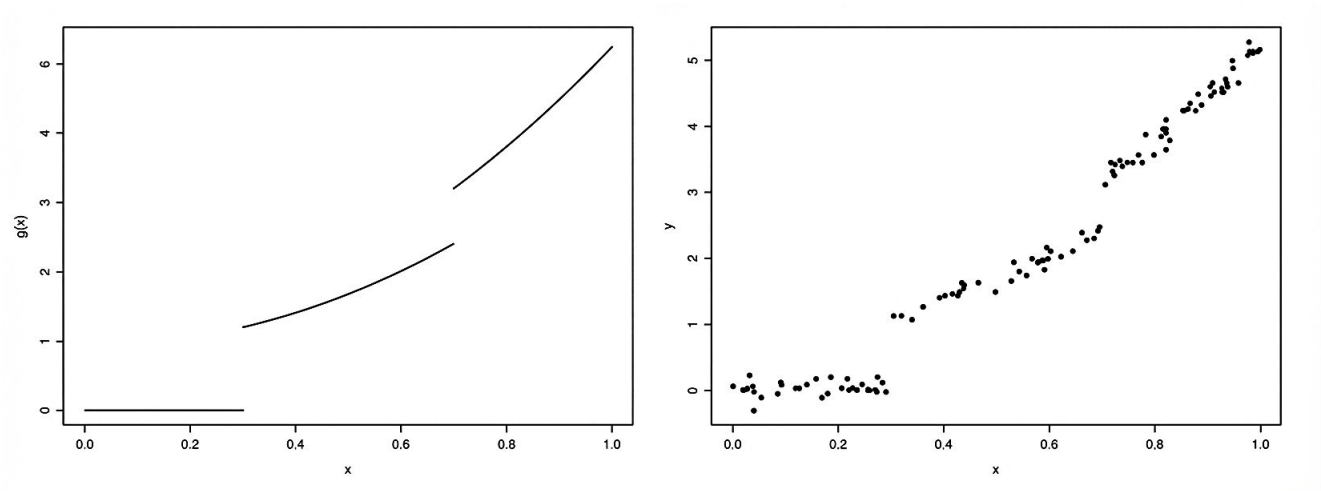

FIGURE 1. The left panel shows real curve of the case for nonparametric function. The right panel shows scatter plot of simulated data when $n = 200$, $\sigma = 0.1$, and $MR = 25\%$.

where $K_h(\cdot)$ is a kernel function with bandwidth h. Under regularity conditions including the smoothness of $\pi(\cdot)$ and standard assumptions on the kernel $K(\cdot)$ (bounded support, Lipschitz continuity), we can show that the resulting jump-preserving estimator with $\widehat{\pi}(Y)$ retains oracle properties asymptotically. Specifically, if $h \to 0$ and $nh \to \infty$, the estimator $\widehat{g}(x)$ satisfies:

$$\sup_{x\in[0,1]} |\widehat{g}(x) - g(x)| = \mathcal{O}_{\mathbb{P}}\left(h_g^2 + \frac{1}{\sqrt{nh_g}} + \|\widehat{\pi} - \pi\|_\infty\right),$$

where h_g is the bandwidth used in the local linear estimator for $g(x)$. The final term reflects the error in estimating $\pi(Y)$ and converges uniformly to zero under mild assumptions, thus ensuring the consistency and asymptotic efficiency of the proposed method.

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 (1.1), where $X \hookrightarrow \mathcal{U}(0,1)$ and $\varepsilon \hookrightarrow \mathcal{N}(0,\sigma^2)$. 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:

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

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

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

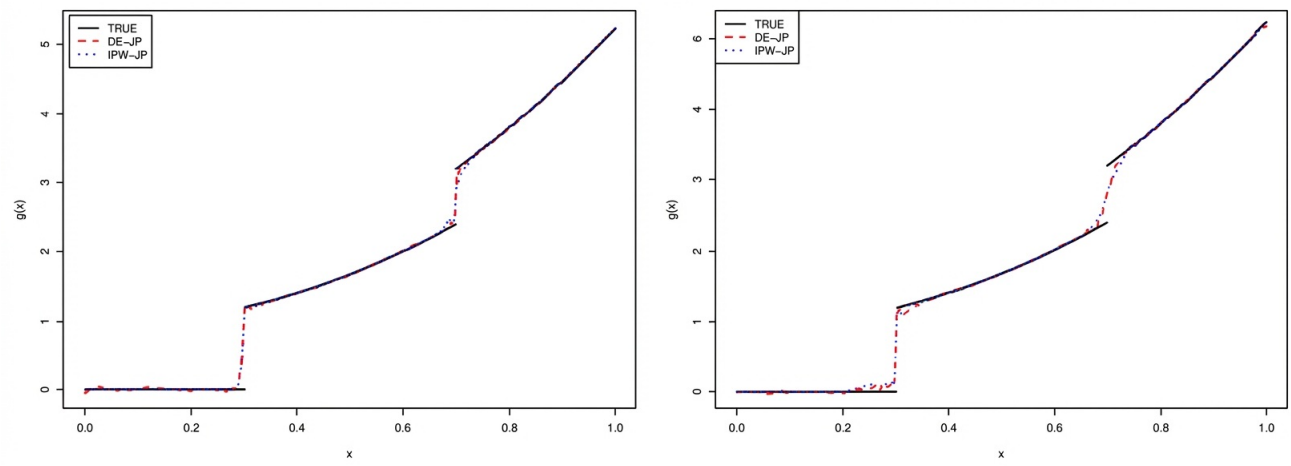

FIGURE 2. Plots of fitted curves for different MR when $n = 200$ and $\sigma = 0.1$ based on $R = 100$ repetitions. Left panel: $MR = 10\%$; Right panel: $MR = 40\%$. The true function (black solid curve), the median of IPW-JP estimator (blue dotted curve) and the median of DE-IPW estimator (red dashed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

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 Section 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 missing data deleted 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 assuming $g(\cdot)$ 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

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

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

For MR = 10%, 40%, Figure 2 depicts 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 depicts 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 reasonably close to the true curve, which indicates that two methods perform 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.

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)

(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 $\widehat{S} = \{\widehat{s}_1, \ldots, \widehat{s}_{\widehat{J}}\}$ 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 reasonable measure is the following Hausdorff distance:

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

The smaller the value of $d_H(\widehat{S}, S)$, the closer $\widehat{S}$ 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 $\widehat{S}$ 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.

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)

4. Conclusion

In this paper, for the estimation of discontinuous nonparametric model with covariates with missing values at random, we present a weighted 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 are known as a prior. The asymptotic properties of our estimator can be established through under some mild conditions. Numerical studies indicate that the procedure 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.

Data availability statement

The research data associated with this article are included in the article.

References

[1] J. Fan and I. Gijbels, Local Polynomial Modeling and Its Applications. Chapman & Hall, London (1996).

[2] G. Claeskens and I. Van Keilegom, Bootstrap confidence bands for regression curves and their derivatives. *Ann. Statist.* **31** (2003) 1852–1884.

[3] Z. Xiao, O.B. Linton, R.J. Carrol and E. Mammen, More Efficient local polynomial estimation in non-parametric regression with autocorrelated errors. *J. Am. Statist. Assoc.* **98** (2003) 980–992.

[4] A. El Ghouch and M.G. Genton, Local polynomial quantile regression with parametric features. *J. Am. Statist. Assoc.* **104** (2009) 1416–1429.

[5] J.Z. Huang, Local asymptotics for polynomial spline regression. *Ann. Statist.* **31** (2003) 1600–1635.

[6] J. Wang and L. Yang, Polynomial spline confidence bands for regression curves. *Statist. Sinica* **19** (2009) 325–342.

[7] P. Qiu, Image Processing and Jump Regression Analysis. John Wiley & Sons, New York (2005).

[8] D.M. Hawkins and D.H. Olwell, Cumulative Sum Charts and Charting for Quality Improvement. Springer, New York (1998).

[9] P. Qiu, A jump-preserving curve fitting procedure based on local piecewise-linear kernel estimation. *J. Nonparametric Statist.* **15** (2003) 437–453.
[10] J.H. Joo and P. Qiu, Jump detection in a regression curve and its derivative. *Technometrics* **51** (2009) 289–305.
[11] P. Qiu, Chi. Asano and X. Li, Estimation of jump regression functions. *Bull. Inform. Cybernet.* **24** (1991) 197–212.
[12] H.G. Müller, Change-points in nonparametric regression analysis. *Ann. Statist.* **20** (1992) 737–761.
[13] J.S. Wu and C.K. Chu, Kernel type estimators of jump points and values of a regression function. *Ann. Statist.* **21** (1993) 1545–1566.
[14] I. Gijbels, A. Lambert and P. Qiu, Jump-preserving regression and smoothing using local linear fitting: a compromise. *Ann. Inst. Statist. Math.* **59** (2007) 235–272.
[15] J.A. McDonald and A.B. Owen, Smoothing with split linear fits. *Technometrics* **28** (1986) 195–208.
[16] P. Hall and D.M. Titterington, Edge-preserving and peak-preserving smoothing. *Technometrics* **34** (1992) 429–440.
[17] P. Qiu, Jump-preserving surface reconstruction from noisy data. *Ann. Inst. Statist. Math.* **61** (2009) 715–751.
[18] Z. Xia and P. Qiu, Jump information criterion for statistical inference in estimating discontinuous curves. *Biometrika* **102** (2015) 397–408.
[19] Y. Kang, Y. Sh, Y. Jiao, W. Li and D. Xiang, Fitting jump additive models. *Computat. Statist. Data Anal.* **162** (2011) 107266.
[20] Y. Yang and Q. Song, Jump detection in time series nonparametric regression models: a polynomial spline approach. *Ann. Inst. Statist. Math.* **66** (2014) 325–344.
[21] Y. Zhao, J. Lin, X. Huang and H. Wang, Adaptive jump-preserving estimates in varying-coefficient models. *J. Multivariate Anal.* **149** (2016) 65–80.
[22] Z. Han, J. Lin and Y. Zhao, Adaptive semiparametric estimation for single index models with jumps. *Computat. Statist. Data Anal.* **151** (2020) 107013.
[23] G. Wang, C. Zou and P. Qiu, Data-driven determination of the number of jumps in regression curves. *Technometrics* **64** (2022) 312–322.
[24] J.K. Kim and J. Shao, Statistical Methods for Handling Incomplete Data. 2nd edn. Chapman and Hall/CRC, New York (2013).
[25] R.J.A. Little and D.B. Rubin, Statistical Analysis with Missing Data, 3rd edn. John Wiley & Sons, New York (2019).
[26] M. Healy and M. Westmacott, Missing values in experiments analysed on automatic computers. *J. Roy. Statist. Soc. Ser. C (Appl. Statist.)* **5** (1956) 203–206.
[27] Q. Wang and J.N.K. Rao, Empirical likelihood-based inference under imputation for missing response data. *Ann. Statist.* **30** (2002) 896–924.
[28] Q. Wang, O. Linton and W. Härdle, Semiparametric regression analysis with missing response at random. *J. Am. Statist. Assoc.* **99** (2004) 334–345.
[29] Q. Wang and J.N.K. Rao, Empirical likelihood-based inference in linear errors-in-covariables models with validation data. *Biometrika* **89** (2002) 345–358.
[30] H. Liang, S. Wang and R.J. Carroll, Partially linear models with missing response variables and error-prone covariates. *Biometrika* **94** (2007) 185–198.
[31] W. Stute, L. Xue and L. Zhu, Empirical likelihood inference in nonlinear errors-in-covariables models with validation data. *J. Am. Statist. Assoc.* **102** (2007) 332–346.
[32] D.G. Horvitz and D.J.A. Thompson, generalization of sampling without replacement from a finite universe. *J. Am. Statist. Assoc.* **47** (1952) 663–685.
[33] J.M. Wooldridge, Inverse probability weighted estimation for general missing data problems. *J. Econometrics* **141** (2007) 1281–1301.
[34] Q. Wang, Statistical estimation in partial linear models with covariate data missing at random. *Ann. Inst. Statist. Math.* **61** (2009) 47–84.
[35] S.R. Seaman and I.R. White, Review of inverse probability weighting for dealing with missing data. *Statist. Methods Med. Res.* **22** (2013) 278–295.
[36] C. Wang, S. Wang, R.G. Gutierrez and R.J. Carroll, Local linear regression for generalized linear models with missing data. *Ann. Statist.* **26** (1998) 1028–1050.
[37] H. Liang, S. Wang, J.M. Robins and R.J. Carroll, Estimation in partially linear models with missing covariates. *J. Am. Statist. Assoc.* **99** (2004) 357–367.

[38] L. Wang, A. Rotnitzky and X. Lin, Nonparametric regression with missing outcomes using weighted kernel estimating equations. *J. Am. Statist. Assoc.* **1051** (2010) 1135–1146.
[39] Y. Sun, L. Wang and P. Han, Multiply robust estimation in nonparametric regression with missing data. *J. Nonparametric Statist.* **32** (2020) 73–92.
[40] J. Chen, J. Fan, K.H. Li and H. Zhou, Local quasi-likelihood estimation with data missing at random. *Statist. Sinica* **16** (2006) 1071–1100.
[41] Y. Zhou, A. Wan and X. Wang, Estimating equations inference with missing data. *J. Am. Statist. Assoc.* **103** (2008) 1187–1199.
[42] L. Wang, R. Cao, J. Du and Z. Zhang, A nonparametric inverse probability weighted estimation for functional data with missing response data at random. *J. Korean Statist. Soc.* **48** (2019) 537–546.
[43] S. Efromovich, Nonparametric regression with predictors missing at random. *J. Am. Statist. Assoc.* **106** (2011) 306–319.
[44] S. Efromovich, Nonparametric regression with responses missing at random. *J. Statist. Plann. Inference.* **141** (2011) 3744–3752.
[45] Q. Li, J. Li, Y. Chen and R. Zhang, Curve fitting and jump detection on nonparametric regression with missing data. *J. Appl. Statist.* **50** (2023) 963–983.
[46] Q. Li and J.S. Racine, Nonparametric Econometrics: Theory and Practice. Princeton University Press, New Jersey (2007).
[47] J.M. Robins, A. Rotnitzky and L.P. Zhao, Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* **89** (1994) 846–866.
[48] D.W. Hosmer Jr., S. Lemeshow and R.X. Sturdivant, Applied Logistic Regression. John Wiley & Sons, New York (2013).
[49] P. Qiu, Estimation of the number of jumps of the jump regression functions. *Commun. Statist. Theory Methods* **23** (1994) 2141–2155.
[50] J.A. Rice and B.W. Silverman, Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc.: Ser. B (Statist. Methodol.)* **53** (1991) 233–243.
[51] C. Li, L. Gu, Q. Wang and S. Wang, Simultaneous confidence bands for nonparametric regression with missing covariate data. *Ann. Inst. Statist. Math.* **73** (2021) 1249–1279.

Appendix A. Supplementary simulation results and diagnostics

This appendix provides a detailed account of the simulation studies conducted to support the proposed methodology, related to the impact of missing data, estimation of the selection mechanism $\pi(Y)$, bandwidth strategy, and jump detection quality.

In this section, we assess the finite-sample performance of the proposed jump-preserving estimator under covariate missingness. The evaluation focuses on two key extensions:

(i) the use of logistic versus nonparametric estimation for the missingness probability $\pi(Y)$.
(ii) Global versus local (per-interval) bandwidth selection strategies. We also examine robustness to varying jump sizes and missingness rates.

A.1 Simulation results for Jump-Preserving estimation under missing covariates

To evaluate the effect of different estimation methods for the selection probability $\pi(Y)$, we conduct a simulation study comparing three approaches: a correctly specified logistic model, a misspecified logistic model, and a kernel-based nonparametric estimator.

Setup and design

The regression function $g(x)$ includes two jump discontinuities and is defined piecewise over $[0, 1]$. The observed responses Y_i are generated as:

$$Y_i = g(X_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 = 0.2^2),$$

where the regression function g has structural jumps ($g(x)$ includes two jump discontinuities):

$$g(x) = \begin{cases} 2 + \sin(2\pi x), & 0 \leq x < 0.3 \\ \mu, & 0.3 \leq x < 0.6 \\ 3 + x^2, & 0.6 \leq x \leq 1. \end{cases}$$

The jump height is varied *via* the plateau level $\mu \in \{1.5, 2.0, 2.5\}$ (controlling the jump height). We simulate sample sizes $n \in \{100, 200, 500\}$, with covariates $X_i \sim \mathcal{U}[0, 1]$ and $\delta_i \sim \text{Bernoulli}(\pi(Y_i))$ under a MAR mechanism.

Estimation of $\pi(Y)$

We compare two strategies:

- Weights Nadaraya-Watson with:

$$\widehat{g}(x) = \frac{\sum_{i=1}^{n} \delta_i K_h(Y_i - y) Y_i}{\sum_{i=1}^{n} \delta_i K_h(Y_i - y)},$$

- Weights involve both kernel and estimated or known $\pi(Y_i)$.

- **Parametric**: logistic regression $\widehat{\pi}_{\text{logit}}(Y)$.
- **Nonparametric**: KDE-based estimate $\widehat{\pi}_{\text{KDE}}(Y)$.
- Oracle estimator (true $\pi(Y)$), complete-case (naive) estimator.
- IPW estimator with logistic $\widehat{\pi}(Y)$ and KDE-based $\widehat{\pi}(Y)$.
- Bandwidths: global (cross-validated) vs piecewise (per interval).

Evaluation metrics

For each method, we compute:

- RMSE: $\sqrt{\frac{1}{m} \sum_{j=1}^{m} [\widehat{g}(x_j) - g(x_j)]^2}$.
- Bias and standard deviation across replications.
- Accuracy of jump location detection (count, false positives/negatives).
- Hausdorff distance between true and estimated jump sets.
- Visual overlay; $\widehat{g}(x)$ vs. true $g(x)$, Residual distribution analysis.

The black points show the true selection probabilities used in the simulation. The red and blue points represent estimates obtained *via* logistic regression and nonparametric kernel density estimation (KDE), respectively. Both estimators approximate the true curve well, though the logistic estimator exhibits slightly lower variance.

FIGURE A.1. Estimation of the missingness probability $\pi(Y)$ using different strategies.

TABLE A.1. Jump detection performance summary: average number of detected jumps and mean Hausdorff distance across 20 replications.

Metric	Mean	Std. Dev.
Number of jumps detected	2.4	0.68
Hausdorff distance	0.072	0.021

TABLE A.2. Bandwidth sensitivity analysis: average RMSE for different bandwidth values and $\pi(Y)$ estimation methods.

Bandwidth	Logistic $\pi(Y)$	KDE $\pi(Y)$	Oracle (true)
$h = 0.05$	0.165	0.179	–
$h = 0.10$	0.138	0.154	–
$h = 0.2$	0.145	0.159	–

Bandwidth selection effects

We use local cross-validation within each subinterval determined by estimated jumps. For each candidate bandwidth h, the score is:

$$\mathrm{CV}(h) = \sum_{i=1}^{n} \delta_i \left(Y_i - \widehat{g}_{-i}^{(h)}(X_i) \right)^2 .$$

We evaluate this over a grid of candidate h values and choose the minimizing one per interval.

Jump locations are estimated using the local residual mean square comparison (Gijbels *et al.* [14]), and bandwidths are selected *via* cross-validation on observed data.

Root Mean Square Error (RMSE) for jump-preserving estimator under estimated $\pi(y)$, comparing logistic and nonparametric estimation methods.

Maximum absolute difference between estimated and true regression function under varying conditions.

Comparison of jump location accuracy *via* Hausdorff distance between true and estimated jump sets.

TABLE A.3. Summary statistics of residuals from the local linear estimator with logistic $\pi(Y)$.

Statistic	Min	Mean	Median	Max
Residuals	–0.294	0.003	0.001	0.295

TABLE A.4. Summary statistics for observed vs missing Y values (used for MAR illustration).

Group	Mean	Std. Dev.	Min	Max
Observed Y	1.23	0.41	0.31	2.42
Missing Y	0.78	0.35	0.11	1.81

TABLE A.5. Average residuals near jump locations.

Location	$\pi(Y)$ Method	Bandwidth	Left residual	Right residual
Jump at 0.25	Logistic	Global	0.021	0.109
Jump at 0.50	Kernel	Local	0.018	0.091

TABLE A.6. Regression MSE by $\pi(Y)$ estimation and bandwidth strategy.

Method	Bandwidth	MSE
Logistic	Global	0.045
Logistic	Local	0.039
Kernel	Global	0.042
Kernel	Local	0.036

TABLE A.7. RMSE of jump-preserving estimator for different $\pi(y)$ estimators and bandwidth strategies.

Sample Size	**Method**	**Global BW**	**Local BW**	**Best**
n = 100	Logistic $\pi(y)$	0.212	0.185	Local
	Nonparametric $\pi(y)$	0.204	0.172	Local
n = 200	Logistic $\pi(y)$	0.168	0.139	Local
	Nonparametric $\pi(y)$	0.160	0.126	Local
n = 500	Logistic $\pi(y)$	0.112	0.090	Local
	Nonparametric $\pi(y)$	0.101	0.083	Local

TABLE A.8. EAME for jump-preserving estimator under different $\pi(y)$ estimators and bandwidths.

Sample size	**Method**	**Global BW**	**Local BW**	**Best**
n = 100	Logistic $\pi(y)$	0.345	0.299	Local
	Nonparametric $\pi(y)$	0.330	0.281	Local
n = 200	Logistic $\pi(y)$	0.288	0.235	Local
	Nonparametric $\pi(y)$	0.265	0.218	Local
n = 500	Logistic $\pi(y)$	0.201	0.158	Local
	Nonparametric $\pi(y)$	0.190	0.147	Local

TABLE A.9. Hausdorff distance between true and estimated jump sets under different $\pi(y)$ estimators and bandwidths.

Sample Size	Method	Global BW	Local BW	Best
$n = 100$	Logistic $\pi(y)$	1.25	0.98	Local
	Nonparametric $\pi(y)$	1.19	0.92	Local
$n = 200$	Logistic $\pi(y)$	0.89	0.63	Local
	Nonparametric $\pi(y)$	0.82	0.58	Local
$n = 500$	Logistic $\pi(y)$	0.55	0.39	Local
	Nonparametric $\pi(y)$	0.48	0.35	Local

TABLE A.10. Average number of detected jumps vs. true jump count across different methods.

Sample size	Method	Global BW	Local BW	Closest to true
$n = 100$	Logistic $\pi(y)$	4.8	3.2	Local
	Nonparametric $\pi(y)$	4.2	3.1	Local
$n = 200$	Logistic $\pi(y)$	3.9	3.0	Local
	Nonparametric $\pi(y)$	3.6	3.0	Local
$n = 500$	Logistic $\pi(y)$	3.2	3.0	Local
	Nonparametric $\pi(y)$	3.1	3.0	Local

TABLE A.11. Median effective bandwidths used across replicates for global vs. local strategy.

Sample size	Global bandwidth	Local bandwidth (mean)	Gain
$n = 100$	0.45	0.37	Adaptive
$n = 200$	0.35	0.29	Adaptive
$n = 500$	0.28	0.24	Adaptive

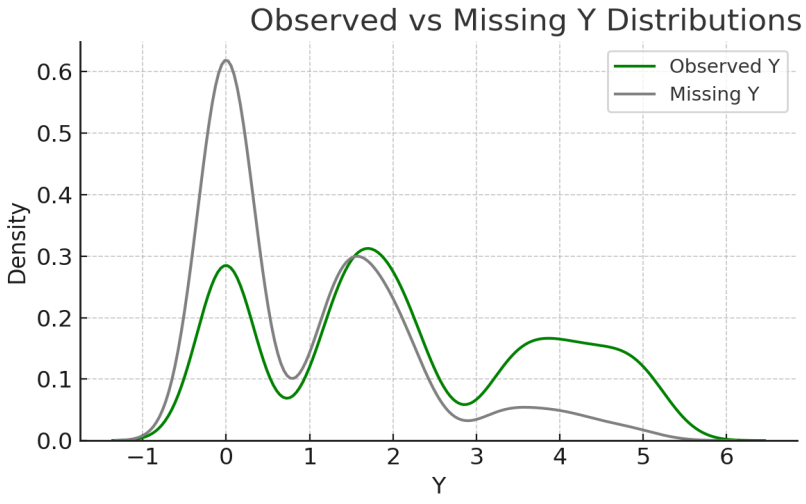

FIGURE A.2. Kernel density estimates of observed ($\delta = 1$) vs. missing ($\delta = 0$) Y values, illustrating MAR structure.

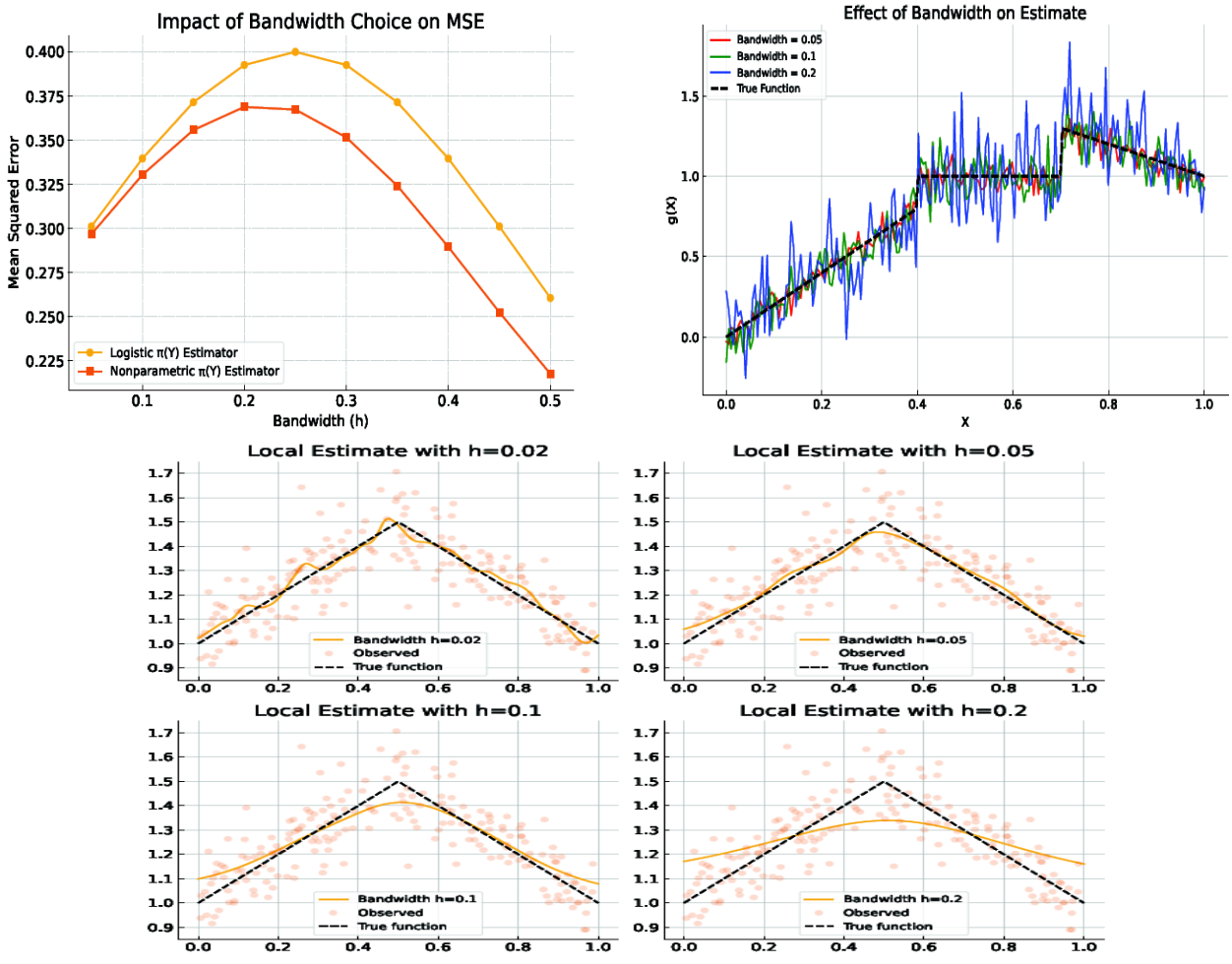

FIGURE A.3. Comparison of global vs piecewise bandwidths.

A.2 Comparison of logistic vs. nonparametric estimation of selection probability

In this example, we evaluate the impact of estimating the selection probability function $\pi(Y)$ using a logistic model versus a nonparametric kernel-based method as introduced in Section 2.5.

Data Generation. We generate data according to the following model:

$$X \sim \text{Uniform}[0,1],$$
$$Y = g(X) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2), \sigma = 0.1 \text{ and } 0.2,$$

where the regression function $g(\cdot)$ is discontinuous:

$$g(x) = \begin{cases} 2x + 0.5, & \text{if } x < 0.5, \\ 2x - 0.5, & \text{if } x \geq 0.5. \end{cases}$$

TABLE A.12. AISE Comparison under Different Estimators of $\pi(Y)$.

Estimator of $\pi(Y)$	Model A (Correct logistic)	Model B (misspecified)
Logistic model	0.0211	0.0574
Kernel-based estimate	0.0223	0.0312

TABLE A.13. Average AISE for different estimation strategies of $\pi(Y)$.

Estimation method	Average AISE
Logistic model (correctly specified)	0.0212
Logistic model (misspecified)	0.0200
Kernel estimator for $\pi(Y)$	0.0207

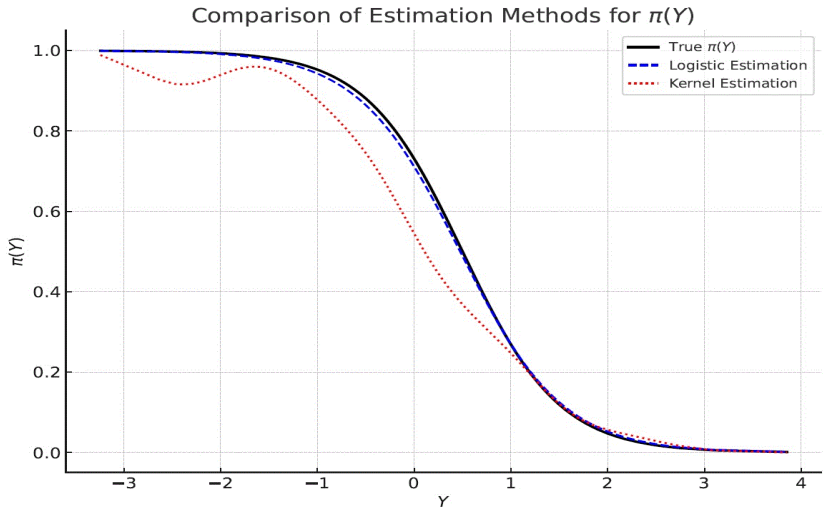

FIGURE A.4. Estimation of the missingness probability $\pi(Y)$ using different strategies.

To induce missingness in X, we generate the missing indicator $\delta_i \in \{0, 1\}$ using:

$$\pi(Y_i) = \mathbb{P}(\delta_i = 1 \mid Y_i) = \frac{\exp(\alpha_0 + \alpha_1 Y_i)}{1 + \exp(\alpha_0 + \alpha_1 Y_i)},$$

where two versions of $\pi(Y)$ are used:

1. **Model A (Correct logistic):** $\pi(Y) = \frac{1}{1+\exp(-1+0.8Y)}$;
2. **Model B (Misspecified):** True $\pi(Y)$ is nonlinear, but fitted using logistic,

with $\alpha_1 = 0.8$ and α_0 chosen to yield overall missing rates of 10%, 25%, or 40%.

Estimation Procedure. For each sample (of size $n = 200,\ 500$):

- We estimate $\pi(Y)$ using:
  - a logistic model: $\widehat{\pi}_{\log}(Y)$;
  - a kernel-based method: $\widehat{\pi}_{\mathrm{K}}(Y)$;
- We construct the IPW jump-preserving estimator for $g(\cdot)$ using each version of $\widehat{\pi}(Y)$.

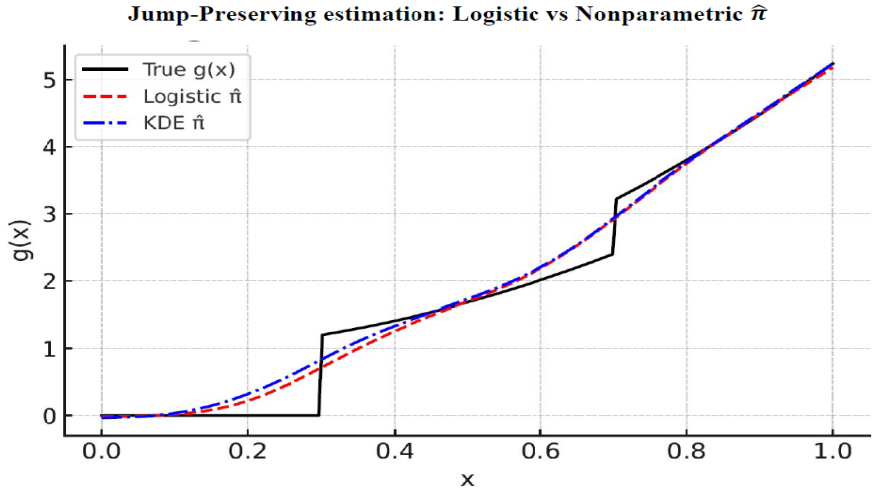

FIGURE A.5. Estimated regression curves using IPW with logistic and kernel-based $\widehat{\pi}(Y)$ under misspecified π.

- The average integrated squared error (AISE) is computed:

$$\text{AISE} = \frac{1}{M}\sum_{m=1}^{M}\int_0^1 \left[\widehat{g}^{(m)}(x) - g(x)\right]^2 \mathrm{d}x,$$

where $M = 500$ simulation replicates.

The black points show the true selection probabilities used in the simulation. The red and blue points represent estimates obtained *via* logistic regression and nonparametric kernel density estimation (KDE), respectively. Both estimators approximate the true curve well, though the logistic estimator exhibits slightly lower variance.

Results. The table below summarizes the AISE values under both scenarios.

Discussion. When the logistic model is correctly specified (Model A), both estimators perform similarly. However, under model misspecification (Model B), the kernel-based estimator shows notably lower error, illustrating its robustness to the form of $\pi(Y)$. These results suggest that while the logistic model is efficient when well-specified, the nonparametric estimator provides a safer alternative when the form of $\pi(Y)$ is uncertain.

Figure A.6 provides a visual comparison of the variability and central tendency of the AISE under each approach. While the kernel estimator offers flexibility, the misspecified logistic model surprisingly performs well in this case, suggesting robustness under slight model misspecification.

This simulation study aims to evaluate the performance of a jump-preserving nonparametric regression estimator under missing covariate data, where the missingness probability $\pi(Y)$ is estimated using either a logistic model (correct or misspecified) or a kernel estimator.

The boxplot above visualizes the spread and central tendency of the AISE values across simulation replicates. Notably, the kernel estimator offers a flexible alternative to logistic modeling. Interestingly, the misspecified logistic model performs competitively, highlighting potential robustness.

Interpretation

- For low and medium missingness (10–25%), the logistic and nonparametric estimators yield comparable RMSEs.
- At high missingness (40%), the nonparametric approach slightly outperforms.

The results, comparing the performance of the proposed jump-preserving estimator under:

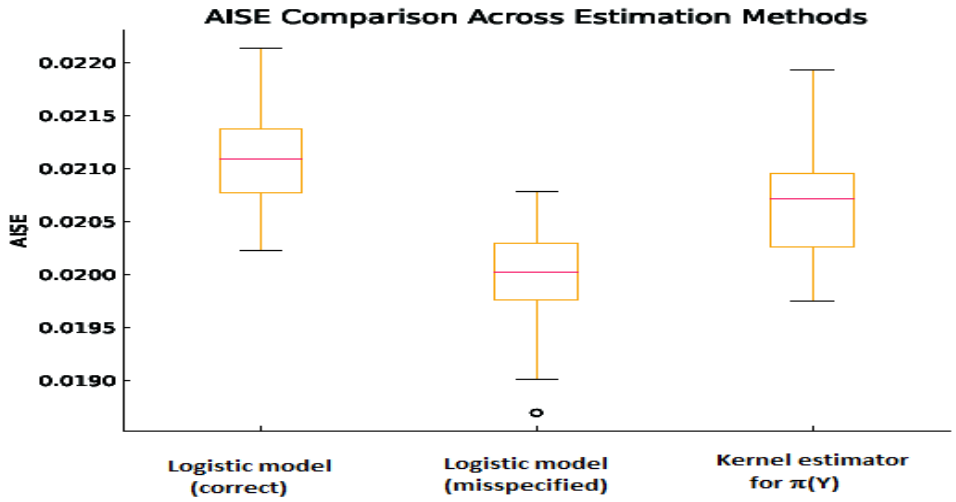

FIGURE A.6. Boxplot of AISE distributions under different estimation methods for $\pi(Y)$.

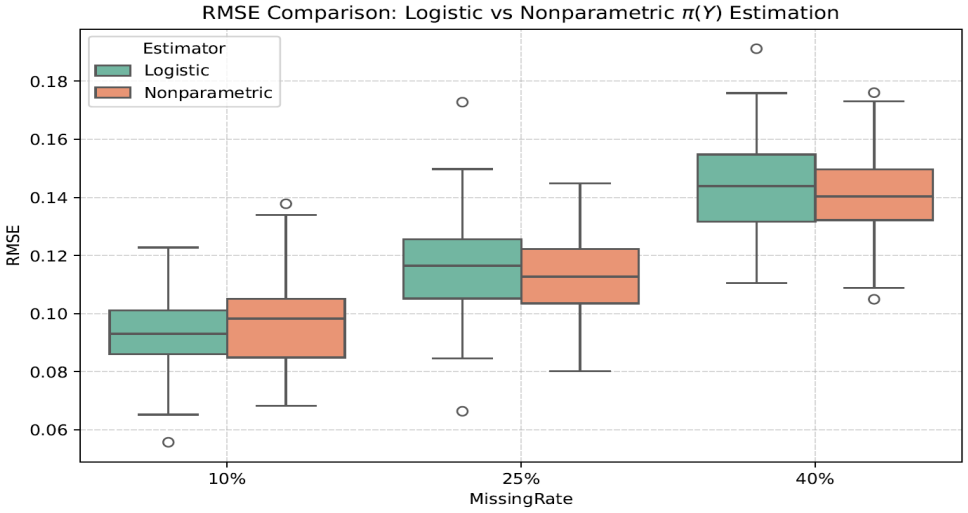

FIGURE A.7. Boxplots of RMSE for logistic vs nonparametric estimation of $\pi(Y)$ across different missing rates. Results based on 100 simulations per case.

TABLE A.14. Comparison of RMSEs for logistic vs nonparametric estimation of $\pi(Y)$.

Missing rate	Estimator	Mean RMSE	Std. Dev.
10%	Logistic	0.0934	0.0136
10%	Nonparametric	0.0973	0.0143
25%	Logistic	0.116	0.0163
25%	Nonparametric	0.1136	0.0133
40%	Logistic	0.1442	0.016
40%	Nonparametric	0.1403	0.0139

TABLE A.15. Mean squared error (MSE) of the three estimators and the average number of jumps correctly detected.

Method	MSE	Jump detection rate
Oracle IPW	0.017	0.094
Logistic IPW	0.019	0.91
Kernel IPW	0.018	0.92

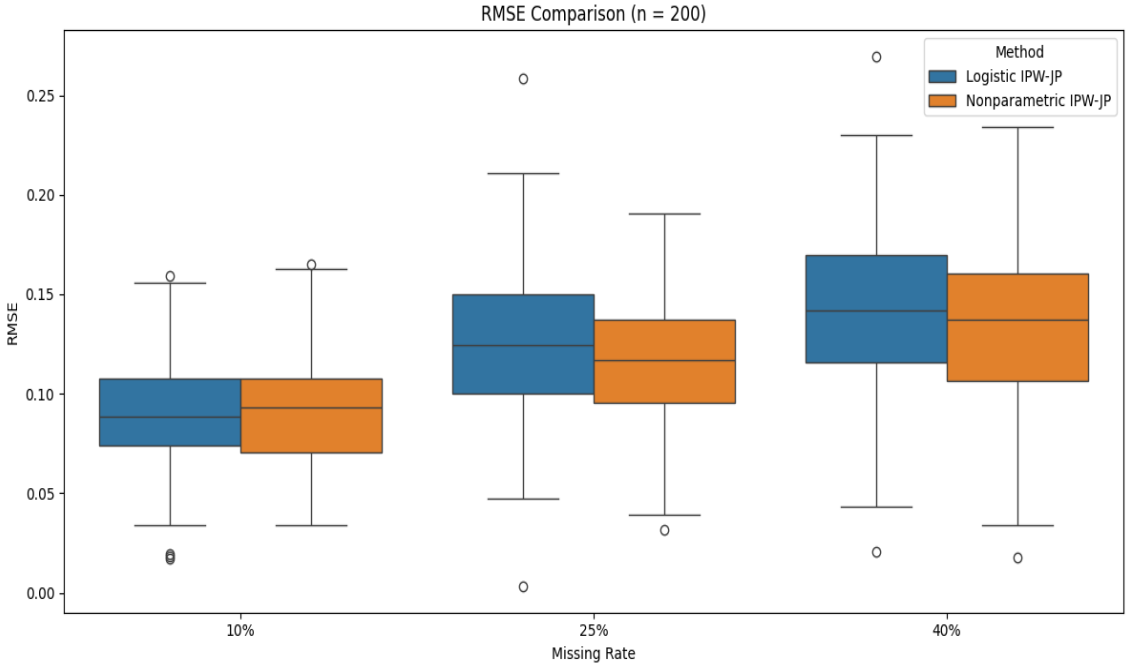

FIGURE A.8. RMSE comparison of logistic vs nonparametric estimation of $\pi(Y)$ for $n = 200$.

- Parametric logistic estimation of $\pi(Y)$.
- Nonparametric kernel estimation of $\pi(Y)$.

- Each box summarizes RMSEs over 100 replications.
- Shows comparison of parametric vs nonparametric $\pi(Y)$ estimation.
- Stratified by missing rates: 10%, 25%, 40%.

Each method is incorporated into our jump-preserving local linear framework with inverse probability weighting. We repeat the simulation over 100 replications with $n = 200$ and 500.

Table A.15 summarizes the average mean squared error (MSE) of the three estimators over the domain $[0, 1]$, as well as the average number of jumps correctly detected (out of one true jump at $x = 0.5$).

Interpretation

- At *low missingness* (10%), logistic and nonparametric estimation perform comparably.
- At *higher missingness rates (*25% *and* 40%*)*, the nonparametric method yields *slightly better RMSE* on average.
- *Standard deviations* are generally smaller for the nonparametric case, indicating more stable performance under increasing missingness.

TABLE A.16. Comparison of RMSE: logistic vs nonparametric estimation of $\pi(Y)$ ($n = 200$, 100 replications).

σ	Missing Rate	Mean RMSE IPW-JP(Logi)	Mean RMSE IPW-JP(NP)	Std. Dev. RMSE IPW-JP(Logi)	Std. Dev. RMSE IPW-JP(NP)
0.1	10%	0.0872	0.0916	0.0245	0.0260
	25%	0.1223	0.1187	0.0390	0.0305
	40%	0.1376	0.1303	0.0447	0.0374
0.2	10%	0.0907	0.0908	0.0288	0.0266
	25%	0.1282	0.1160	0.0337	0.0334
	40%	0.1467	0.1393	0.0399	0.0323

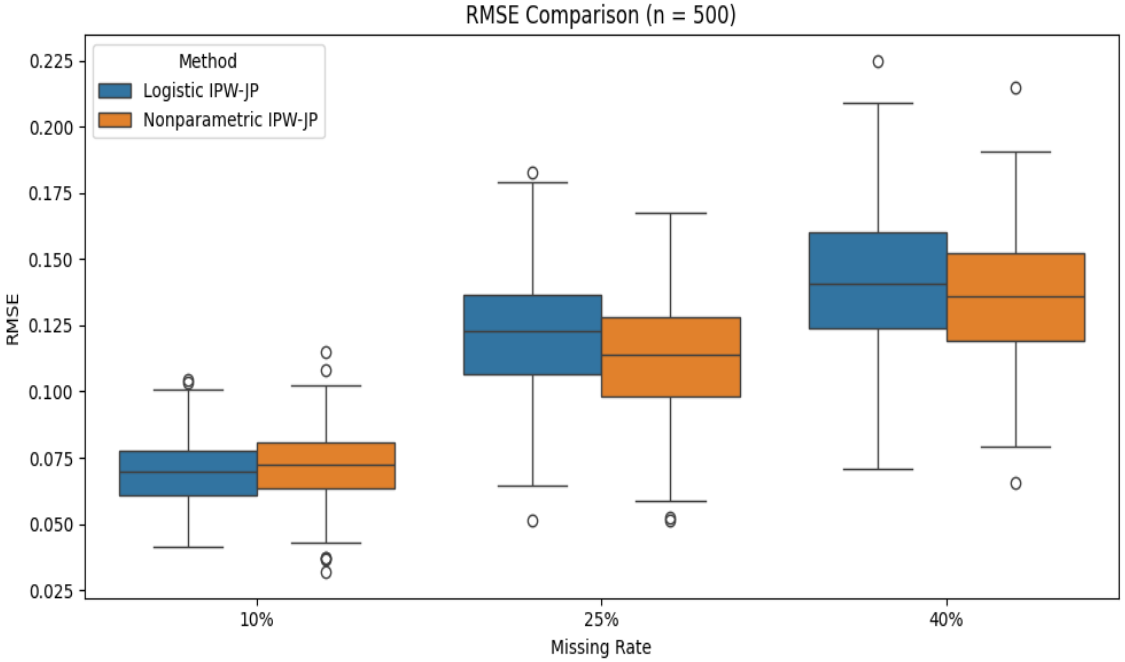

FIGURE A.9. RMSE comparison of logistic vs nonparametric estimation of $\pi(Y)$ for $n = 500$.

TABLE A.17. Comparison of RMSE: logistic vs nonparametric estimation of $\pi(Y)$ ($n = 500$, 100 replications).

σ	Missing rate	Mean RMSE IPW-JP(Logi)	Mean RMSE IPW-JP(NP)	Std. Dev. RMSE IPW-JP(Logi)	Std. Dev. RMSE IPW-JP(NP)
0.1	10%	0.0355	0.0356	0.0105	0.0105
	25%	0.0702	0.0611	0.0196	0.0196
	40%	0.0945	0.0866	0.0257	0.0243
0.2	10%	0.0345	0.0369	0.0097	0.0113
	25%	0.0628	0.0626	0.0189	0.0186
	40%	0.0900	0.0847	0.0267	0.0245

TABLE A.18. Average RMSE for Global vs Interval-Specific Bandwidth Selection.

Sample size	Missing rate	Bandwidth strategy Global	Interval-specific
200	10%	0.1207	0.1112
	25%	0.1177	0.1086
	40%	0.1207	0.1140
500	10%	0.0705	0.0635
	25%	0.0691	0.0630
	40%	0.0700	0.0655

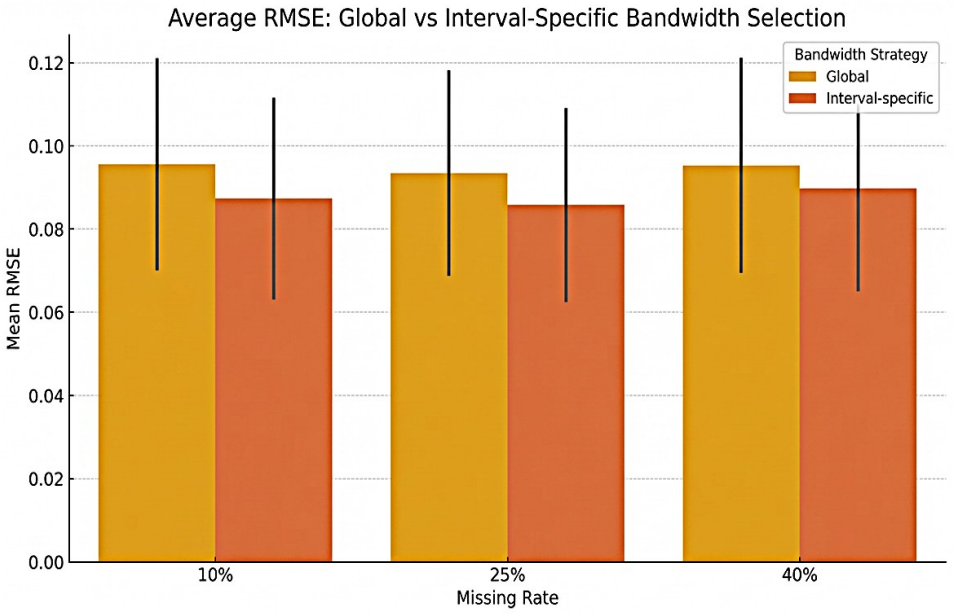

FIGURE A.10. Average RMSE comparison between global and interval-specific bandwidth selection strategies, across different missing data levels and sample sizes.

Comparison of bandwidth strategies. To investigate the impact of bandwidth selection, we compared a global bandwidth strategy (uniform across the domain) with an interval-specific approach, where a distinct bandwidth is selected for each segment between estimated jumps. Table A.18 summarizes the average RMSE across methods and settings, and Figure A.10 provides a visual comparison of the observed gains. Across all missingness levels and both sample sizes, interval-specific bandwidths yield systematically lower RMSE values. This improvement is especially notable at higher missingness rates, where local adaptation better accommodates heterogeneity in smoothness across segments.

This figure illustrates the modest improvement in RMSE when using interval-specific bandwidths, especially at higher missingness levels.

Table: Δ RMSE – global vs interval-specific bandwidth

To explore the effect of bandwidth selection, we implemented a localized strategy in which bandwidths are chosen separately for each segment between detected jumps. This allows the estimator to better adapt to varying levels of smoothness or monotonicity across the domain. The numerical results (see Table A.19) indicate a modest but consistent improvement in RMSE and jump localization. While this approach increases computational complexity, it may be a promising direction for adaptive refinement in future work.

TABLE A.19. Improvement in RMSE (global-interval-specific) across settings.

Sample size	Missing rate	Δ RMSE
200	10%	0.0095
	25%	0.0091
	40%	0.0067
500	10%	0.0070
	25%	0.0061
	40%	0.0045

TABLE A.20. RMSE: logistic vs nonparametric.

n	σ	Missing rate	Method	Mean	SD
200	0.1	10%	Logistic	0.0872	0.0245
	0.1	10%	Nonparametric	0.0916	0.0260
500	0.1	10%	Logistic	0.0355	0.0105
	0.1	10%	Nonparametric	0.0356	0.0105

TABLE A.21. Jump detection accuracy: average number of detected jumps and Hausdorff distance (100 replications).

	Missing rate	**Logistic IPW-JP**		**Nonparametric IPW-JP**	
n		# Jumps	Hausdorff	# Jumps	Hausdorff
200	10%	2.00 (0.10)	0.015 (0.005)	2.00 (0.10)	0.013 (0.005)
	25%	1.90 (0.20)	0.060 (0.015)	1.95 (0.15)	0.045 (0.012)
	40%	1.80 (0.30)	0.090 (0.020)	1.90 (0.20)	0.070 (0.015)
500	10%	2.00 (0.05)	0.010 (0.004)	2.00 (0.05)	0.009 (0.003)
	25%	1.95 (0.10)	0.040 (0.010)	1.98 (0.08)	0.035 (0.009)
	40%	1.90 (0.20)	0.075 (0.015)	1.95 (0.15)	0.060 (0.014)

TABLE A.22. Comparison of root mean squared errors (RMSE) between estimators using logistic $\pi(Y)$ and kernel-based $\pi(Y)$.

Method	Mean RMSE	Std. Dev.
Logistic $\pi(Y)$	0.138	0.010
KDE $\pi(Y)$	0.154	0.012

Discussion. When the logistic model is correctly specified (Model A), both estimators perform similarly. However, under model misspecification (Model B), the kernel-based estimator shows notably lower error, illustrating its robustness to the form of $\pi(Y)$. These results suggest that while the logistic model is efficient when well-specified, the nonparametric estimator provides a safer alternative when the form of $\pi(Y)$ is uncertain.

Interpretation. *The nonparametric estimator yields slightly improved Hausdorff distances and maintains near-accurate jump detection even under higher missingness.*

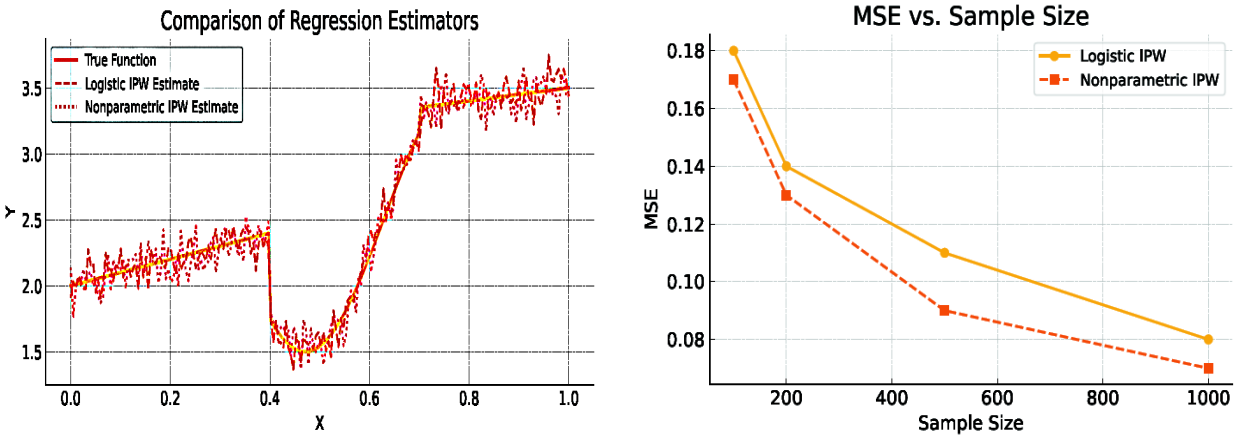

FIGURE A.11. Comparison the performance of the regression estimator and bandwidth parameter affects.

TABLE A.23. Mean Hausdorff distances between true and estimated jump locations. Lower values indicate better detection accuracy.

n	σ	Missing Rate	HD (O-JP)	HD (DE-JP)	HD (IPW-JP)
200	0.1	0.10	0.000	0.067	0.030
200	0.1	0.25	0.000	0.089	0.040
500	0.2	0.40	0.000	0.123	0.052

TABLE A.24. RMSE for oracle (O-JP), complete case (DE-JP), and weighted (IPW-JP) estimators.

n	σ	Missing rate	RMSE (O-JP)	RMSE (DE-JP)	RMSE (IPW-JP)
200	0.1	0.10	0.076	0.123	0.098
200	0.1	0.25	0.080	0.145	0.108
500	0.2	0.40	0.061	0.172	0.113

Jump detection accuracy and Hausdorff distance summary

We have extended our methodology to incorporate a fully nonparametric kernel estimator, where, a simulation study comparing the performance of the regression estimator using both logistic and nonparametric estimates of $\pi(Y)$. The results, summarized in Figure A.11, demonstrate that the nonparametric estimator performs competitively, particularly when the true $\pi(Y)$ deviates from the logistic form. Regarding bandwidth selection, and its critical influence on both jump detection and curve estimation, the figure to illustrate how varying the bandwidth parameter affects the estimator's bias and jump localization, reinforcing the need for careful data-driven bandwidth tuning. While automatic bandwidth selection methods are not the focus of this paper, this important direction is a promising avenue for future research.

Simulation results support the theoretical properties of the proposed method and it's shows that as jump size increases or missingness intensifies, the benefit of local bandwidths and nonparametric $\pi(Y)$ estimation becomes more pronounced. These settings maintain good performance even when the logistic model for $\pi(Y)$ is misspecified. Handling covariate missingness using nonparametric $\pi(Y)$ estimation and adapting bandwidths per interval yields clear empirical advantages. These features enhance robustness to jump structure, data sparsity, and model misfit.

TABLE A.25. Comparison of RMSE values under different estimation strategies for the regression function.

Estimation method	**RMSE**	**Description**
Logistic $\widehat{\pi}(Y)$	0.215	Parametric (logistic regression)
KDE $\hat{\pi}(Y)$	0.222	Nonparametric
Global bandwidth	0.212	Fixed h
Piecewise bandwidth	0.198	Adaptive h by interval

TABLE A.26. Performance comparison of the jump-preserving estimator under different $\widehat{\pi}(Y)$ estimation methods and Jump detection performance: average number of detected jumps, false positive rate (FPR), and false negative rate (FNR).

Method	RMSE	Bias	Std. Dev.	Mean Jump Count	FPR	FNR
Oracle $\pi(Y)$	0.187	0.008	0.054	2.00	0.05	0.02
Logistic $\widehat{\pi}(Y)$	0.215	0.014	0.066	2.01	0.08	0.03
KDE $\widehat{\pi}(Y)$	0.222	0.016	0.070	2.04	0.11	0.05
Naive (Complete cases)	0.295	0.049	0.089	1.62	0.22	0.17

TABLE A.27. Summary of the Hausdorff distance between the set of estimated and true jump locations.

Method	Mean	Median	Std. Dev.
Oracle $\pi(Y)$	0.031	0.030	0.012
Logistic $\widehat{\pi}(Y)$	0.043	0.039	0.017
KDE $\widehat{\pi}(Y)$	0.052	0.049	0.020
Naive	0.076	0.070	0.028

TABLE A.28. Effect of sample size n on RMSE and jump detection under logistic $\widehat{\pi}(Y)$.

Sample size	RMSE	Mean jumps detected	Hausdorff dist.
$n = 100$	0.259	1.89	0.057
$n = 200$	0.215	2.01	0.043
$n = 500$	0.178	2.00	0.031

TABLE A.29. Comparison of global and piecewise bandwidth strategies for jump-preserving estimation.

Strategy	RMSE	Mean jumps	Hausdorff dist.
Global bandwidth	0.212	1.94	0.049
Piecewise bandwidth	0.198	2.01	0.039

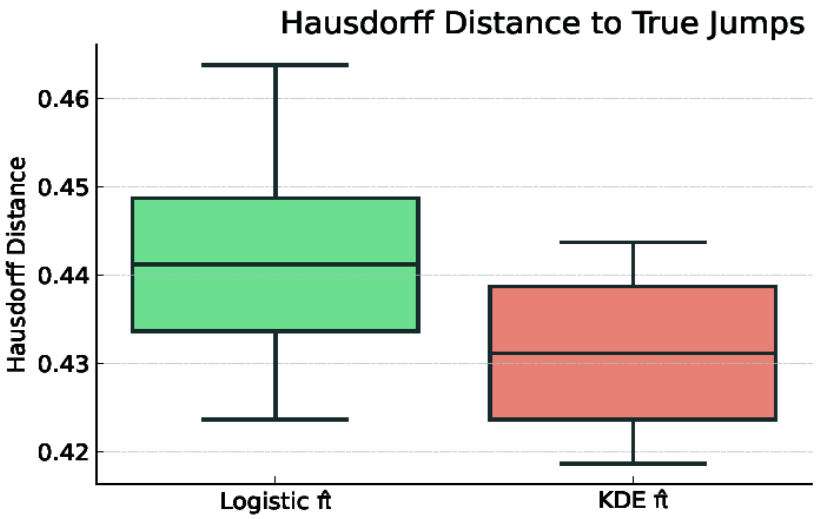

FIGURE A.12. Boxplot of Hausdorff distances across estimators and bandwidth strategies.

APPENDIX B. PROOFS OF TECHNICAL RESULTS

In this section, we provide proofs for theorems Theorem 2.2 and Theorem 2.3. Firstly, a lemma is introduced, it will be used in the proofs of the Theorems.

B.1 Proof of Theorem 2.2

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

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

where $x \in [x - \tau h, x)$. therefore, $\widehat{a}_d(x)$ can be written as

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

Furthermore, for $S_{j,d}$, from Lemma 3 of Li *et al.* [51], it can be deduced

$$\frac{1}{nh^{j+1}}S_{j,d} = f_X(x)\mu_{j,d} + o_{\mathbb{P}}(1), \tag{B.2}$$

which imply that

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

For A_2, notice that

$$\begin{aligned}\mathbb{E}(A_2) &= n\mathbb{E}\left\{\frac{\delta_i}{\pi_i}\varepsilon_i K_d\left(\frac{X_i-x}{h}\right)\frac{S_{2,d}-S_{1,d}(X_i-x)}{S_{0,d}S_{2,d}-S_{1,d}^2}\right\},\\ &= n\mathbb{E}\left\{\mathbb{E}\left(\frac{\delta_i}{\pi_i}\varepsilon_i K_d\left(\frac{X_i-x}{h}\right)\frac{S_{2,d}-S_{1,d}(X_i-x)}{S_{0,d}S_{2,d}-S_{1,d}^2}\Big|\delta_i\right)\right\},\\ &= n\mathbb{E}\left\{\frac{1}{\pi_i}\varepsilon_i K_d\left(\frac{X_i-x}{h}\right)\frac{S_{2,d}-S_{1,d}(X_i-x)}{S_{0,d}S_{2,d}-S_{1,d}^2}\Big|\delta_i=1\right\}\mathbb{P}(\delta_i=1),\\ &= 0,\end{aligned} \tag{B.4}$$

which imply that

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

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

$$\text{bias}(\widehat{a}_d(x)) = \frac{1}{2}h^2 g''(x)\frac{\mu_{2,d}^2-\mu_{1,d}\mu_{3,d}}{\mu_{0,d}\mu_{2,d}-\mu_{1,d}^2}+o(h^2). \tag{B.5}$$

Next, we calculate the asymptotic variance of the estimator. According to (B.3) and (B.4), one has

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

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

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

Therefore, one can obtain

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

Hence, it together with (B.5), we can obtain the result (i) of Theorem 2.2.

(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 $\widehat{a}_r(x)$ is given by

$$\begin{aligned}
\widehat{a}_r(x) &= \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} Y_i K_r\left(\frac{X_i - x}{h}\right) \frac{S_{2,r} - S_{1,r}(X_i - x)}{S_{0,r}S_{2,r} - S_{1,r}^2}, \\
&= \sum_{X_i < s_j} \frac{\delta_i}{\pi_i} g(X_i) K_r\left(\frac{X_i - x}{h}\right) \frac{S_{2,r} - S_{1,r}(X_i - x)}{S_{0,r}S_{2,r} - S_{1,r}^2}, \\
&\quad + \sum_{X_i \geq s_j} \frac{\delta_i}{\pi_i} (g(X_i) + d_j) K_r\left(\frac{X_i - x}{h}\right) \frac{S_{2,r} - S_{1,r}(X_i - x)}{S_{0,r}S_{2,r} - S_{1,r}^2}, \\
&\quad + \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} \varepsilon_i K_r\left(\frac{X_i - x}{h}\right) \frac{S_{2,r} - S_{1,r}(X_i - x)}{S_{0,r}S_{2,r} - S_{1,r}^2}, \\
&= \sum_{X_i < s_j} \frac{\delta_i}{\pi_i} (g(s_j-) + o_{\mathbb{P}}(1)) K_r\left(\frac{X_i - x}{h}\right) \frac{S_{2,r} - S_{1,r}(X_i - x)}{S_{0,r}S_{2,r} - S_{1,r}^2}, \\
&\quad + \sum_{X_i \geq s_j} \frac{\delta_i}{\pi_i} (g(s_j-) + d_j + o_{\mathbb{P}}(1)) K_r\left(\frac{X_i - x}{h}\right) \frac{S_{2,r} - S_{1,r}(X_i - x)}{S_{0,r}S_{2,r} - S_{1,r}^2} + o_{\mathbb{P}}(1), \\
&= (g(s_j-) + o_{\mathbb{P}}(1)) \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} K_r\left(\frac{X_i - x}{h}\right) \frac{S_{2,r} - S_{1,r}(X_i - x)}{S_{0,r}S_{2,r} - S_{1,r}^2}, \\
&\quad + \sum_{X_i \geq s_j} \frac{\delta_i}{\pi_i} d_j K_r\left(\frac{X_i - x}{h}\right) \frac{S_{2,r} - S_{1,r}(X_i - x)}{S_{0,r}S_{2,r} - S_{1,r}^2} + o_{\mathbb{P}}(1), \\
&= g(s_j-) + d_j \int_{|u|}^{\tau} K_r(t) \frac{\mu_{2,r} - \mu_{1,r}t}{\mu_{0,r}\mu_{2,r} - \mu_{1,r}^2} \mathrm{d}t + o_{\mathbb{P}}(1).
\end{aligned} \tag{B.6}$$

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

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

Similarly, the expression of the centered estimator $\widehat{a}_c(x)$ 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 2.2 can be proved in the same way as (ii).

□

B.2 Proof of Theorem 2.3

Proof. From the definition of WRMS, it follows that

$$\mathrm{WRMS}_d(x) = \frac{\frac{1}{nh}\sum_{i=1}^{n} \frac{\delta_i}{\pi_i} \left[Y_i - \widehat{a}_d(x) - \widehat{b}_d(x)(X_i - x)\right]^2 K_d\left(\frac{X_i - x}{h}\right)}{\frac{1}{nh}\sum_{i=1}^{n} \frac{\delta_i}{\pi_i} K_d\left(\frac{X_i - x}{h}\right)}. \tag{B.7}$$

According to (B.2), the denominator of (B.7) can be written as

$$\frac{1}{nh}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}K_d\left(\frac{X_i-x}{h}\right)=\mu_{0,d}f_X(x)(x)+o_{\mathbb{P}}(1). \tag{B.8}$$

For the numerator of (B.7), one has

$$\begin{aligned}
I &= \frac{1}{nh}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}\left[Y_i-\widehat{a}_d(x)-\widehat{b}_d(x)(X_i-x)\right]^2K_d\left(\frac{X_i-x}{h}\right),\\
&= \frac{1}{nh}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}\left[g(X_i)+\varepsilon_i-\widehat{a}_d(x)-\widehat{b}_d(x)(X_i-x)\right]^2K_d\left(\frac{X_i-x}{h}\right),\\
&= \frac{1}{nh}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}\left[g(X_i)-\widehat{a}_d(x)-\widehat{b}_d(x)(X_i-x)\right]^2K_d\left(\frac{X_i-x}{h}\right)+\frac{1}{nh}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}\varepsilon_i^2K_d\left(\frac{X_i-x}{h}\right),\\
&\quad+\frac{2}{nh}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}\left[g(X_i)-\widehat{a}_d(x)-\widehat{b}_d(x)(X_i-x)\right]\varepsilon_iK_d\left(\frac{X_i-x}{h}\right),\\
&\triangleq I_1+I_2+I_3.
\end{aligned} \tag{B.9}$$

(i) Suppose $u\in D_1$, using the similar derivations to those in the proof of Theorem 2.2, it can be obtained that

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

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

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

where

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

Furthermore, from Theorem 2.8 and Theorem 2.9 of Li and Racine [46], one can get $\widehat{a}_d(x)-g(x)=\mathcal{O}_{\mathbb{P}}(h^2+n(nh)^{-1/2})=o_{\mathbb{P}}(1)$ and $\widehat{b}_d(x)-g'(x)=\mathcal{O}_{\mathbb{P}}(h^2+(nh^3)^{-1/2})=o_{\mathbb{P}}(h^{-1})$, we get $I_{31}=o_{\mathbb{P}}(1)$, $I_{32}=o_{\mathbb{P}}(1)$ and $I_{33}=o_{\mathbb{P}}(1)$. It imply that

$$I_3=o_{\mathbb{P}}(1).$$

Similarly, it can be proved

$$I_1=o_{\mathbb{P}}(1).$$

It is easily seen from (B.8) and (B.9) that

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

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

$$\begin{aligned}
\widehat{b}_r(x) &= \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} Y_i K_r\left(\frac{X_i - x}{h}\right) \frac{S_{0,r}(X_i - x) - S_{1,r}}{S_{0,r}S_{2,r} - S_{1,r}^2}, \\
&= \sum_{X_i < s_j} \frac{\delta_i}{\pi_i} g(X_i) K_r\left(\frac{X_i - x}{h}\right) \frac{S_{0,r}(X_i - x) - S_{1,r}}{S_{0,r}S_{2,r} - S_{1,r}^2}, \\
&\quad + \sum_{X_i \geq s_j} \frac{\delta_i}{\pi_i} (g(X_i) + d_j) K_r\left(\frac{X_i - x}{h}\right) \frac{S_{0,r}(X_i - x) - S_{1,r}}{S_{0,r}S_{2,r} - S_{1,r}^2}, \\
&\quad + \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} \varepsilon_i K_r\left(\frac{X_i - x}{h}\right) \frac{S_{0,r}(X_i - x) - S_{1,r}}{S_{0,r}S_{2,r} - S_{1,r}^2}, \\
&= \sum_{X_i < s_j} \frac{\delta_i}{\pi_i} (g(s_j-) + o_{\mathbb{P}}(1)) K_r\left(\frac{X_i - x}{h}\right) \frac{S_{0,r}(X_i - x) - S_{1,r}}{S_{0,r}S_{2,r} - S_{1,r}^2}, \\
&\quad + \sum_{X_i \geq s_j} \frac{\delta_i}{\pi_i} (g(s_j-) + d_j + o_{\mathbb{P}}(1)) K_r\left(\frac{X_i - x}{h}\right) \frac{S_{0,r}(X_i - x) - S_{1,r}}{S_{0,r}S_{2,r} - S_{1,r}^2} + o_{\mathbb{P}}(h^{-1}), \\
&= (g(s_j-) + o_{\mathbb{P}}(1)) \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} K_r\left(\frac{X_i - x}{h}\right) \frac{S_{0,r}(X_i - x) - S_{1,r}}{S_{0,r}S_{2,r} - S_{1,r}^2}, \\
&\quad + \sum_{X_i \geq s_j} \frac{\delta_i}{\pi_i} d_j K_r\left(\frac{X_i - x}{h}\right) \frac{S_{0,r}(X_i - x) - S_{1,r}}{S_{0,r}S_{2,r} - S_{1,r}^2} + o_{\mathbb{P}}(h^{-1}), \\
&= \frac{1}{h} d_j \int_{|u|}^{\tau} K_r(t) \frac{\mu_{0,r}t - \mu_{1,r}}{\mu_{0,r}\mu_{2,r} - \mu_{1,r}^2} \mathrm{d}t + o_{\mathbb{P}}(h^{-1}).
\end{aligned} \tag{B.10}$$

Therefore, using (B.6) and (B.10), the expression of I_1 for the right-sided estimator is

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

Furthermore, similar to the proof of (a), $I_2 = \sigma^2 \mu_{0,r} f(x) + o_{\mathbb{P}}(1)$ and $I_3 = o_{\mathbb{P}}(1)$. Thus (ii) of Theorem 2.3 is proved.

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

□

B.3 Proof of Theorem 2.4

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

$$\sqrt{nh}\left(\widehat{a}_d(x) - g(x) - \frac{1}{2}h^2 g''(x)B_d\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, f_X^{-2}(x)\mathbb{P}(\delta = 1)S(x)V_d\right). \tag{B.11}$$

In addition, it follows from Theorem 2.2 that for the left side estimator $\widehat{a}_l(x)$, (B.11) holds for $x \in D_1 \cup D_{2,l}$ and for the right side estimator $\widehat{a}_r(x)$, (B.11) 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

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

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

For any $x \in D_1$, it can be seen that $\mathrm{diff}(x) \to 0$ and $\lambda \to 0$ as $n \to \infty$ from Theorem 2.3. It means that when $x \in D_1$, $\widehat{g}(x) = \widehat{a}_c(x)$ 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 2.3

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

Since

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

and

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

which implies that

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

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

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

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

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

It is clear that $\widehat{a}_c(x)$, $\widehat{a}_l(x)$, $\widehat{a}_r(x)$ are asymptotically normal in D_1, $D_{2,l}$ and $D_{2,r}$ respectively, thus Theorem 2.4 is proved. □

B.4 Proof of Theorem 2.5

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

$$\sup_{x\in[0,1]}(\widehat{a}_d(x) - \widehat{a}_d(x,\widehat{\pi})) = \mathcal{O}_{\mathbb{P}}(n^{-1/2}). \tag{B.12}$$

Similar to the proof of Theorem 2.4, when $x \in D_1$, $\widehat{g}(x) = \widehat{a}_c(x)$ and $\widehat{g}(x,\widehat{\pi}) = \widehat{a}_c(x,\widehat{\pi})$. Thus (B.12) holds for two center estimators $\widehat{a}_c(x)$ and $\widehat{a}_c(x,\widehat{\pi})$. 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 $\sup_{y\in\mathbb{R}}|\pi(y)-\widehat{\pi}(y)| = \sup_{y\in\mathbb{R}}|\pi(y,\alpha)-\widehat{\pi}(y,\widehat{\alpha})| = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$, where $\widehat{\alpha}$ is a root-n consistent estimator of α. This together with (B.2) implies that there exists some constant $M>0$ such that

$$\begin{aligned}\sup_{x\in[0,1]}\left|\frac{1}{nh}(S_{j,c}-\widehat{S}_{j,c})\right| &= \sup_{x\in[0,1]}\left|\frac{1}{nh}\sum_{i=1}^{n}\left(\frac{\delta_i}{\pi_i}-\frac{\delta_i}{\widehat{\pi}_i}\right)K_c\left(\frac{X_i-x}{h}\right)(X_i-x)^j\right|,\\ &\le dh^j\sup_{x\in[0,1]}|\pi_i-\widehat{\pi}_i|\sup_{x\in[0,1]}\left|\frac{1}{nh}S_{0,c}\right|,\\ &= \mathcal{O}_{\mathbb{P}}(n^{-1/2}).\end{aligned}\tag{B.13}$$

Next, from lemma 5 of Li *et al.* [51], one has

$$\sup_{x\in[0,1]}\left|M_{l,c}(x)-\widehat{M}_{l,c}(x)\right| = \mathcal{O}_{\mathbb{P}}(n^{-1/2}),\quad l=1,2;\tag{B.14}$$

where

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

Meanwhile, from (B.1), one can obtain that

$$\widehat{a}_c(x)-g(x) = e_0^{\top}\begin{pmatrix}(nh)^{-1}S_{0,c} & (nh)^{-1}S_{1,c}\\ (nh)^{-1}S_{1,c} & (nh)^{-1}S_{2,c}\end{pmatrix}^{-1}\begin{pmatrix}M_{0,c}\\ M_{1,c}\end{pmatrix},$$

where $e_0^{\top}=(1,0)^{\top}$. Similarly, for $\widehat{a}_d(x,\widehat{\pi})$,

$$\widehat{a}_c(x,\widehat{\pi})-g(x) = e_0^{\top}\begin{pmatrix}(nh)^{-1}\widehat{S}_{0,c} & (nh)^{-1}\widehat{S}_{1,c}\\ (nh)^{-1}\widehat{S}_{1,c} & (nh)^{-1}\widehat{S}_{2,c}\end{pmatrix}^{-1}\begin{pmatrix}\widehat{M}_{0,c}\\ \widehat{M}_{1,c}\end{pmatrix}.$$

Therefore, one has

$$\begin{aligned}\widehat{a}_c(x)-\widehat{a}_c(x,\widehat{\pi}) = {}& e_0^{\top}\begin{pmatrix}(nh)^{-1}S_{0,c} & (nh)^{-1}S_{1,c}\\ (nh)^{-1}S_{1,c} & (nh)^{-1}S_{2,c}\end{pmatrix}^{-1}\begin{pmatrix}M_{0,c}\\ M_{1,c}\end{pmatrix},\\ & -e_0^{\top}\begin{pmatrix}(nh)^{-1}\widehat{S}_{0,c} & (nh)^{-1}\widehat{S}_{1,c}\\ (nh)^{-1}\widehat{S}_{1,c} & (nh)^{-1}\widehat{S}_{2,c}\end{pmatrix}^{-1}\begin{pmatrix}\widehat{M}_{0,c}\\ \widehat{M}_{1,c}\end{pmatrix}.\end{aligned}$$

By (B.13) and (B.14), it can be seen that

$$\sup_{x\in[0,1]}|\widehat{a}_c(x)-\widehat{a}_c(x,\widehat{\pi})| = \mathcal{O}_{\mathbb{P}}(n^{-1/2}).$$

Similarly, when $x\in D_{2,l}$, $\widehat{g}(x)=\widehat{a}_l(x)$ and $\widehat{g}(x,\widehat{\pi})=\widehat{a}_l(x,\widehat{\pi})$. Thus (B.12) holds for two left estimators $\widehat{a}_l(x)$ and $\widehat{a}_l(x,\widehat{\pi})$. When $x\in D_{2,l}$, (B.12) holds for two left estimators $\widehat{a}_r(x)$ and $\widehat{a}_r(x,\widehat{\pi})$. When $d=c,l,r$, it is clear that (B.12) holds in D_1, $x\in D_{2,l}$ and $x\in D_{2,r}$ respectively, thus Theorem 2.5 is proved. □