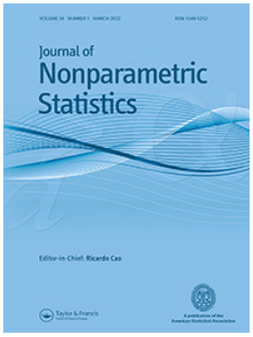

Taylor & Francis
Taylor & Francis Group

Journal of Nonparametric Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gnst20

Nonparametric quantile regression with missing data using local estimating equations

Chunyu Wang, Maozai Tian & Man-Lai Tang

To cite this article: Chunyu Wang, Maozai Tian & Man-Lai Tang (2022) Nonparametric quantile regression with missing data using local estimating equations, Journal of Nonparametric Statistics, 34:1, 164-186, DOI: 10.1080/10485252.2022.2026353

To link to this article: https://doi.org/10.1080/10485252.2022.2026353

Published online: 31 Jan 2022.

JOURNAL OF NONPARAMETRIC STATISTICS
2022, VOL. 34, NO. 1, 164–186
https://doi.org/10.1080/10485252.2022.2026353

Taylor & Francis
Taylor & Francis Group

Nonparametric quantile regression with missing data using local estimating equations

Chunyu Wang[a], Maozai Tian [a,b,c,d] and Man-Lai Tang[e,f]

[a]Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, People's Republic of China; [b]Department of Medical Engineering and Technology, Xinjiang Medical University, Ürümqi, People's Republic of China; [c]School of Statistics, Lanzhou University of Finance and Economics, Lanzhou, People's Republic of China; [d]School of Statistics and Information, Xinjiang University of Finance, Ürümqi, People's Republic of China; [e]Department of Mathematics, College of Engineering, Design Physical Sciences, Brunel University London, Uxbridge, UK; [f]Department of Mathematics, Statistics and Insurance, Hang Seng University of Hong Kong, Siu Lek Yuen, Hong Kong

ABSTRACT

In this paper, we propose augmented inverse probability weighted (AIPW) local estimating equations in dealing with missing data in nonparametric quantile regression context. The missing mechanism here is missing at random. To avoid the problem of misspecification, we adopt nonparametric approach to estimate the propensity score and conditional expectations of estimating functions. The asymptotic properties of our proposed estimator are studied. Majorisation–minimisation algorithm is used to circumvent the nonsmoothness of check function at the origin. When it comes to the choice of bandwidth, the theoretical expression of local optimal bandwidth is derived based on asymptotic properties. Moreover, we apply smoothed bootstrap method to obtain the empirical mean square error and use cross-validation to determine the bandwidth in practice. Simulations are conducted to compare the performance of our proposed methods with other existing methods. Finally, we illustrate our methodology with an analysis of non-insulin-dependent diabetes mellitus data set.

ARTICLE HISTORY
Received 28 February 2020
Accepted 28 December 2021

KEYWORDS
Missing data; augmented inverse probability weighted method; local estimating equations; nonparametric quantile regression

2020 MATHEMATICS SUBJECT CLASSIFICATION
62G08

1. Introduction

The prevalence of diabetes mellitus among Pima Indians has been given continuous study. Although it is known that the high incidence rate of diabetes is strongly related to obesity and parental diabetes (genetic factor) (Knowler, Pettitt, Savage, and Bennett 1981), the functional form of the dependence is unknown and expected to be nonlinear. We are hence interested in modelling the effect of obesity and genetics on the plasma glucose level nonparametrically. Instead of exploring the effects on mean values of distribution of plasma glucose, we prefer to examine the effects on various quantiles of the distribution using quantile regression (QR) methods. However, different with the traditional nonparametric QR analysis, our study here is complicated by the occurrence of missing values. Among all

CONTACT Man-Lai Tang man.tang@brunel.ac.uk, mltang@hsu.edu.hk

the 632 observations in the data set, the obesity indicator (triceps skin fold thickness) is not available for 98 subjects. Furthermore, it appears that the most ideal missing mechanism, missing completely at random (MCAR), is not applicable in this situation, since the propensity of missingness is to some extent associated with the magnitude of obesity and the level of plasma glucose. Motivated by this example, we develop flexible nonparametric quantile regression models in the presence of missing data.

Missing data are often encountered in many studies. Standard analyses that rely solely on subjects with complete data may yield biased conclusion as these subjects may form an unrepresentative subgroup. We assume that the data are missing at random (MAR; Little and Rubin 2002), which implies that, conditional on the observed, missingness and the unobserved data are independent. Imputation and inverse probability weighting (IPW) are two widely adopted approaches for handling missing data. The basic idea of imputation is to replace missing values with imputed values and then to perform statistical inference based on the imputed sample. The imputation method often requires the specification of a joint or conditional likelihood and the final estimation obtained in this way tends to be biased when the likelihood is misspecified (Zhou, Wan, and Wang 2008; Han 2016). IPW method, however, weights the contribution of each complete individual by the inverse probability of being fully observed and then performs the analysis on the weighted data. In most literatures, this pivotal probability is called propensity score. Robins, Rotnitzky, and Zhao (1994) assumed a parametric model, logistic regression model, for the propensity score and demonstrated that consistent estimators can be obtained if the model is correctly specified. Wang, Wang, Zhao, and Ou (1997) proposed a nonparametric kernel smoother to overcome the problem of misspecification and proved that the convergence rate of final estimators is the same as that in Robins et al. (1994). In other word, estimating the propensity score via nonparametric method does not lower the final convergence rate.

Augmented IPW (AIPW) method is a natural generalisation of IPW method by adding an augmented term which includes additional contributions from individuals with missing data (Tsiatis 2007). Compared to parametric IPW method, parametric AIPW method has the advantage of yielding estimators with double consistency (i.e. they are consistent if either the propensity score or the distribution of missing variables conditional on the observed variables is correctly specified). Furthermore, if both are correctly specified, parametric AIPW estimators will be more efficient. Although double consistency makes parametric AIPW method more flexible, correct specification is still an inevitable challenge. Chen, Wan, and Zhou (2015) proposed the nonparametric AIPW method based on the projection approach in Zhou et al. (2008). They considered nonparametric smoother in estimating both the propensity score and the augmented term.

Existing research on missing data mainly focuses on parametric or nonparametric regression models in conditional mean framework. Wang, Wang, Gutierrez, and Carroll (1998) investigated the missing covariate issue in nonparametric generalised linear model for mean function. Wang, Rotnitzky, and Lin (2010) considered nonparametric mean regression with missing outcomes using weighted kernel estimating equations. Limited attention has been paid to nonparametric QR context. For nonparametric QR model with full data, Fan, Hu, and Truong (1994) proposed the kernel estimator based on local linear fitting and investigated its theoretical properties in detail. Hu, Yang, Wang, and Tian (2017) developed a multiple imputation procedure in nonparametric QR when discrete covariates are MAR. However, their approach is applicable only to cases where

missing comes from covariates. Further explorations are needed when responses or both responses and covariates suffer from missing.

In this paper, we focus on nonparametric quantile regression models in the presence of missing data and develop AIPW local estimating equations to deal with this problem. The outline of this paper is as follows. In Section 2, we present the nonparametric QR model with missing data. AIPW method is proposed in Section 3 with kernel smoother to avoid misspecification for estimating the propensity score and the augmented term. In Section 4, we utilise majorisation–minimisation (MM) algorithm to circumvent the nonsmoothness of the corresponding equations. Asymptotic properties of the estimators obtained by solving local estimating equations are presented in Section 5 and a smoothed booststrap approach is given to implement bandwidth selection in practice. In Section 6, we compare our proposed method with other methods through simulation studies. An application of our methodology to analyse diabetes data is shown in Section 7, followed by a discussion in Section 8.

2. Nonparametric QR with missing data

Let $\{(\boldsymbol{Z}_1, Y_1), \ldots, (\boldsymbol{Z}_n, Y_n)\}$ be a random sample from the population $(\boldsymbol{Z}, Y)$ with Y being a scalar response variable and $\boldsymbol{Z}$ being a d-dimensional explanatory vector. We consider the following nonparametric QR model

$$Y = \theta_\tau(\boldsymbol{Z}) + \epsilon, \tag{1}$$

where $\theta_\tau(\cdot)$ is an unknown smooth function and ϵ is an unobserved random variable whose τth conditional quantile is 0 (i.e. $P(\epsilon \leq 0 \mid \boldsymbol{Z}) = \tau$). It can be easily seen from (1) that the τth quantile of Y given $\boldsymbol{Z} = \boldsymbol{z}$ depends on $\boldsymbol{Z}$ through the following model

$$Q_{Y|\boldsymbol{Z}=\boldsymbol{z}}(\tau) \equiv F^{-1}_{Y|\boldsymbol{Z}=\boldsymbol{z}}(\tau) = \theta_\tau(\boldsymbol{z}).$$

This paper aims to estimate $\theta_\tau(\boldsymbol{z})$ in the presence of missing data. In order to distinguish between fully observed and missing variables, we rearrange $(\boldsymbol{Z}, Y)$ into $(\boldsymbol{X}^o, \boldsymbol{X}^m)$, where $\boldsymbol{X}^o$, a d_1-dimensional random vector with $0 < d_1 \leq d$, contains those variables that can be observed on each individual while $\boldsymbol{X}^m$ contains those variables that cannot be observed (i.e. missing) for some individuals. It should be noted that missing observations may be induced from the response and/or explanatory variables under this unified framework (Chen et al. 2015). Let δ_i be an indicator variable such that $\delta_i = 1$ if $\boldsymbol{X}_i^m$ is fully observed, and $\delta_i = 0$ otherwise. As a result, the observed data can be denoted by $\{(\delta_1, \boldsymbol{X}_1^o, \delta_1\boldsymbol{X}_1^m), \ldots, (\delta_n, \boldsymbol{X}_n^o, \delta_n\boldsymbol{X}_n^m)\}$. See Appendix 1 for an illustration of notations described above. In this article, we assume that the missing mechanism is MAR; i.e. the propensity score depends only on those variables that are observed

$$P(\delta_i = 1 \mid Y_i, \boldsymbol{Z}_i) = P(\delta_i = 1 \mid \boldsymbol{X}_i^o, \boldsymbol{X}_i^m) = P(\delta_i = 1 \mid \boldsymbol{X}_i^o). \tag{2}$$

That is, conditional on $\boldsymbol{X}_i^o$, δ_i is independent of $\boldsymbol{X}_i^m$.

3. Augmented inverse probability weighted (AIPW) methods

Let $\mathcal{K}_{\mathbf{H}}(\cdot) = \frac{1}{\det(\mathbf{H})}\mathcal{K}(\mathbf{H}^{-1}\cdot)$, where $\mathcal{K}(\cdot)$ is a multidimensional kernel function and $\mathbf{H}$ is a bandwidth matrix. For simplicity, we use the special form: $\mathcal{K}(\boldsymbol{u}) = K(u_1)\cdots K(u_d)$ and

$\mathbf{H} = hI_d$, where $K(\cdot)$ is an ordinary univariate kernel function and h is a common bandwidth. If there is no missing data, the local linear kernel estimator for $\theta_\tau(\boldsymbol{z})$ can be readily obtained by solving the following minimisation problem

$$\min_{\alpha,\boldsymbol{\beta}} \sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}(\boldsymbol{Z}_i - \boldsymbol{z})\rho_\tau(Y_i - \alpha - (\boldsymbol{Z}_i - \boldsymbol{z})^\top \boldsymbol{\beta}), \tag{3}$$

where $\rho_\tau(t) = t(\tau - I(t < 0))$ is the check function and $\boldsymbol{\beta}$ is a d-dimensional vector. Obviously, the first component of the solution to the above optimisation problem is the estimator of $\theta_\tau(\boldsymbol{z})$, which is asymptotically unbiased and achieves $\sqrt{nh^d}$ consistency (Fan et al. 1994; Ruppert and Wand 1994). Like estimating equations for parametric problems, Carroll, Ruppert, and Welsh (1998) developed a similar method (i.e. local estimating equations) to deal with the nonparametric problem. Let

$$g_{\boldsymbol{z}}(Y_i, \boldsymbol{Z}_i, \alpha, \boldsymbol{\beta}) \equiv \mathcal{K}_{\mathbf{H}}(\boldsymbol{Z}_i - \boldsymbol{z})\varphi_\tau(Y_i - \alpha - (\boldsymbol{Z}_i - \boldsymbol{z})^\top \boldsymbol{\beta})U(\boldsymbol{Z}_i - \boldsymbol{z}), \tag{4}$$

where $\varphi_\tau(t) = \tau - I(t < 0)$ and $U(\boldsymbol{u}) = (1, \boldsymbol{u}^\top)^\top$ for any d-dimensional vector $\boldsymbol{u}$. The resultant minimiser of (3) is also the solution of the following local estimating equations

$$\sum_{i=1}^{n} g_{\boldsymbol{z}}(Y_i, \boldsymbol{Z}_i, \alpha, \boldsymbol{\beta}) = \mathbf{0}. \tag{5}$$

To adopt the above local estimating equations for nonparametric QR with missing data, we define the following AIPW kernel estimating equations

$$\sum_{i=1}^{n} \left[\frac{\delta_i}{\hat{\pi}_i} g_{\boldsymbol{z}}(Y_i, \boldsymbol{Z}_i, \alpha, \boldsymbol{\beta}) + \left(1 - \frac{\delta_i}{\hat{\pi}_i}\right) \hat{m}_{\boldsymbol{z}}(\boldsymbol{X}_i^o, \alpha, \boldsymbol{\beta}) \right] = \mathbf{0}, \tag{6}$$

where $\hat{\pi}_i$ is the estimator of the propensity score $\pi_i \equiv \pi(\boldsymbol{X}_i^o) = P(\delta_i = 1 \mid \boldsymbol{X}_i^o)$, and $\hat{m}_{\boldsymbol{z}}(\boldsymbol{X}_i^o, \alpha, \boldsymbol{\beta})$ is the estimator of $m_{\boldsymbol{z}}(\boldsymbol{X}_i^o, \alpha, \boldsymbol{\beta}) = E[g_{\boldsymbol{z}}(Y_i, \boldsymbol{Z}_i, \alpha, \boldsymbol{\beta}) \mid \boldsymbol{X}_i^o]$; i.e. the projection of $g_{\boldsymbol{z}}(Y_i, \boldsymbol{Z}_i, \alpha, \boldsymbol{\beta})$ into the space generated by the fully observed data (Zhou et al. 2008). For the computation of $\hat{\pi}_i$ and $\hat{m}_{\boldsymbol{z}}(\boldsymbol{X}_i^o, \alpha, \boldsymbol{\beta})$, there are two main approaches, namely, the parametric and nonparametric approaches. Since π_i is the conditional expectation of a binary variable, it is natural to resort to logistic regression to investigate the dependence relationship. However, it takes the risk of misspecification, which may then lead to biased estimate, if a parametric method is used. To circumvent this issue, we consider the following Nadaraya–Watson estimator, which is the most intuitive kernel smoother

$$\hat{\pi}_i = \frac{\sum_{j=1}^{n} \widetilde{\mathcal{K}}_{\tilde{h}}(\boldsymbol{X}_j^o - \boldsymbol{X}_i^o)\delta_j}{\sum_{j=1}^{n} \widetilde{\mathcal{K}}_{\tilde{h}}(\boldsymbol{X}_j^o - \boldsymbol{X}_i^o)}, \tag{7}$$

where $\widetilde{\mathcal{K}}_{\tilde{h}}(\boldsymbol{u}) = \widetilde{\mathcal{K}}(\boldsymbol{u}/\tilde{h})/\tilde{h}^{d_1}$, $\widetilde{\mathcal{K}}(\cdot)$ is a d_1-dimensional kernel function and $\tilde{h}$ is a bandwidth parameter. Under the MAR assumption, δ is conditionally independent of $\boldsymbol{X}^m$ given $\boldsymbol{X}^o$. Consequently, we have

$$E(g_{\boldsymbol{z}}(Y, \boldsymbol{Z}, \alpha, \boldsymbol{\beta}) \mid \boldsymbol{X}^o) = E(g_{\boldsymbol{z}}(Y, \boldsymbol{Z}, \alpha, \boldsymbol{\beta}) \mid \delta, \boldsymbol{X}^o) = E(g_{\boldsymbol{z}}(Y, \boldsymbol{Z}, \alpha, \boldsymbol{\beta}) \mid \delta = 1, \boldsymbol{X}^o),$$

which implies that it is sufficient to construct the estimator of $m_{\boldsymbol{z}}(\boldsymbol{X}_i^o, \alpha, \boldsymbol{\beta})$ with the complete cases $\{i : \delta_i = 1\}$ only. For this purpose, we adopt the estimator suggested in Zhou

et al. (2008),

$$\hat{m}_{\boldsymbol{z}}(\boldsymbol{X}_i^o, \alpha, \boldsymbol{\beta}) = \frac{\sum_{j=1}^n \overline{\mathcal{K}}_{\bar{h}}(\boldsymbol{X}_j^o - \boldsymbol{X}_i^o)\delta_j g_{\boldsymbol{z}}(Y_j, \boldsymbol{Z}_j, \alpha, \boldsymbol{\beta})}{\sum_{j=1}^n \overline{\mathcal{K}}_{\bar{h}}(\boldsymbol{X}_j^o - \boldsymbol{X}_i^o)\delta_j}, \tag{8}$$

where $\overline{\mathcal{K}}_{\bar{h}}(\boldsymbol{u}) = \overline{\mathcal{K}}(\boldsymbol{u}/\bar{h})/\bar{h}^{d_1}$, $\overline{\mathcal{K}}(\cdot)$ is a d_1-dimensional kernel function and $\bar{h}$ is a bandwidth parameter.

It is worth noting that our method is restricted to the situation, where $\boldsymbol{X}^o$ is nonnull ($d_1 > 0$ as mentioned in Section 2). That is, there must be at least one variable which is observable for all individuals. For the random vector $\boldsymbol{X}^m$, a variable which is missing for some (maybe a small portion) individuals should be incorporated as an element in $\boldsymbol{X}^m$. Moreover, $\{(\boldsymbol{X}_i^o, \boldsymbol{X}_i^m)\}_{i=1}^n$ are in fact i.i.d. samples from the population $(\boldsymbol{X}^o, \boldsymbol{X}^m)$, which is required when using the kernel-based estimators such as (7) and (8).

4. Computation

To circumvent the nonsmoothness issue encountered in solving AIPW kernel equations (6), we utilise the generalised MM algorithm developed by Hunter and Lange (2000). The idea of a standard MM algorithm is to create a surrogate function (majoriser) and then to iteratively optimise the sequence of surrogate functions instead of the original objective function. A typical example of MM algorithm is EM algorithm, in which the conditional expectation of complete-data log-likelihood is chosen as the surrogate of true log-likelihood function. When it comes to quantile regression, the situation becomes more complicated due to the fact that the check function is nondifferentiable at the origin. Hunter and Lange (2000) proposed to first construct a differentiable approximation of the objective function (sum of check functions) and then use a standard MM algorithm to minimise the approximating function. Specifically, consider the following quadratic function

$$\xi_\tau^\varepsilon(r_{iz}(\alpha, \boldsymbol{\beta}) \mid r_{iz}^{(k)}) = \frac{1}{4}\left[\frac{r_{iz}^2(\alpha, \boldsymbol{\beta})}{\varepsilon + |r_{iz}^{(k)}|} + (4\tau - 2)r_{iz}(\alpha, \boldsymbol{\beta}) + c\right],$$

as the surrogate function for $\rho_\tau^\varepsilon(r_{iz}(\alpha, \boldsymbol{\beta}))$, where $\rho_\tau^\varepsilon(r_{iz}(\alpha, \boldsymbol{\beta})) = \rho_\tau(r_{iz}(\alpha, \boldsymbol{\beta})) - \frac{\varepsilon}{2}\log(\varepsilon + |r_{iz}(\alpha, \boldsymbol{\beta})|)$ is a perturbation (approximating function) of the check function $\rho_\tau(r_{iz}(\alpha, \boldsymbol{\beta}))$, ε is the perturbation constant, $r_{iz}(\alpha, \boldsymbol{\beta}) = Y_i - \alpha - (\boldsymbol{Z}_i - \boldsymbol{z})^\top \boldsymbol{\beta}$ is the residual with $r_{iz}^{(k)} = r_{iz}(\alpha^{(k)}, \boldsymbol{\beta}^{(k)})$ being the residue value at iteration k, and c is a constant satisfying $\xi_\tau^\varepsilon(r_{iz}^{(k)} \mid r_{iz}^{(k)}) = \rho_\tau^\varepsilon(r_{iz}^{(k)})$. Generally, ε should be chosen to be close to 0 to ensure that $\rho_\tau^\varepsilon(r_{iz}(\alpha, \boldsymbol{\beta}))$ is a good approximation of $\rho_\tau(r_{iz}(\alpha, \boldsymbol{\beta}))$. Then a standard MM algorithm operates by minimising $\sum_{i=1}^n \xi_\tau^\varepsilon(r_{iz}(\alpha, \boldsymbol{\beta}) \mid r_{iz}^{(k)})$ with respect to $(\alpha, \boldsymbol{\beta})$. Let κ denote the convergence tolerance specified to control relative changes of parameters between the two most recent iterations. A simple simulation is conducted to illustrate the impact of κ and ε on the performance of MM algorithm (see Table A1 in Appendix 2). In the following simulation and case studies, we set $\kappa = 10^{-5}$ and choose ε around κ/n, where n is the sample size.

Based on their work, we can then use the derivative of $\xi_\tau^\varepsilon(r_{iz}(\alpha, \boldsymbol{\beta}) \mid r_{iz}^{(k)})$ as the surrogate function of $\varphi_\tau(r_{iz}(\alpha, \boldsymbol{\beta}))$ involved in $g_{\boldsymbol{z}}(Y_i, \boldsymbol{Z}_i, \alpha, \boldsymbol{\beta})$ and therefore construct a smooth surrogate estimating equations for (6). That is, if $(\alpha^{(k)}, \boldsymbol{\beta}^{(k)})$ is the solution to equations (6)

from the kth iteration, then the surrogate estimating equations at the $k+1$th iteration is

$$\sum_{i=1}^{n}\left\{\frac{\delta_i}{2\hat{\pi}_i}\mathcal{K}_{\mathbf{H}}(\boldsymbol{Z}_i-\boldsymbol{z})\left[\frac{r_{iz}(\alpha,\boldsymbol{\beta})}{\varepsilon+|r_{iz}^{(k)}|}+2\tau-1\right]\nabla_{r_{iz}}(\alpha,\boldsymbol{\beta})\right.$$
$$\left.+\left(1-\frac{\delta_i}{\hat{\pi}_i}\right)\frac{\sum_{j=1}^{n}\overline{\mathcal{K}}_{\bar{h}}(\boldsymbol{X}_j^o-\boldsymbol{X}_i^o)\delta_j\mathcal{K}_{\mathbf{H}}(\boldsymbol{Z}_j-\boldsymbol{z})\left[\frac{r_{jz}(\alpha,\boldsymbol{\beta})}{\varepsilon+|r_{jz}^{(k)}|}+2\tau-1\right]\nabla_{r_{jz}}(\alpha,\boldsymbol{\beta})}{2\sum_{j=1}^{n}\overline{\mathcal{K}}_{\bar{h}}(\boldsymbol{X}_j^o-\boldsymbol{X}_i^o)\delta_j}\right\}=0,$$

where $\nabla_{r_{iz}}(\alpha,\boldsymbol{\beta})$ denotes the derivative of $r_{iz}(\cdot)$ with respect to $(\alpha,\boldsymbol{\beta})$, and $(\alpha^{(k+1)},\boldsymbol{\beta}^{(k+1)})$ is obtained by solving the above equation through Gauss-Newton algorithm.

5. Main results

5.1. Asymptotic properties

Let $\phi(s\mid\boldsymbol{t})\equiv f(\theta_\tau(\boldsymbol{t})+s\mid\boldsymbol{t})$, where $f(\cdot\mid\boldsymbol{t})$ denotes the conditional density function of Y given $\boldsymbol{Z}=\boldsymbol{t}$. Let $\mathcal{H}_{\theta_\tau}(\cdot)$ denote the Hessian matrix of the function $\theta_\tau(\cdot)$. Let $g(\cdot)$ be the density function of those fully observed variables $\boldsymbol{X}^o$.

In this article, we assume the following conditions.

(C1) $\phi(s\mid\boldsymbol{t})$ as a function of s is continuous in a neighbourhood of 0, uniformly for $\boldsymbol{t}$ in a neighbourhood of $\boldsymbol{z}$. On the other hand, $\phi(s\mid\boldsymbol{t})$ as a function of $\boldsymbol{t}$ is bounded and continuous in a neighbourhood of $\boldsymbol{z}$ for all small s and satisfies $\phi(0\mid\boldsymbol{z})\neq 0$.

(C2) The density function of $\boldsymbol{Z}$, i.e. $p(\cdot)$, is continuous and $p(\boldsymbol{z})>0$.

(C3) The function $\theta_\tau(\cdot)$ has a continuous second derivative and $\|\mathcal{H}_{\theta_\tau}(\cdot)\|_2$ is bounded in a neighbourhood of $\boldsymbol{z}$, where the matrix norm $\|\cdot\|_2$ is defined as $\|A\|_2=\lambda_{\max}(A)$, the largest singular value for a given matrix A.

(C4) $g(\cdot)$ and $\pi(\cdot)$ have bounded partial derivatives up to an order l with $l\geq 2$, $l>d_1$, $\inf_{\boldsymbol{x}}\pi(\boldsymbol{x})\geq c_0$ and $\inf_{\boldsymbol{x}} r(\boldsymbol{x})\geq\bar{c}_0$, where $r(\boldsymbol{x})=g(\boldsymbol{x})\pi(\boldsymbol{x})$, and c_0 and $\bar{c}_0$ are positive constants.

(C5) $m_{\boldsymbol{z}}(\boldsymbol{x},\alpha,\boldsymbol{\beta})$ has bounded partial derivatives with respect to $\boldsymbol{x}$ up to an order l and $m_{\boldsymbol{z}}(\boldsymbol{x},\alpha,\boldsymbol{\beta})$ has bounded Hessian matrix with respect to $(\alpha,\boldsymbol{\beta})$ at $(\theta_\tau(\boldsymbol{z}),\nabla_{\theta_\tau}(\boldsymbol{z}))$.

(C6 − 1) $\mathcal{K}(\cdot)$ is a d-dimensional multivariate kernel function with compact support and order 2, satisfying $\int\boldsymbol{t}\boldsymbol{t}^\top\mathcal{K}(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t}=\mu_2(\mathcal{K})\mathrm{I}_d$ with $\mu_2(\mathcal{K})=\int s^2K(s)\,\mathrm{d}s$, and the bandwidth h satisfies $h\to 0$, $nh^d\to\infty$ as $n\to\infty$.

(C6 − 2) $\widetilde{\mathcal{K}}(\cdot)$ is a d_1-dimensional multivariate kernel function with compact support and order l and the bandwidth $\tilde{h}$ satisfies $\tilde{h}\to 0$, $n^\nu\tilde{h}^{d_1}\to\infty$ for some $0<\nu<1/2$ and $n\tilde{h}^{2l}\to 0$ as $n\to\infty$.

(C6 − 3) $\overline{\mathcal{K}}(\cdot)$ is a d_1-dimensional multivariate kernel function with compact support and order l and the bandwidth $\bar{h}$ satisfies $\bar{h}\to 0$, $n^\nu\bar{h}^{d_1}\to\infty$ for some $0<\nu<1/2$ and $n\bar{h}^{2l}\to 0$ as $n\to\infty$.

Remark 5.1: (C1), (C2), (C3) and (C6-1) are commonly used conditions in nonparametric quantile regression literature (Fan et al. 1994). Condition (C4) may not hold for probability density functions with unbounded support. In these cases, we can always truncate the support by setting arbitrarily large intervals and rescale the density to make Condition (C4) be satisfied. Conditions (C6-2) and (C6-3) are twofold. On the one hand, $n^{\nu}\tilde{h}^{d_1} \to \infty$ and $n^{\nu}\bar{h}^{d_1} \to \infty$ for some $0 < \nu < 1/2$ guarantee the uniform convergence of kernel estimators $\hat{\pi}_i$ and $\hat{m}_{\boldsymbol{z}}(\boldsymbol{X}_i^o, \alpha, \boldsymbol{\beta})$, respectively, as stated in Lemma A.1 in Appendix 3 (Mack and Silverman 1982; Stute 1984). On the other hand, $n\tilde{h}^{2l} \to 0$ and $n\bar{h}^{2l} \to 0$ ensure that the bias terms resulted from $\hat{\pi}_i$ and $\hat{m}_{\boldsymbol{z}}(\boldsymbol{X}_i^o, \alpha, \boldsymbol{\beta})$ are negligible. Together with Conditions (C4) and (C5), we can ignore the resultant extra terms when π_i and $m_{\boldsymbol{z}}(\boldsymbol{X}_i^o, \alpha, \boldsymbol{\beta})$ are replaced by $\hat{\pi}_i$ and $\hat{m}_{\boldsymbol{z}}(\boldsymbol{X}_i^o, \alpha, \boldsymbol{\beta})$ in the local estimating equations.

Theorem 5.1: *If Conditions* (C1)–(C6-3) *hold, the solution to Equation* (6) *satisfies*

$$\begin{pmatrix} \hat{\alpha} - \theta_\tau(\boldsymbol{z}) \\ h\left(\hat{\boldsymbol{\beta}} - \nabla_{\theta_\tau}(\boldsymbol{z})\right) \end{pmatrix} \xrightarrow{p} \mathbf{0},$$

where ∇_{θ_τ} *denotes the gradient of* $\theta_\tau(\cdot)$.

Theorem 5.2: *If Conditions (C1)–(C6-3) holds, then*

$$\sqrt{nh^d}\left(\hat{\alpha} - \theta_\tau(\boldsymbol{z}) - \frac{h^2}{2}\mu_2(\mathcal{K})\mathrm{tr}(\mathcal{H}_{\theta_\tau}(\boldsymbol{z})) + o(h^2)\right) \xrightarrow{d} N\left(0, \frac{c(\mathcal{K})}{f^2(\theta_\tau(\boldsymbol{z}) \mid \boldsymbol{z})p(\boldsymbol{z})}\right),$$

where $p(\cdot)$ *denotes the density of the explanatory vector,* $f(\cdot \mid \boldsymbol{z})$ *denotes the density function of* Y *conditional on* $\boldsymbol{Z} = \boldsymbol{z}$, *and*

$$c(\mathcal{K}) = \begin{cases} \tau(1-\tau)\|\mathcal{K}\|_2^2/\pi(\boldsymbol{z}^{(1)}), & \text{if } Y \text{ is missing;} \\ \tau^2\|\mathcal{K}\|_2^2\left[\int \dfrac{f(y \mid \boldsymbol{z})}{\pi(y, \boldsymbol{z}^{(1)})}\,\mathrm{d}y\right] + (1-2\tau)\|\mathcal{K}\|_2^2\left[\displaystyle\int_{\mathcal{D}} \frac{f(y \mid \boldsymbol{z})}{\pi(y, \boldsymbol{z}^{(1)})}\,\mathrm{d}y\right], & \text{otherwise,} \end{cases}$$

with $\|\mathcal{K}\|_2^2 = \int \mathcal{K}^2(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t}$, $\mathcal{D} = \{y : y - \theta_\tau(\boldsymbol{z}) < 0\}$ *and* $\boldsymbol{z}^{(1)}$ *being the corresponding value of the fully observed explanatory subvector* $\boldsymbol{Z}^{(1)}$ ($\boldsymbol{Z}^{(1)} \subseteq \boldsymbol{X}^o$) *when* $\boldsymbol{Z} = \boldsymbol{z}$.

Remark 5.2: Theorem 5.2 shows that the propensity score may affect the asymptotic variance. Specifically, a decrease in the propensity score will lead to increases in both $c(\mathcal{K})$ and variance, which thus worsen the efficiency of the AIPW estimator and explain the necessity of Condition (C4) (i.e. $\pi(\boldsymbol{x})$ is strictly greater than zero for all values of $\boldsymbol{x}$ in the support of $\boldsymbol{X}^o$). When the propensity score equals 1, the asymptotic variance in both cases can be written as $\tau(1-\tau)\|\mathcal{K}\|_2^2/(f^2(\theta_\tau(\boldsymbol{z}) \mid \boldsymbol{z})p(\boldsymbol{z}))$, which is identical to the expression in the existing nonparametric QR literature.

5.2. Bandwidth selection

The choice of bandwidth plays an important role in the performance of nonparametric estimation. As indicated in Section 3, three bandwidth parameters (i.e. $h, \bar{h}$ and $\tilde{h}$) are considered for our proposed method. For $\bar{h}$ and $\tilde{h}$, it is noted that the classical optimal rate

for bandwidth (i.e. $n^{-1/(d_1+2l)}$) is not applicable in this paper since Conditions (C6-2) and (C6-3) require $n\bar{h}^{2l} \to 0, n\tilde{h}^{2l} \to 0$. According to the method proposed in Sepanski, Knickerbocker, and Carroll (1994), an appropriate choice of bandwidth is $Cn^{-1/(d_1+l)}$, where C can be determined by the sample standard deviation of $\boldsymbol{X}^o$. As for h, an asymptotic local optimal bandwidth $h(\boldsymbol{z})$ can be chosen by minimising the corresponding mean squared error induced in Theorem 5.2. In our case, it is given as

$$h(\boldsymbol{z}) = n^{-1/(d+4)} \left\{ \frac{dc(\mathcal{K})}{f^2(\theta_\tau(\boldsymbol{z}) \mid \boldsymbol{z}) p(\boldsymbol{z}) \mu_2^2(\mathcal{K}) \operatorname{tr}^2(\mathcal{H}_{\theta_\tau}(\boldsymbol{z}))} \right\}^{1/(d+4)}.$$

Since $h(\boldsymbol{z})$ depends on some unknown values, the above expression is not applicable in practice. To choose h in practice, we suggest deriving a data-driven bandwidth by minimising empirical mean squared errors $\mathrm{EMSE}\{\boldsymbol{z}; h(\boldsymbol{z})\}$, where $\mathrm{EMSE}\{\boldsymbol{z}; h(\boldsymbol{z})\} = \widehat{\mathrm{bias}}^2\{\theta_\tau(\boldsymbol{z})\} + \widehat{\mathrm{Var}}\{\theta_\tau(\boldsymbol{z})\}$. Here, $\widehat{\mathrm{bias}}\{\theta_\tau(\boldsymbol{z})\}$ and $\widehat{\mathrm{Var}}\{\theta_\tau(\boldsymbol{z})\}$ respectively denote the empirical bias and variance, and they can be computed through smoothed bootstrap resampling method (Cao-Abad and González-Manteiga 1993) as shown below:

(a) Reconstruct the unobservable data set $\{(\delta_i, \boldsymbol{X}_i^o, \hat{\boldsymbol{X}}_i^m)\}_{i=1}^n$ from the observed data $\{(\delta_i, \boldsymbol{X}_i^o, \delta_i \boldsymbol{X}_i^m)\}_{i=1}^n$ by

$$\hat{\boldsymbol{X}}_i^m = \frac{\sum_{j=1}^n \overline{\mathcal{K}}_{\bar{h}}(\boldsymbol{X}_j^o - \boldsymbol{X}_i^o)\delta_j \boldsymbol{X}_j^m}{\sum_{j=1}^n \overline{\mathcal{K}}_{\bar{h}}(\boldsymbol{X}_j^o - \boldsymbol{X}_i^o)\delta_j}, \tag{9}$$

if $\delta_i = 0$ and $\hat{\boldsymbol{X}}_i^m = \boldsymbol{X}_i^m$ otherwise. Rearrange $\{(\delta_i, \boldsymbol{X}_i^o, \hat{\boldsymbol{X}}_i^m)\}_{i=1}^n$ to get $\{(\delta_i, \hat{Y}_i, \hat{\boldsymbol{Z}}_i)\}_{i=1}^n$ in which some components of $(\hat{Y}_i, \hat{\boldsymbol{Z}}_i)$ are constructed using (9) when $\delta_i = 0$.

(b) Draw a bootstrap sample of size n, $\{(\delta_i^*, Y_i^*, \boldsymbol{Z}_i^*)\}_{i=1}^n$, from the following multivariate distribution function

$$F_n(\delta, y, \boldsymbol{z}) = \frac{1}{n} \sum_{i=1}^n I(\delta_i \le \delta) \int_{-\infty}^{y} \frac{1}{l_1} K_1\left(\frac{v - \hat{Y}_i}{l_1}\right) \mathrm{d}v \int_{-\infty}^{\boldsymbol{z}} \frac{1}{l_2} K_2\left(\frac{\boldsymbol{u} - \hat{\boldsymbol{Z}}_i}{l_2}\right) \mathrm{d}\boldsymbol{u},$$

where $K_1(\cdot)$ and $K_2(\cdot)$ are respectively univariate and d-dimensional kernel functions with corresponding bandwidth l_1 and l_2. Delete the components which correspond to $\boldsymbol{X}_i^{m*}$ in $(Y_i^*, \boldsymbol{Z}_i^*)$ if $\delta_i^* = 0$.

(c) Calculate $\hat{\pi}_i^*$ based on the bootstrap sample and obtain $\hat{\alpha}^*$ by solving the following estimating equations:

$$\sum_{i=1}^n \left[\frac{\delta_i^*}{\hat{\pi}_i^*} g_{\boldsymbol{z}}(Y_i^*, \boldsymbol{Z}_i^*, \alpha, \boldsymbol{\beta}) + \left(1 - \frac{\delta_i^*}{\hat{\pi}_i^*}\right) \hat{m}_{\boldsymbol{z}}(\boldsymbol{X}_i^{o*}, \alpha, \boldsymbol{\beta}) \right] = \boldsymbol{0}.$$

(d) Repeat (b) and (c) for N times to get the collection of bootstrap estimates $\{\hat{\alpha}^{*(j)}\}_{j=1}^N$.

(e) Denote $\theta_\tau(\boldsymbol{z}; F_n)$ as the τth quantile of conditional distribution $F_n(y \mid \boldsymbol{Z} = z)$. According to the definition, $\theta_\tau(\boldsymbol{z}; F_n)$ can be obtained by solving the following equation

$$\tau = \frac{\sum_{i=1}^n K_{l_2}(\boldsymbol{z} - \hat{\boldsymbol{Z}}_i) \int_{-\infty}^{\theta_\tau(\boldsymbol{z};F_n)} K_{l_1}(y - \hat{Y}_i)\, \mathrm{d}y}{\sum_{i=1}^n K_{l_2}(\boldsymbol{z} - \hat{\boldsymbol{Z}}_i)}.$$

Based on $\{\hat{\alpha}^{*(1)}, \ldots, \hat{\alpha}^{*(N)}\}$ and $\theta_\tau(\boldsymbol{z}; F_n)$, we calculate the bootstrap bias $\bar{\alpha}^* - \theta_\tau(\boldsymbol{z}; F_n)$ and the bootstrap variance $\sum_{j=1}^N (\hat{\alpha}^{*(j)} - \bar{\alpha}^*)^2/N$, where $\bar{\alpha}^* = \sum_{j=1}^N \hat{\alpha}^{*(j)}/N$.

Remark 5.3: If we denote the distribution from which $\{(Y_i, \boldsymbol{Z}_i)\}_{i=1}^n$ are generated by $F(y, \boldsymbol{z})$, then $F_n(y, \boldsymbol{z})(F_n(1, y, \boldsymbol{z})$ defined in step (b)) is a smoothed approximation of F. In step (b), a sample $(\delta^*, Y^*, \boldsymbol{Z}^*)$ from F_n can be generated in a simple way: If $(\delta, \hat{Y}, \hat{\boldsymbol{Z}})$ is a sample from $\hat{F}_n$, the empirical distribution of $\{(\delta_i, \hat{Y}_i, \hat{\boldsymbol{Z}}_i)\}_{i=1}^n$, and U and V are independently generated from distributions determined by kernel function K_1 and K_2 respectively, then we may take $\delta^* = \delta$, $Y^* = \hat{Y} + l_1 U$ and $\boldsymbol{Z}^* = \hat{\boldsymbol{Z}} + l_2 V$. Additionally, bandwidths l_1 and l_2 can be chosen via function 'npcdistbw' in R package 'np'.

6. Simulations

In this section, we conduct simulation studies to evaluate the finite-sample performance of our proposed method and compare it with some existing methods.

Case 1. We consider data to be generated from the following model:

$$Y_i = 5\sin(2\pi Z_{1i}) - 2Z_{2i} + \epsilon_i,$$

where Z_{1i} and Z_{2i} are generated from uniform$(0, 1)$ independently, and ϵ_i is generated according to $\epsilon_i + \Phi^{-1}(\tau) \sim N(0, 1)$ and is independent of $\boldsymbol{Z}_i$. Therefore, the τth conditional quantile of Y at a given point is $\theta_\tau(\boldsymbol{z}) = 5\sin(2\pi z_1) - 2z_2$. The propensity score function considered here is $\pi(z_2, y) = 0.4$ if $y \le 0.25$th sample quantile of Y, and $\pi(z_2, y) = 0.7$ otherwise. In this case, the missing proportion is roughly 37%. It can be seen that the fully observed variables are Y and Z_2 and that the missing observations only come from covariate Z_1. The kernel function used to estimate $\pi(z_2, y)$ in (7) and $m_z(z_2, y; \alpha, \boldsymbol{\beta})$ in (8) is chosen to be the 4th order Gaussian kernel, which is consistent with the assumption in Section 5.1 (i.e. $\widetilde{K}(x) = \overline{K}(x) = (3/2 - x^2/2)\exp(-x^2/2)/\sqrt{2\pi}$). For $K(\cdot)$, which is used to construct the local estimating functions in (5), we choose the ordinary Epanechnikov kernel.

Figure 1 shows the true surface $\theta_\tau(\boldsymbol{z})$ and the estimated surfaces $\hat{\theta}_\tau(\boldsymbol{z})$ (based on one replication) obtained by solving AIPW local estimating equations with $\tau = 0.25, 0.5$ and 0.75. It can be seen that the potential surface can be restored in different quantiles and the performance of our proposed estimator for $\tau = 0.5$ is more favourable than those for $\tau = 0.25$ and 0.75 at the edge points. This is most likely due to the fact that larger amount of data are available around the center of the distribution than those available around the tails.

Next, we evaluate the performance of various methods with respect to the magnitude of average absolute relative bias (AARB) and average square error (ASE) based on 100 Monte

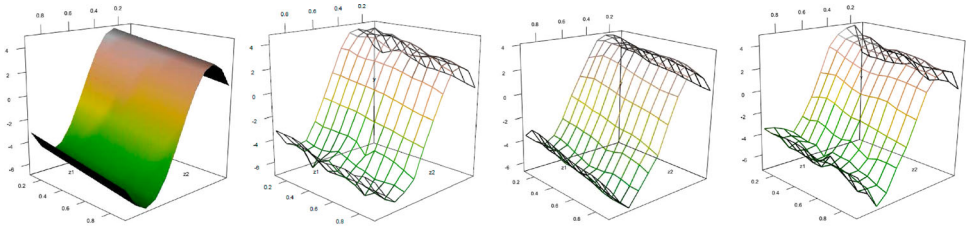

Figure 1. The true surface and the estimated surfaces for $\tau = 0.25, 0.5, 0.75$ (from left to right).

Carlo replications; i.e.

$$\text{AARB} = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{\hat{\theta}_\tau(\boldsymbol{z}_i) - \theta_\tau(\boldsymbol{z}_i)}{\theta_\tau(\boldsymbol{z}_i)}\right|, \quad \text{ASE} = \frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_\tau(\boldsymbol{z}_i) - \theta_\tau(\boldsymbol{z}_i))^2,$$

where $\{\boldsymbol{z}_i, i = 1, \ldots, m\}$ are grid points on $[0,1]^2$. Besides, bootstrap ASE's (denoted by ASE^{bs}) are also calculated. In this example, the following five methods are considered:

- Full: estimating based on the original full data set (i.e. the original data set before missing data are generated).
- CC: estimating based on those complete cases (i.e. those observations with $\{i : \delta_i = 1\}$).
- AIPW: estimating by solving equations (6) with propensity score and augmented term estimated by (7) and (8), respectively.
- IPW: estimating by solving the following IPW kernel equations:

$$\sum_{i=1}^{n}\left[\frac{\delta_i}{\hat{\pi}_i}g_{\boldsymbol{z}}(Y_i, \boldsymbol{Z}_i, \alpha, \boldsymbol{\beta})\right] = \mathbf{0},$$

  where $\hat{\pi}_i$ is defined in (7).
- EE: estimating by solving the following estimating equations (Zhou et al. 2008):

$$\sum_{i=1}^{n}\left[\delta_i g_{\boldsymbol{z}}(Y_i, \boldsymbol{Z}_i, \alpha, \boldsymbol{\beta}) + (1-\delta_i)\,\hat{m}_{\boldsymbol{z}}(\boldsymbol{X}_i^o, \alpha, \boldsymbol{\beta})\right] = \mathbf{0},$$

  where $\hat{m}_{\boldsymbol{z}}(\boldsymbol{X}_i^o, \alpha, \boldsymbol{\beta})$ is defined in (8).

Table 1 shows the performance of CC, AIPW, IPW and EE methods for the model given in Case 1 with Full method serving as the benchmark. Numbers in parentheses are standard deviations of corresponding quantities. In general, CC method is inferior to all other methods due to the loss of information in the estimation procedure. Our proposed AIPW method generally performs the best in the sense that it yields the smallest AARBs, ASEs and ASE^{bs}s in most cases. Additionally, we observe no substantial difference between ASE^{bs} and its corresponding ASE, which indicates that using the smoothed bootstrap procedure to calculate ASE is an appropriate method. Therefore, bandwidth selection based on the empirical ASE is reasonable in practice.

Table 1. The performance of CC, AIPW, IPW and EE methods in different quantiles for the model given in Case 1 with 'Full' serving as the benchmark.

Case 1, $n = 500$				
τ	Method	AARB	ASE	ASE^{bs}
$\tau = 0.5$	Full	0.0923 (0.0201)	0.0477 (0.0089)	0.0481 (0.0134)
	CC	0.1209 (0.0206)	0.0895 (0.0161)	0.1149 (0.0195)
	AIPW	0.1064 (0.0196)	0.0779 (0.0133)	0.0842 (0.0150)
	IPW	0.1094 (0.0197)	0.0802 (0.0148)	0.0910 (0.0153)
	EE	0.1103 (0.0199)	0.0846 (0.0154)	0.0922 (0.0170)
$\tau = 0.25$	Full	0.1189 (0.0242)	0.0693 (0.0082)	0.0731 (0.0126)
	CC	0.1517 (0.0284)	0.1068 (0.0163)	0.1255 (0.0155)
	AIPW	0.1307 (0.0273)	0.0925 (0.0150)	0.1049 (0.0144)
	IPW	0.1388 (0.0270)	0.0981 (0.0155)	0.1132 (0.0137)
	EE	0.1315 (0.0281)	0.1014 (0.0157)	0.1120 (0.0192)
$\tau = 0.75$	Full	0.1021 (0.0211)	0.0767 (0.0091)	0.0740 (0.0176)
	CC	0.1332 (0.0296)	0.1210 (0.0151)	0.1280 (0.0202)
	AIPW	0.1259 (0.0225)	0.1058 (0.0123)	0.1054 (0.0165)
	IPW	0.1249 (0.0198)	0.1073 (0.0138)	0.1090 (0.0171)
	EE	0.1264 (0.0239)	0.1087 (0.0141)	0.1105 (0.0173)

Note: The comparison is conducted in terms of average absolute relative bias (AARB), average square error (ASE) and bootstrap ASE (ASE^{bs}). Numbers in parentheses are standard deviations of corresponding quantities.

Case 2. We consider data to be generated from the following model:

$$Y_i = \frac{4\Gamma(8)\Gamma(8)}{\Gamma(8+8)} Z_i^{8-1}(1-Z_i)^{8-1} - 4 + \epsilon_i,$$

where $Z_i \sim U(0,1)$ and $\epsilon_i + \Phi^{-1}(\tau) \sim N(0,1)$ with ϵ_i independent of Z_i. The τth conditional quantile of Y_i at a given point is thus $\theta_\tau(z) = 4\Gamma(8)\Gamma(8)/\Gamma(8+8)z^{8-1}(1-z)^{8-1} - 4$. Here, we set the propensity score to be $\pi(z) = \exp(8z^2 - 4z - 1)/(1 + \exp(8z^2 - 4z - 1))$ and the missing data only come from the response variable Y. Roughly, 57% of the observations are missing. Besides the aforementioned five methods, another three methods are also considered:

- AIPW^1: estimating by solving equations (6) with misspecification of the propensity score. Here, the propensity score $\pi(z)$ is estimated by $\pi(z, \hat{\boldsymbol{\eta}})$, where $\pi(z, \boldsymbol{\eta}) = \{1 + \exp(-\eta_1 - z\eta_2)\}^{-1}$ and $\hat{\boldsymbol{\eta}}$ is the maximum likelihood estimate (MLE) of $\boldsymbol{\eta}$ based on $\{(\delta_i, Z_i)\}_{i=1}^n$; i.e. $\hat{\boldsymbol{\eta}}$ is obtained by maximising $\prod_{i=1}^n \pi(Z_i, \boldsymbol{\eta})^{\delta_i}(1 - \pi(Z_i, \boldsymbol{\eta}))^{1-\delta_i}$.
- AIPW^2: estimating by solving equations (6) with misspecification of the augmented term. Here, we set the conditional distribution $p_{Y|Z}(y \mid z, \boldsymbol{\zeta})$ as $N(\zeta_1 + Z\zeta_2, 1)$ and estimate $m_z(Z_i, \alpha, \boldsymbol{\beta})$ by $\int g_z(y, Z_i, \alpha, \boldsymbol{\beta}) p(y \mid Z_i, \hat{\boldsymbol{\zeta}})\, \mathrm{d}y$, where $\hat{\boldsymbol{\zeta}}$ is the MLE of $\boldsymbol{\zeta}$ based on the complete cases.
- IPW^1: estimating by solving IPW kernel equations with misspecification of the propensity score.

Table 2 reports the performance of different methods in Case 2. It can be seen that both AIPW^1 and AIPW^2 methods perform quite well in terms of AARB and ASE, which provides further verification on double robustness of our proposed AIPW method. Additionally, the CC method does well in terms of bias in this example, which is consistent

Table 2. The performance of CC, AIPW, AIPW^1, AIPW^2, IPW and IPW^1 and EE methods in different quantiles for the model given in Case 2 with 'Full' serving as the benchmark.

Example 2, $n = 500$				
τ	Method	AARB	ASE	ASE^{bs}
$\tau = 0.5$	Full	0.0215 (0.0120)	0.0117 (0.0050)	0.0142 (0.0079)
	CC	0.0634 (0.0246)	0.0772 (0.0419)	0.0780 (0.0395)
	AIPW	0.0519 (0.0249)	0.0335 (0.0315)	0.0350 (0.0302)
	AIPW^1	0.0526 (0.0248)	0.0347 (0.0314)	0.0366 (0.0331)
	AIPW^2	0.0537 (0.0249)	0.0364 (0.0315)	0.0375 (0.0406)
	IPW	0.0520 (0.0246)	0.0346 (0.0316)	0.0361 (0.0353)
	IPW^1	0.0638 (0.0247)	0.0727 (0.0321)	0.0743 (0.0382)
	EE	0.0529 (0.0250)	0.0353 (0.0314)	0.0370 (0.0343)
$\tau = 0.25$	Full	0.0636 (0.0302)	0.0398 (0.0285)	0.0407 (0.0315)
	CC	0.1150 (0.0366)	0.0907 (0.0549)	0.1018 (0.0550)
	AIPW	0.1117 (0.0318)	0.0609 (0.0550)	0.0625 (0.0505)
	AIPW^1	0.1121 (0.0309)	0.0673 (0.0561)	0.0670 (0.0514)
	AIPW^2	0.1120 (0.0327)	0.0677 (0.0595)	0.0685 (0.0517)
	IPW	0.1136 (0.0370)	0.0647 (0.0510)	0.0679 (0.0529)
	IPW^1	0.1147 (0.0376)	0.1085 (0.0522)	0.1080 (0.0615)
	EE	0.1119 (0.0381)	0.0672 (0.0527)	0.0681 (0.0536)
$\tau = 0.75$	Full	0.0705 (0.0316)	0.0429 (0.0301)	0.0447 (0.0360)
	CC	0.1142 (0.0387)	0.0839 (0.0597)	0.0890 (0.0598)
	AIPW	0.1123 (0.0356)	0.0627 (0.0439)	0.0643 (0.0552)
	AIPW^1	0.1125 (0.0347)	0.0629 (0.0503)	0.0657 (0.0529)
	AIPW^2	0.1159 (0.0352)	0.0639 (0.0521)	0.0660 (0.0494)
	IPW	0.1134 (0.0358)	0.0631 (0.0504)	0.0648 (0.0501)
	IPW^1	0.1130 (0.0339)	0.0735 (0.0532)	0.0752 (0.0590)
	EE	0.1145 (0.0355)	0.0622 (0.0477)	0.0632 (0.0558)

Note: The comparison is conducted in terms of average absolute relative bias (AARB), average square error (ASE) and bootstrap ASE (ASE^{bs}). Numbers in parentheses are standard deviations of corresponding quantities.

with the arguments provided in Wang et al. (1997) that the CC method gives valid inference when the missingness depends only on the regressors. Again, our proposed AIPW method generally performs the best.

7. Real data analysis

In this section, we apply our proposed method to analyse a non-insulin-dependent diabetes mellitus data set, in which it contains information of a population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona (Smith, Everhart, Dickson, Knowler, and Johannes 1988). All subjects were tested for diabetes according to World Health Organization criteria. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases. Since our main interest here is to investigate the effects of obesity and genetics on the plasma glucose level, we restrict our analysis to the following attributes: $Z_1 = \log(\text{skin})$, denoting the logarithm of triceps skin fold thickness; $Z_2 = \text{ped}$, referring to the diabetes pedigree function and providing a measure of the expected genetic influence of affected and unaffected relatives on the subject's eventual diabetes risk; and the dependent variable $Y = \text{glu}$, representing the plasma glucose concentration at 2 hours in an oral glucose tolerance test. The present data set comprises 632 observations, amongst 98 subjects have their triceps skin fold thickness records unavailable

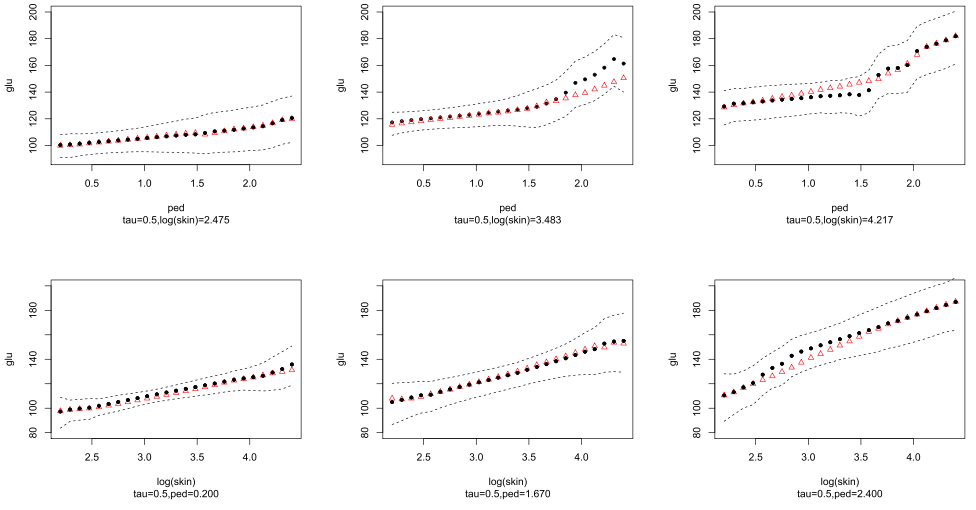

Figure 2. The effect of genetics (obesity) on 0.5-quantile of plasma glucose level when obesity (genetics) is fixed. The solid black circles are obtained by AIPW method while the hallow triangles are obtained by CC method. The area between dashed lines in each subfigure illustrates the corresponding 95% pointwise confidence intervals calculated by AIPW smoothed bootstrap procedure.

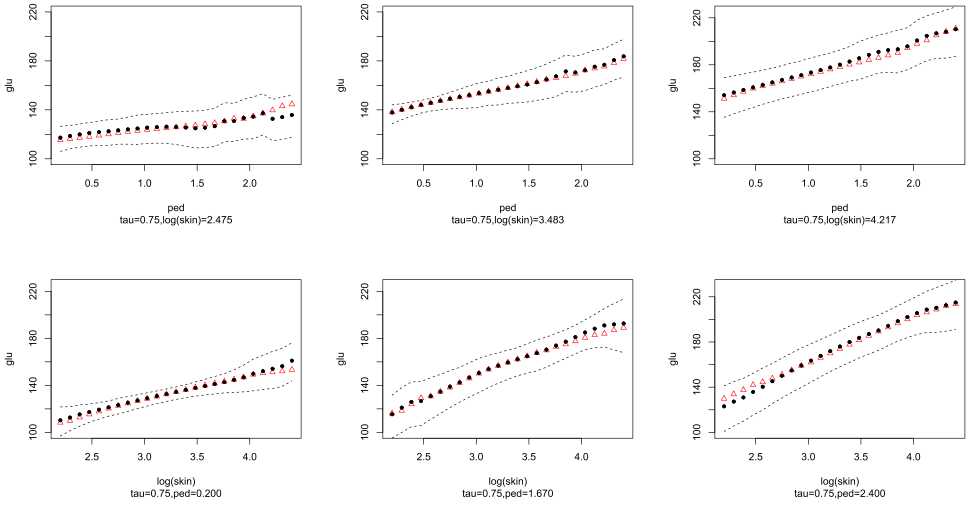

Figure 3. The effect of genetics (obesity) on 0.75-quantile of plasma glucose level when obesity (genetics) is fixed. The solid black circles are obtained by AIPW method while the hallow triangles are obtained by CC method. The area between dashed lines in each subfigure illustrates the corresponding 95% pointwise confidence intervals calculated by AIPW smoothed bootstrap procedure.

(i.e. $\boldsymbol{X}^o = (Y, Z_2)$ and $\boldsymbol{X}^m = Z_1$). Here, we consider the nonparametric QR models based on the AIPW local estimating equations at $\tau = 0.5$ and 0.75, respectively, to study the effects of obesity and genetics on median and upper quantile of plasma glucose level.

In order to have a better understanding about the effect of individual covariate on the dependent variable, we plot the quantile regressions between the dependent variable and

the covariate of interest with the other covariate being fixed (at certain value). The corresponding plots are reported in Figures 2 and 3. At $\tau = 0.5$ (i.e. median), it is interesting to notice that there is a linear association between glu and ped with log(skin) being fixed at 2.475, while there is a piece-wise linear association between glu and ped with log(skin) being fixed at 3.483 and 4.217. For the latter case, a turning point around ped = 1.5 is observed (see first row of Figure 2) and the two estimation methods yield substantially different results. This supports that the nonparametric regression model is an appropriate choice and ignorance of missing data could yield biased results. Unlike the results being observed at $\tau = 0.5$, it is noteworthy that there are always strong linear relationships between glu and ped regardless of the value of log(skin) at $\tau = 0.75$ (see first row of Figure 3). Additionally, the slope (between glu and ped) increases with the fixed value of log(skin). This supports that quantile regression model is an appropriate choice. Finally, second rows of Figures 2 and 3 show a significant linear trend between glu and log(skin) regardless of the τ-value. Besides, the slopes (between glu and log(skin)) increases with the fixed value of ped.

8. Discussion

In this paper, AIPW local estimating equations have been constructed to deal with nonparametric quantile regression models for data missing at random. To circumvent misspecification, we use kernel smoother to estimate both the propensity score and the augmented term. The theoretical properties of our proposed estimator are derived and the corresponding computation algorithm is presented. Besides, our method exhibits satisfactory finite sample performance.

The nonparametric regression model considered in this paper tends to fall into the curse of dimensionality since the convergence rate decrease rapidly as the dimension of $\boldsymbol{Z}$ increase, which is indicated by Theorem 5.2. The extension of our method to semiparametric models such as additive models and partial linear models can be considered to circumvent the dimensionality issue. Another limitation is that our research is based on the MAR assumption. However, it is difficult or even impossible to identify the exact missing mechanism of real data unless the missingness is by design. When the MAR assumption does not hold, our methods may not be appropriate. Developing proper methods to handle above issues will be challenging and deserves further research.

Acknowledgments

The authors would like to thank two anonymous reviewers for providing insightful comments and suggestions which improved the quality of this paper, especially on the smoothed bootstrap part.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was partially supported by the National Natural Science Foundation of China [grant number 11861042], and the China Statistical Research Project [grant number 2020LZ25]. The work of Man-Lai Tang was partially supported through grants from the Research Grant Council of the Hong Kong Special Administrative Region [grant number UGC/FDS14/P01/16, UGC/FDS14/P02/18 and

The Research Matching Grant Scheme (RMGS)] and a grant from the National Natural Science Foundation of China [grant number 11871124]. The computing facilities/software were supported from SAS Viya and the Big Data Intelligence Centre at The Hang Seng University of Hong Kong.

ORCID

Maozai Tian http://orcid.org/0000-0002-0515-4477

References

Cao-Abad, R., and González-Manteiga, W. (1993), 'Bootstrap Methods in Regression Smoothing', *Journal of Nonparametric Statistics*, 2(4), 379–388.

Carroll, R.J., Ruppert, D., and Welsh, A.H. (1998), 'Local Estimating Equations', *Journal of the American Statistical Association*, 93, 214–227.

Chen, X., Wan, A.T., and Zhou, Y. (2015), 'Efficient Quantile Regression Analysis With Missing Observations', *Journal of the American Statistical Association*, 110, 723–741.

Fan, J., Hu, T.C., and Truong, Y.K. (1994), 'Robust Non-parametric Function Estimation', *Scandinavian Journal of Statistics*, 21(4), 433–446.

Han, P. (2016), 'Combining Inverse Probability Weighting and Multiple Imputation to Improve Robustness of Estimation', *Scandinavian Journal of Statistics*, 43(1), 246–260.

Hu, Y., Yang, Y., Wang, C., and Tian, M (2017), 'Imputation in Nonparametric Quantile Regression With Complex Data', *Statistics & Probability Letters*, 127, 120–130.

Hunter, D.R., and Lange, K. (2000), 'Quantile Regression Via An MM Algorithm', *Journal of Computational and Graphical Statistics*, 9, 60–77.

Knowler, W.C., Pettitt, D.J., Savage, P.J., and Bennett, P.H. (1981), 'Diabetes Incidence in Pima Indians: Contributions of Obesity and Parental Diabetes', *American Journal of Epidemiology*, 113(2), 144–156.

Little, R.J.A., and Rubin, D.B. (2002), *Statistical Analysis With Missing Data* (2nd ed.), New York: Wiley.

Mack, Y.P., and Silverman, B.W. (1982), 'Weak and Strong Uniform Consistency of Kernel Regression Estimates', *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 61(3), 405–415.

Robins, J.M., Rotnitzky, A., and Zhao, L.P (1994), 'Estimation of Regression Coefficients When Some Regressors Are Not Always Observed', *Journal of the American Statistical Association*, 89(427), 846–866.

Ruppert, D., and Wand, M.P. (1994), 'Multivariate Locally Weighted Least Squares Regression', *The Annals of Statistics*, 22(3), 1346–1370.

Sepanski, J.H., Knickerbocker, R., and Carroll, R.J. (1994), 'A Semiparametric Correction for Attenuation', *Journal of the American Statistical Association*, 89(428), 1366–1373.

Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., and Johannes, R.S. (1988, November 7–11), 'Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus', in *Proceedings of the Annual Symposium on Computer Application in Medical Care*, Orlando, pp. 261–265. Washington, DC.

Stute, W. (1984), 'The Oscillation Behavior of Empirical Processes: The Multivariate Case', *The Annals of Probability*, 12, 361–379.

Tsiatis, A. (2007), *Semiparametric Theory and Missing Data*, New York, NY: Springer Science & Business Media.

Van der Vaart, A.W. (1998), *Asymptotic Statistics* (Vol. 3), Cambridge: Cambridge University Press.

Wang, L., Rotnitzky, A., and Lin, X. (2010), 'Nonparametric Regression With Missing Outcomes Using Weighted Kernel Estimating Equations', *Journal of the American Statistical Association*, 105, 1135–1146.

Wang, C.Y., Wang, S., Gutierrez, R.G., and Carroll, R.J. (1998), 'Local Linear Regression for Generalized Linear Models With Missing Data', *The Annals of Statistics*, 26(3), 1028–1050.

Wang, C.Y., Wang, S., Zhao, L.P., and Ou, S.T. (1997), 'Weighted Semiparametric Estimation in Regression Analysis With Missing Covariate Data', *Journal of the American Statistical Association*, 92, 512–525.

Zhou, Y., Wan, A.T.K., and Wang, X. (2008), 'Estimating Equations Inference With Missing Data', *Journal of the American Statistical Association*, 103, 1187–1199.

Appendices

Appendix 1

There are five observations ($n = 5$) denoted as $(1, \ldots, 5)$ with Y being the response variable and $\boldsymbol{Z} = (Z^{(1)}, Z^{(2)}, Z^{(3)})$ the three-dimensional vector of covariates. The completely observed variables (i.e. which can be observed for all five observations) are $Z^{(1)}$ and $Z^{(3)}$. That is, $\boldsymbol{X}^o = (Z^{(1)}, Z^{(3)})$ while the missing variables (i.e. missing in some observations) are Y and $Z^{(2)}$, and $\boldsymbol{X}^c = (Y, Z^{(2)})$. Although only Y is missing and $Z^{(2)}$ is observed for the second observation, we can still define $\boldsymbol{X}_2^c$ as $\boldsymbol{X}_2^c = (Y, Z^{(2)})$. Therefore, the notations $\boldsymbol{X}^o$ and $\boldsymbol{X}^c$ are well-defined according to the whole data set. In this example, we have $\boldsymbol{X}_i^o = (Z_i^{(1)}, Z_i^{(3)})$ and $\boldsymbol{X}_i^c = (Y_i, Z_i^{(2)})$ for all the i' s.

Appendix 2

We conduct a simple simulation to illustrate the performance of MM algorithm (see Section 4) under different settings of (κ, ε). Data are generated from the following standard parametric quantile regression model:

$$Y_i = Z_{i1}\beta_1 + Z_{i2}\beta_2 + \epsilon_i, \quad i = 1, \ldots, n,$$

where Z_{i1} and Z_{i2} are independently generated from the uniform distribution $U(0, 1)$, the vector of regression coefficients is set as $(\beta_1, \beta_2) = (1, 2)$, and ϵ_i is generated according to $\epsilon_i + 0.1\Phi^{-1}(\tau) \sim N(0, 0.1^2)$ and is independent with (Z_{i1}, Z_{i2}). The sample size here is set to be $n = 200$. To reduce computation burden and focus on the impact of (ε, κ), we restrict our attention to this parametric model with no missing data. The linear regression form makes the MM algorithm perform efficiently in this case since $\mathrm{d}\xi_\tau^\varepsilon(r_i(\boldsymbol{\beta}) \mid r_i^{(k)})/\mathrm{d}\boldsymbol{\beta} = 0$ can be solved analytically. As the least-squares starting value is quite close to the minimum point when $\tau = 0.5$, we only consider the situation for $\tau = 0.1$ and $\tau = 0.9$, as shown in Table A1. Results in Table A1 are based on 100 repetitions; that is, K is the average iteration number across 100 repetitions and $\hat{\theta}_1$ and $\hat{\theta}_2$ are average estimated values of θ_1 and θ_2, respectively.

It can be seen from Table A1 that the smaller the convergence tolerance κ, the more iterations required. Also, smaller κ usually means better accuracy. Compared with the case that $\kappa = 10^{-3}$, estimation results become better when $\kappa = 10^{-5}$ or 10^{-7}. Hence, the value of κ represents a compromise between accuracy and computation cost. For permutation constant ε, estimates show larger biases when $\varepsilon = 10^{-1}$ under different values of κ. Estimation results are quite similar when $\varepsilon = 10^{-3}, 10^{-5}$ and 10^{-7} in this simple simulation study.

Appendix 3

A.1 Notations

Let $a = \alpha, \boldsymbol{b} = h\boldsymbol{\beta}, a^* = \alpha^*, \boldsymbol{b}^* = h\boldsymbol{\beta}^*, \boldsymbol{r} = (a, \boldsymbol{b}^\top)^\top$ and $\boldsymbol{r}^* = (a^*, \boldsymbol{b}^{*\top})^\top = (\theta_\tau(\boldsymbol{z}), h\nabla_{\theta_\tau}^\top(\boldsymbol{z}))^\top$. The local estimating equations of $\boldsymbol{r}$ can then be written as

$$g_{\boldsymbol{z}}^\sharp(Y_i, \boldsymbol{Z}_i, \boldsymbol{r}) = \mathcal{K}_{\mathbf{H}}(\boldsymbol{Z}_i - \boldsymbol{z})\varphi_\tau\left(Y_i - a - \left(\frac{\boldsymbol{Z}_i - \boldsymbol{z}}{h}\right)^\top \boldsymbol{b}\right) U_{\mathbf{H}}(\boldsymbol{Z}_i - \boldsymbol{z}),$$

where $U_{\mathbf{H}} = (1, (\boldsymbol{Z}_i - \boldsymbol{z})^\top/h)^\top$. Let $m_{\boldsymbol{z}}^\sharp(\boldsymbol{X}_i^o, \boldsymbol{r})$ denote the projection of $g_{\boldsymbol{z}}^\sharp(Y_i, \boldsymbol{Z}_i, \boldsymbol{r})$ into the space generated by $\boldsymbol{X}_i^o$; i.e. $m_{\boldsymbol{z}}^\sharp(\boldsymbol{X}_i^o, \boldsymbol{r}) = E[g_{\boldsymbol{z}}^\sharp(Y_i, \boldsymbol{Z}_i, \boldsymbol{r}) \mid \boldsymbol{X}_i^o]$. Define

$$G_n^{AIPW}(\hat{\pi}, \hat{m}_{\boldsymbol{z}}^\sharp, \boldsymbol{r}) = \frac{1}{n}\sum_{i=1}^n \left[\frac{\delta_i}{\hat{\pi}_i} g_{\boldsymbol{z}}^\sharp(Y_i, \boldsymbol{Z}_i, \boldsymbol{r}) + \left(1 - \frac{\delta_i}{\hat{\pi}_i}\right)\hat{m}_{\boldsymbol{z}}^\sharp(\boldsymbol{X}_i^o, \boldsymbol{r})\right],$$

$$G_n^{AIPW}(\pi, m_z^\sharp, \boldsymbol{r}) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\delta_i}{\pi_i} g_z^\sharp(Y_i, \boldsymbol{Z}_i, \boldsymbol{r}) + \left(1 - \frac{\delta_i}{\pi_i}\right) m_z^\sharp(\boldsymbol{X}_i^o, \boldsymbol{r})\right],$$

and $G^{AIPW}(\pi, m_z^\sharp, \boldsymbol{r}) = E[G_n^{AIPW}(\pi, m_z^\sharp, \boldsymbol{r})]$.

A.2 Proofs

Lemma A.1: *Assume that* (C1), (C2), (C6-1) *are satisfied, $\eta(\cdot)$ is a real function, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are d-dimensional iid random vectors and $\tilde{\boldsymbol{X}}_1, \ldots, \tilde{\boldsymbol{X}}_n$ are independent random vectors either identically or nonidentically distributed. Assume further that $E|\eta(\boldsymbol{X}_i, \tilde{\boldsymbol{X}}_i)|^s < \infty$ and $\sup_{\boldsymbol{x}} \int \eta(\boldsymbol{x}, \tilde{\boldsymbol{x}}) f_{\boldsymbol{X}, \tilde{\boldsymbol{X}}_i}(\boldsymbol{x}, \tilde{\boldsymbol{x}})\, \mathrm{d}\tilde{\boldsymbol{x}} < \infty$, with $f_{\boldsymbol{X}, \tilde{\boldsymbol{X}}_i}(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ denoting the joint density of $(\boldsymbol{X}_i, \tilde{\boldsymbol{X}}_i)$. If $n \to \infty$, $h \to 0$ and $n^\nu h^d \to \infty$ for some $0 < \nu < 1 - s^{-1}$, then we have*

$$\sup_{\boldsymbol{x}} \left| \frac{1}{n}\sum_{i=1}^{n} \left\{ \mathcal{K}_{\mathbf{H}}(\boldsymbol{X}_i - \boldsymbol{x}) \prod_{j=1}^{d} \left(\frac{X_{ij} - x_j}{h}\right)^k \eta(\boldsymbol{X}_i, \tilde{\boldsymbol{X}}_i) \right.\right.$$
$$\left.\left. - E\left[\mathcal{K}_{\mathbf{H}}(\boldsymbol{X}_i - \boldsymbol{x}) \prod_{j=1}^{d} \left(\frac{X_{ij} - x_j}{h}\right)^k \eta(\boldsymbol{X}_i, \tilde{\boldsymbol{X}}_i) \right] \right\} \right| = o_p(1)$$

Additionally, if $E[\eta(\boldsymbol{X}_i, \tilde{\boldsymbol{X}}_i) \mid \boldsymbol{X}_i = \boldsymbol{x}], i = 1, \ldots, n$, is continuous and twice differentiable at $\boldsymbol{X}_i = \boldsymbol{x}$, then

$$E\left[\mathcal{K}_{\mathbf{H}}(\boldsymbol{X}_i - \boldsymbol{x}) \prod_{j=1}^{d} \left(\frac{X_{ij} - x_j}{h}\right)^k \eta(\boldsymbol{X}_i, \tilde{\boldsymbol{X}}_i) \right]$$
$$= \mu_k f(\boldsymbol{x}) E(\eta(\boldsymbol{X}_i, \tilde{\boldsymbol{X}}_i) \mid \boldsymbol{X}_i = \boldsymbol{x}) + h\mu_{k+1} \frac{\partial (f(\boldsymbol{x}) E(\eta(\boldsymbol{X}_i, \tilde{\boldsymbol{X}}_i) \mid \boldsymbol{X}_i = \boldsymbol{x}))}{\partial \boldsymbol{x}} + O(h^2),$$

where μ_k denotes the multiple integral $\int t_1^k \cdots t_d^k \mathcal{K}(\boldsymbol{t})\, \mathrm{d}\boldsymbol{t}$. Specially, when $k = 0$ and $\mathcal{K}(\cdot)$ is a kernel function of order l, we have

$$E\left[\mathcal{K}_{\mathbf{H}}(\boldsymbol{X}_i - \boldsymbol{x})\, \eta(\boldsymbol{X}_i, \tilde{\boldsymbol{X}}_i)\right] = f(\boldsymbol{x}) E(\eta(\boldsymbol{X}_i, \tilde{\boldsymbol{X}}_i) \mid \boldsymbol{X}_i = \boldsymbol{x}) + O(h^l).$$

This lemma is a natural generalisation of Lemma 1 in Chen et al. (2015) to multiple cases. Similar results with detailed proof can be found in Stute (1984) and Mack and Silverman (1982).

Lemma A.2: *Assume conditions* (C1)–(C6-3) *are satisfied. We have*

$$\sqrt{nh^d} G_n^{AIPW}(\hat{\pi}, \hat{m}_z^\sharp, \boldsymbol{r}^*) = \frac{\sqrt{h^d}}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{\delta_i}{\pi_i} g_z^\sharp(Y_i, \boldsymbol{Z}_i, \boldsymbol{r}^*) + \left(1 - \frac{\delta_i}{\pi_i}\right) m_z^\sharp(\boldsymbol{X}_i, \boldsymbol{r}^*) \right\} + o_p(\mathbf{1}).$$

Proof: $G_n^{AIPW}(\hat{\pi}, \hat{m}_z^\sharp, \boldsymbol{r}^*)$ is a vector of $(1 + d)$ dimension, in which the the first component corresponds to α^* (or a^*) and the other components correspond to $h\boldsymbol{\beta}^*$ (or $\boldsymbol{b}^*$). Note that $h\boldsymbol{\beta}^*$ is in a similar position to α^* in the expression of G_n^{AIPW}. We here focus on the derivation of the first component in $G_n^{AIPW}(\hat{\pi}, \hat{m}_z^\sharp, \boldsymbol{r}^*)$ and similar proof can be performed on components with respect to $h\boldsymbol{\beta}^*$. Let $g_z^{\sharp(1)}(Y_i, \boldsymbol{Z}_i, \boldsymbol{r}^*)$ and $m_z^{\sharp(1)}(\boldsymbol{X}_i, \boldsymbol{r}^*)$ ($\hat{m}_z^{\sharp(1)}(\boldsymbol{X}_i, \boldsymbol{r}^*)$) be the first components in $g_z^\sharp(Y_i, \boldsymbol{Z}_i, \boldsymbol{r}^*)$ and $m_z^\sharp(\boldsymbol{X}_i, \boldsymbol{r}^*)$ ($\hat{m}_z^\sharp(\boldsymbol{X}_i, \boldsymbol{r}^*)$), respectively. Denote $G_n^{(1)}(\hat{\pi}, \hat{m}_z^\sharp, \boldsymbol{r}^*)$ as the first component in $G_n^{AIPW}(\hat{\pi}, \hat{m}_z^\sharp, \boldsymbol{r}^*)$; that is

$$G_n^{(1)}(\hat{\pi}, \hat{m}_z^\sharp, \boldsymbol{r}^*) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\delta_i}{\hat{\pi}_i} g_z^{\sharp(1)}(Y_i, \boldsymbol{Z}_i, \boldsymbol{r}^*) + \left(1 - \frac{\delta_i}{\hat{\pi}_i}\right) \hat{m}_z^{\sharp(1)}(\boldsymbol{X}_i^o, \boldsymbol{r}^*)\right].$$

By simple calculations, we have

$$\begin{aligned}\sqrt{nh^d}G_n^{(1)}(\hat{\pi},\hat{m}_z^\sharp,\boldsymbol{r}^*) &= \frac{\sqrt{h^d}}{\sqrt{n}}\sum_{i=1}^n m_z^{\sharp(1)}\left(\boldsymbol{X}_i^o,\boldsymbol{r}^*\right)+\frac{\sqrt{h^d}}{\sqrt{n}}\sum_{i=1}^n\frac{\delta_i}{\pi_i}\left\{g_z^{\sharp(1)}\left(Y_i,\boldsymbol{Z}_i,\boldsymbol{r}^*\right)-m_z^{\sharp(1)}\left(\boldsymbol{X}_i^o,\boldsymbol{r}^*\right)\right\}\\ &\quad+\frac{\sqrt{h^d}}{\sqrt{n}}\sum_{i=1}^n\left(\frac{1}{\hat{\pi}_i}-\frac{1}{\pi_i}\right)\delta_i\left\{g_z^{\sharp(1)}\left(Y_i,\boldsymbol{Z}_i,\boldsymbol{r}^*\right)-m_z^{\sharp(1)}\left(\boldsymbol{X}_i^o,\boldsymbol{r}^*\right)\right\}\\ &\quad+\frac{\sqrt{h^d}}{\sqrt{n}}\sum_{i=1}^n\left(1-\frac{\delta_i}{\hat{\pi}_i}\right)\left\{\hat{m}_z^{\sharp(1)}\left(\boldsymbol{X}_i^o,\boldsymbol{r}^*\right)-m_z^{\sharp(1)}\left(\boldsymbol{X}_i^o,\boldsymbol{r}^*\right)\right\}\\ &=: L_1+L_2+L_3,\end{aligned}$$

where $L_1=\frac{\sqrt{h^d}}{\sqrt{n}}\sum_{i=1}^n m_z^{\sharp(1)}(\boldsymbol{X}_i^o,\boldsymbol{r}^*)+\frac{\sqrt{h^d}}{\sqrt{n}}\sum_{i=1}^n\frac{\delta_i}{\pi_i}\{g_z^{\sharp(1)}(Y_i,\boldsymbol{Z}_i,\boldsymbol{r}^*)-m_z^{\sharp(1)}(\boldsymbol{X}_i^o,\boldsymbol{r}^*)\}$ is the term we want to retain in the last. In the following, we will prove L_2 and L_3 are negligible under conditions (C6-2, C6-3) on bandwidths.

To deal with L_3 further, let $\hat{r}(X_i^o)=\frac{1}{n}\sum_{j=1}^n\tilde{\mathcal{K}}_{\bar{h}}(\boldsymbol{X}_j^o-\boldsymbol{X}_i^o)\delta_j$, an estimator of $r(\boldsymbol{X}_i^o)=\pi_i p(\boldsymbol{X}_i^o)$; and define

$$\xi_n\left(\boldsymbol{x}^o\right)=\frac{1}{n}\sum_{j=1}^n\overline{\mathcal{K}}_{\bar{h}}\left(\boldsymbol{x}^o-\boldsymbol{X}_j^o\right)\sqrt{h^d}\delta_j\left[g_z^{\sharp(1)}\left(Y_j,\boldsymbol{Z}_j,\boldsymbol{r}^*\right)-m_z^{\sharp(1)}\left(\boldsymbol{X}_j^o,\boldsymbol{r}^*\right)\right],$$

and

$$\psi_n\left(\boldsymbol{x}^o\right)=\frac{1}{n}\sum_{j=1}^n\overline{\mathcal{K}}_{\bar{h}}\left(\boldsymbol{x}^o-\boldsymbol{X}_j^o\right)\sqrt{h^d}\delta_j\left[m_z^{\sharp(1)}\left(\boldsymbol{X}_j^o,\boldsymbol{r}^*\right)-m_z^{\sharp(1)}\left(\boldsymbol{x}^o,\boldsymbol{r}^*\right)\right],$$

where $\boldsymbol{x}^o$ can be any fixed point in the support of $\boldsymbol{X}^o$. We put $\sqrt{h^d}$ inside the expressions of $\xi_n(x^o)$ and $\psi_n(x^o)$ to make them satisfy the condition $E|\eta(\boldsymbol{X}_i,\tilde{\boldsymbol{X}}_i)|^s<\infty$ in Lemma A.1 with $s=2$. Then for L_3 we have

$$\begin{aligned}&\frac{\sqrt{h^d}}{\sqrt{n}}\sum_{i=1}^n\left(1-\frac{\delta_i}{\hat{\pi}_i}\right)\left\{\hat{m}_z^{\sharp(1)}\left(\boldsymbol{X}_i^o,\boldsymbol{r}^*\right)-m_z^{\sharp(1)}\left(\boldsymbol{X}_i^o,\boldsymbol{r}^*\right)\right\}\\ &=\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^n\left(1-\frac{\delta_i}{\hat{\pi}_i}\right)\frac{\xi_n\left(\boldsymbol{X}_i^o\right)}{\hat{r}\left(\boldsymbol{X}_i^o\right)}+\frac{1}{n}\sum_{i=1}^n\left(1-\frac{\delta_i}{\hat{\pi}_i}\right)\frac{\psi_n\left(\boldsymbol{X}_i^o\right)}{\hat{r}\left(\boldsymbol{X}_i^o\right)}\right\}\\ &=\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^n\left(1-\frac{\delta_i}{\pi_i}\right)\frac{\xi_n\left(\boldsymbol{X}_i^o\right)}{r\left(\boldsymbol{X}_i^o\right)}+\frac{1}{n}\sum_{i=1}^n\left(1-\frac{\delta_i}{\pi_i}\right)\frac{\psi_n\left(\boldsymbol{X}_i^o\right)}{r\left(\boldsymbol{X}_i^o\right)}\right\}\\ &\quad+\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^n\xi_n\left(\boldsymbol{X}_i^o\right)\left[\left(\frac{1}{\hat{r}\left(\boldsymbol{X}_i^o\right)}-\frac{1}{r\left(\boldsymbol{X}_i^o\right)}\right)-\left(\frac{\delta_i}{\hat{r}\left(\boldsymbol{X}_i^o\right)\hat{\pi}_i}-\frac{\delta_i}{r\left(\boldsymbol{X}_i^o\right)\pi_i}\right)\right]\right.\\ &\quad\left.+\frac{1}{n}\sum_{i=1}^n\psi_n\left(\boldsymbol{X}_i^o\right)\left[\left(\frac{1}{\hat{r}\left(\boldsymbol{X}_i^o\right)}-\frac{1}{r\left(\boldsymbol{X}_i^o\right)}\right)-\left(\frac{\delta_i}{\hat{r}\left(\boldsymbol{X}_i^o\right)\hat{\pi}_i}-\frac{\delta_i}{r\left(\boldsymbol{X}_i^o\right)\pi_i}\right)\right]\right\}\\ &=:\sqrt{n}(L_{31}+L_{32})+\sqrt{n}(L_{33}+L_{34}).\end{aligned}$$

For L_{31}, applying Lemma A.1 to $\xi_n(\boldsymbol{x}^o)$ yields $\sup_{\boldsymbol{x}^o\in\mathcal{X}}|\xi_n(\boldsymbol{x}^o)|=o_p(1)$ since $E(\xi_n(\boldsymbol{x}^o))=0$. Then $L_{31}=o_p(1)O_p(1/\sqrt{n})=o_p(1/\sqrt{n})$. Therefore, the term $\sqrt{n}L_{31}$ is negligible.

Similarly, for L_{32}, applying Lemma A.1 to $\psi_n(\boldsymbol{x}^o)$ gives $\sup_{\boldsymbol{x}^o\in\mathcal{X}}|\psi_n(\boldsymbol{x}^o)-O(h^{d/2}\bar{h}^l)|=o_p(1)$ since $E(\psi_n(\boldsymbol{x}^o))=O(h^{d/2}\bar{h}^l)$. In order to make $\sqrt{n}L_{32}$ negligible, the condition $nh^d\bar{h}^{2l}\to 0$ is required. We instead impose a stronger condition $n\bar{h}^{2l}\to 0$ as specified in (C6-3), to make the bandwidth selection for $\bar{h}$ independent of h and therefore easier to be performed in practice.

For L_{33} and L_{34}, similar calculations as shown above with Lemma A.1 applied to $\widehat{r}(X_i^o)$ and $\hat{\pi}_i$ lead to $L_{33} = o_p(1/\sqrt{n}), L_{34} = o_p(1/\sqrt{n})$. Therefore, we have $L_3 = o_p(1)$.

Similarly, we can prove $L_2 = o_p(1)$. Thus,

$$\begin{aligned}\sqrt{nh^d}G_n^{(1)}(\hat{\pi}, \hat{m}_{\boldsymbol{z}}^{\sharp}, \boldsymbol{r}^*) &= L_1 + o_p(1)\\ &= \frac{\sqrt{h^d}}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{\delta_i}{\pi_i}g_{\boldsymbol{z}}^{\sharp(1)}\left(Y_i, Z_i, \boldsymbol{r}^*\right) + \left(1 - \frac{\delta_i}{\pi_i}\right)m_{\boldsymbol{z}}^{\sharp(1)}\left(X_i^o, \boldsymbol{r}^*\right)\right\} + o_p(1).\end{aligned}$$

■

Proof of Theorem 5.1: According to Theorem 5.9 of Van der Vaart (1998), we need to verify that $\inf_{\|\boldsymbol{r}-\boldsymbol{r}^*\|_2>\delta}\|G^{AIPW}(\pi, m_{\boldsymbol{z}}^{\sharp}, \boldsymbol{r}) - G^{AIPW}(\pi, m_{\boldsymbol{z}}^{\sharp}, \boldsymbol{r}^*)\|_2 > 0$ and $J_1 := \sup_{\boldsymbol{r}}\|G_n^{AIPW}(\hat{\pi}, \hat{m}_{\boldsymbol{z}}^{\sharp}, \boldsymbol{r}) - G^{AIPW}(\pi, m_{\boldsymbol{z}}^{\sharp}, \boldsymbol{r})\|_2 = o_p(1)$. First,

$$\begin{aligned}&\inf_{\|\boldsymbol{r}-\boldsymbol{r}^*\|_2>\delta}\|G^{AIPW}(\pi, m_{\boldsymbol{z}}^{\sharp}, \boldsymbol{r}) - G^{AIPW}(\pi, m_{\boldsymbol{z}}^{\sharp}, \boldsymbol{r}^*)\|_2\\ &= \inf_{\|\boldsymbol{r}-\boldsymbol{r}^*\|_2>\delta}\left\|E\left\{E\left[\varphi_\tau\left(Y_i - a - \left(\frac{\boldsymbol{Z}_i - \boldsymbol{z}}{h}\right)^{\top}\boldsymbol{b}\right) - \varphi_\tau\left(Y_i - a^*\right.\right.\right.\right.\\ &\quad \left.\left.\left.\left. - \left(\frac{\boldsymbol{Z}_i - \boldsymbol{z}}{h}\right)^{\top}\boldsymbol{b}^*\right)\middle|\boldsymbol{Z}_i\right]U_{\mathbf{H}}(\boldsymbol{Z}_i - \boldsymbol{z})\mathcal{K}_{\mathbf{H}}(\boldsymbol{Z}_i - \boldsymbol{z})\right\}\right\|_2\\ &= \inf_{\|\boldsymbol{r}-\boldsymbol{r}^*\|_2>\delta}\left\|E\left\{\left[F\left(a + \left(\frac{\boldsymbol{Z}_i - \boldsymbol{z}}{h}\right)^{\top}\boldsymbol{b}\middle|\boldsymbol{Z}_i\right)\right.\right.\right.\\ &\quad \left.\left.\left. - F\left(a^* + \left(\frac{\boldsymbol{Z}_i - \boldsymbol{z}}{h}\right)^{\top}\boldsymbol{b}^*\middle|\boldsymbol{Z}_i\right)\right]U_{\mathbf{H}}(\boldsymbol{Z}_i - \boldsymbol{z})\mathcal{K}_{\mathbf{H}}(\boldsymbol{Z}_i - \boldsymbol{z})\right\}\right\|_2\\ &= \inf_{\|\boldsymbol{r}-\boldsymbol{r}^*\|_2>\delta}\left\|E\left\{f\left(\tilde{a} + \left(\frac{\boldsymbol{Z}_i - \boldsymbol{z}}{h}\right)^{\top}\tilde{b}\middle|\boldsymbol{Z}_i\right)U_{\mathbf{H}}^{\top}(\boldsymbol{Z}_i - \boldsymbol{z})U_{\mathbf{H}}(\boldsymbol{Z}_i - \boldsymbol{z})\mathcal{K}_{\mathbf{H}}(\boldsymbol{Z}_i - \boldsymbol{z})\right\}\right\|_2\|\boldsymbol{r} - \boldsymbol{r}^*\|_2\\ &> 0,\end{aligned}$$

where $(\tilde{a}, \tilde{\boldsymbol{b}})$ lies between $\boldsymbol{r}$ and $\boldsymbol{r}^*$. Next, we consider J_1. Note that

$$\begin{aligned}J_1 &\le \sup_{\boldsymbol{r}}\|G_n^{AIPW}(\hat{\pi}, \hat{m}_{\boldsymbol{z}}^{\sharp}, \boldsymbol{r}) - G_n^{AIPW}(\pi, m_{\boldsymbol{z}}^{\sharp}, \boldsymbol{r})\|_2\\ &\quad + \sup_{\boldsymbol{r}}\|G_n^{AIPW}(\pi, m_{\boldsymbol{z}}^{\sharp}, \boldsymbol{r}) - G^{AIPW}(\pi, m_{\boldsymbol{z}}^{\sharp}, \boldsymbol{r})\|_2\\ &:= J_2 + J_3.\end{aligned}$$

Applying Lemma A.1 to $\hat{\pi}_i$ and $\hat{m}_{\boldsymbol{z}}^{\sharp}(\boldsymbol{X}_i^o, \boldsymbol{r})$, we have

$$\begin{aligned}J_2 &\le \sup_{\boldsymbol{r}}\left\|\frac{1}{n}\sum_{i=1}^{n}\delta_i\frac{\pi_i - \hat{\pi}_i}{\hat{\pi}_i\pi_i}g_{\boldsymbol{z}}^{\sharp}(Y_i, \boldsymbol{Z}_i, \boldsymbol{r})\right\|_2 + \left\|\frac{1}{n}\sum_{i=1}^{n}\delta_i\frac{\hat{\pi}_i - \pi_i}{\hat{\pi}_i\pi_i}m_{\boldsymbol{z}}^{\sharp}(\boldsymbol{X}_i^o, \boldsymbol{r})\right\|_2\\ &\quad + \left\|\frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{\delta_i}{\hat{\pi}_i}\right)(\hat{m}_{\boldsymbol{z}}^{\sharp}(\boldsymbol{X}_i^o, \boldsymbol{r}) - m_{\boldsymbol{z}}^{\sharp}(\boldsymbol{X}_i^o, \boldsymbol{r}))\right\|_2\\ &= o_p(1).\end{aligned}$$

For J_3, we have $J_3 = o_p(n^{-1/2})$ since $g_{\boldsymbol{z}}^{\sharp}(Y_i, \boldsymbol{Z}_i, \boldsymbol{r})$ is a Donsker class (see, Van der Vaart 1998; Chen et al. 2015). Thus, $J_1 = o_p(1)$ and this completes the proof of Theorem 5.1. ■

Proof of Theorem 5.2: Let $J(\boldsymbol{r}^*) = -\frac{\partial E^{(a^*,\boldsymbol{b}^*)}\{\mathcal{K}_{\mathbf{H}}(\boldsymbol{Z}_i-\boldsymbol{z})\varphi_\tau(Y_i-\alpha-\boldsymbol{b}^\top(\boldsymbol{Z}_i-\boldsymbol{z})/h)U_{\mathbf{H}}(\boldsymbol{Z}_i-\boldsymbol{z})\}}{\partial(a,\boldsymbol{b}^\top)}|_{a=a^*,\boldsymbol{b}=\boldsymbol{b}^*}$. Note that

$$\begin{aligned}\sqrt{nh^d}(\hat{\boldsymbol{r}}-\boldsymbol{r}^*) &= J^{-1}(\boldsymbol{r}^*)\sqrt{nh^d}G_n^{AIPW}(\hat{\pi},\hat{m}_{\boldsymbol{z}}^\sharp,\boldsymbol{r}^*)+o_p(\mathbf{1})\\ &= J^{-1}(\boldsymbol{r}^*)\frac{\sqrt{h^d}}{\sqrt{n}}\sum_{i=1}^n\left\{\frac{\delta_i}{\pi_i}g_{\boldsymbol{z}}^\sharp(Y_i,\boldsymbol{Z}_i,\boldsymbol{r}^*)+\left(1-\frac{\delta_i}{\pi_i}\right)m_{\boldsymbol{z}}^\sharp(\boldsymbol{X}_i,\boldsymbol{r}^*)\right\}+o_p(\mathbf{1}),\end{aligned}\tag{A1}$$

where the second equation holds due to Lemma A.2. It is easy to verify that

$$J(\boldsymbol{r}^*) = -f(a^*\mid\boldsymbol{z})p(\boldsymbol{z})D(\mathcal{K})+o(1),$$

where $D(\mathcal{K})$ is a $(1+d)\times(1+d)$ block diagonal matrix indicated as $\mathrm{diag}(1, D_d(\mathcal{K}))$ with $D_d(\mathcal{K}) = \mu_2(\mathcal{K})I_d$. As α (or a) is the first element in $\boldsymbol{r}$, we here restrict our attention to the first element of the right-hand side vector in (A1). We first consider the bias term:

$$\begin{aligned}&E\left\{\frac{\delta_i}{\pi_i}g_{\boldsymbol{z}}^{\sharp(1)}(Y_i,\boldsymbol{Z}_i,\boldsymbol{r}^*)+\left(1-\frac{\delta_i}{\pi_i}\right)m_{\boldsymbol{z}}^{\sharp(1)}(\boldsymbol{X}_i^o,\boldsymbol{r}^*)\right\}\\ &= E\left\{\mathcal{K}_{\mathbf{H}}(\boldsymbol{Z}_i-\boldsymbol{z})\varphi_\tau\left(Y_i-a^*-\boldsymbol{b}^{*\top}(\boldsymbol{Z}_i-\boldsymbol{z})/h\right)\right\}\\ &= E\left\{\mathcal{K}_{\mathbf{H}}(\boldsymbol{Z}_i-\boldsymbol{z})\left(\tau-I\left\{Y_i-\theta_\tau(\boldsymbol{Z}_i)+\theta_\tau(\boldsymbol{Z}_i)-a^*-\boldsymbol{b}^{*\top}(\boldsymbol{Z}_i-\boldsymbol{z})/h<0\right\}\right)\right\}.\end{aligned}\tag{A2}$$

Let $R(\boldsymbol{Z}_i)\equiv\theta_\tau(\boldsymbol{Z}_i)-a^*-\boldsymbol{b}^{*\top}(\boldsymbol{Z}_i-\boldsymbol{z})/h$. Using the law of iterated expectations, (A2) can be rewritten as

$$\begin{aligned}&E\Big\{\mathcal{K}_{\mathbf{H}}(\boldsymbol{Z}_i-\boldsymbol{z})\left[\tau-F\left(\theta_\tau(\boldsymbol{Z}_i)-R(\boldsymbol{Z}_i)\mid\boldsymbol{Z}_i\right)\right]\Big\}\\ &= E\left\{\mathcal{K}_{\mathbf{H}}(\boldsymbol{Z}_i-\boldsymbol{z})[\tau-F(\theta_\tau(\boldsymbol{Z}_i)\mid\boldsymbol{Z}_i)]\right\}+E\left\{\mathcal{K}_{\mathbf{H}}(\boldsymbol{Z}_i-\boldsymbol{z})f(\theta_\tau(\boldsymbol{Z}_i)\mid\boldsymbol{Z}_i)R(\boldsymbol{Z}_i)\right\}\\ &\quad-E\left\{\mathcal{K}_{\mathbf{H}}(\boldsymbol{Z}_i-\boldsymbol{z})\left[f(\theta_\tau(\boldsymbol{Z}_i)\mid\boldsymbol{Z}_i)R(\boldsymbol{Z}_i)+F(\theta_\tau(\boldsymbol{Z}_i)-R(\boldsymbol{Z}_i)\mid\boldsymbol{Z}_i)-F(\theta_\tau(\boldsymbol{Z}_i)\mid\boldsymbol{Z}_i)\right]\right\}\\ &:= W_1+W_2+W_3.\end{aligned}$$

Obviously, $W_1=0$. Applying Taylor expansion to $R(\boldsymbol{Z}_i)$, we have

$$\begin{aligned}W_2 &= \int\mathcal{K}(\boldsymbol{t})f(\theta_\tau(\boldsymbol{z}+\boldsymbol{t}h)\mid\boldsymbol{z}+\boldsymbol{t}h)\left(\frac{h^2}{2}\boldsymbol{t}^\top\mathcal{H}_{\theta_\tau}(\boldsymbol{z})\boldsymbol{t}+o(h^2)\right)p(\boldsymbol{z}+\boldsymbol{t}h)\,\mathrm{d}\boldsymbol{t}\\ &= \frac{h^2}{2}f(\theta_\tau(\boldsymbol{z})\mid\boldsymbol{z})p(\boldsymbol{z})\mu_2(\mathcal{K})\mathrm{tr}(\mathcal{H}_{\theta_\tau}(\boldsymbol{z}))+o(h^2).\end{aligned}$$

Similarly, it is to easy to prove that $W_3=o(h^2)$.

Applying variance decomposition formula, the variance can be calculated as follows

$$\begin{aligned}&\mathrm{Var}\left\{\frac{\delta_i}{\pi_i}g_{\boldsymbol{z}}^{\sharp(1)}(Y_i,\boldsymbol{Z}_i,\boldsymbol{r}^*)+\left(1-\frac{\delta_i}{\pi_i}\right)m_{\boldsymbol{z}}^{\sharp(1)}(\boldsymbol{X}_i^o,\boldsymbol{r}^*)\right\}\\ &= E\left\{\mathrm{Var}\left[\frac{\delta_i}{\pi_i}g_{\boldsymbol{z}}^{\sharp(1)}(Y_i,\boldsymbol{Z}_i,\boldsymbol{r}^*)+\left(1-\frac{\delta_i}{\pi_i}\right)m_{\boldsymbol{z}}^{\sharp(1)}(\boldsymbol{X}_i^o,\boldsymbol{r}^*)\mid\boldsymbol{X}_i^o\right]\right\}\\ &\quad+\mathrm{Var}\left\{E\left[\frac{\delta_i}{\pi_i}g_{\boldsymbol{z}}^{\sharp(1)}(Y_i,\boldsymbol{Z}_i,\boldsymbol{r}^*)+\left(1-\frac{\delta_i}{\pi_i}\right)m_{\boldsymbol{z}}^{\sharp(1)}(\boldsymbol{X}_i^o,\boldsymbol{r}^*)\mid\boldsymbol{X}_i^o\right]\right\}\\ &= E\left[\frac{\sigma_{g_z}^2(\boldsymbol{X}_i^o)}{\pi_i}\right]+\mathrm{Var}[m_{\boldsymbol{z}}^{\sharp(1)}(\boldsymbol{X}_i^o,\boldsymbol{r}^*)],\end{aligned}\tag{A3}$$

where $\sigma_{g_z}^2(\boldsymbol{X}_i^o)=\sigma_{g_z}^2(\boldsymbol{X}_i^o,\boldsymbol{r}^*)=\mathrm{Var}\{g_{\boldsymbol{z}}^{\sharp(1)}(Y_i,\boldsymbol{Z}_i,\boldsymbol{r}^*)\mid\boldsymbol{X}_i^o\}$, which is the conditional covariance of $g_{\boldsymbol{z}}^{\sharp(1)}(Y_i,\boldsymbol{Z}_i,\boldsymbol{r}^*)$ given $\boldsymbol{X}_i^o$. In order to determine the final converge rate of our proposed estimator, we need to deal with $\sigma_{g_z}^2(\boldsymbol{X}_i^o)$ further. Here, we consider two cases:

- *Case I*: Y is not included in $\boldsymbol{X}^o$ (i.e. some values of $\{Y_i\}_{i=1}^n$ are missing). Thus, the explanatory vector $\boldsymbol{Z}$ can be partitioned into two parts $(\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)})$, where $\boldsymbol{Z}^{(2)}$ denotes the components of $\boldsymbol{Z}$ that are missing for some individuals. In this case, $\boldsymbol{Z}^{(1)}$ is exactly $\boldsymbol{X}^o$, which is d_1 dimensional, and the dimension of $\boldsymbol{Z}^{(2)}$ denoted by d_2 is exactly $(d - d_1)$. Similarly, the given point $\boldsymbol{z}$ can be partitioned into $(\boldsymbol{z}^{(1)\top}, \boldsymbol{z}^{(2)\top})^\top$.
- *Case II*: Y is included in $\boldsymbol{X}^o$ (i.e. $\{Y_i\}_{i=1}^n$ are fully observed and the missing only comes from explanatory variables). Partition $\boldsymbol{Z}$ into two parts $(\boldsymbol{Z}^{(1)\top}, \boldsymbol{Z}^{(2)\top})^\top$, denoting fully observed variables and missing variables of $\boldsymbol{Z}$ respectively. Actually in this case, $\boldsymbol{X}^o = (Y, \boldsymbol{Z}^{(1)\top})^\top$.

Under Case I, we have

$$\begin{aligned}\sigma^2_{g_{\boldsymbol{z}}}(\boldsymbol{X}_i^o) &= E\left\{\mathrm{Var}\Big[\mathcal{K}_{\mathbf{H}}(\boldsymbol{Z}_i - \boldsymbol{z})\varphi_\tau(Y_i - a^* - \boldsymbol{b}^{*\top}(\boldsymbol{Z}_i - \boldsymbol{z})/h) \mid \boldsymbol{Z}_i\Big] \mid \boldsymbol{X}_i^o\right\} \\ &\quad + \mathrm{Var}\left\{E\Big[\mathcal{K}_{\mathbf{H}}(\boldsymbol{Z}_i - \boldsymbol{z})\varphi_\tau(Y_i - a^* - \boldsymbol{b}^{*\top}(\boldsymbol{Z}_i - \boldsymbol{z})/h) \mid \boldsymbol{Z}_i\Big] \mid \boldsymbol{X}_i^o\right\} \\ &:= I_1 + I_2. \end{aligned} \tag{A4}$$

Partition $\boldsymbol{b}$ (or $\boldsymbol{\beta}$) into $(\boldsymbol{b}_1, \boldsymbol{b}_2)$ (or $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$) to match with $(\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)})$ in Case I. The first term in (A4) becomes

$$\begin{aligned} I_1 &= E\left\{\mathcal{K}^2_{\mathbf{H}}(\boldsymbol{Z}_i - \boldsymbol{z})F\left(a^* + \boldsymbol{b}^{*\top}(\boldsymbol{Z}_i - \boldsymbol{z})/h \mid \boldsymbol{Z}_i\right)\left[1 - F\left(a^* + \boldsymbol{b}^{*\top}(\boldsymbol{Z}_i - \boldsymbol{z})/h \mid \boldsymbol{Z}_i\right)\right] \mid \boldsymbol{Z}_i^{(1)}\right\} \\ &= \mathcal{K}^2_{\mathbf{H}_{d_1}}(\boldsymbol{Z}_i^{(1)} - \boldsymbol{z}^{(1)})E\left\{\mathcal{K}^2_{\mathbf{H}_{d_2}}\left(\boldsymbol{Z}_i^{(2)} - \boldsymbol{z}^{(2)}\right)\right. \\ &\qquad F\left(a^* + \boldsymbol{b}_1^{*\top}\left(\boldsymbol{Z}_i^{(1)} - \boldsymbol{z}^{(1)}\right)/h + \boldsymbol{b}_2^{*\top}\left(\boldsymbol{Z}_i^{(2)} - \boldsymbol{z}^{(2)}\right)/h \mid \boldsymbol{Z}_i\right) \\ &\qquad \left.\times\left[1 - F\left(a^* + \boldsymbol{b}_1^{*\top}\left(\boldsymbol{Z}_i^{(1)} - \boldsymbol{z}^{(1)}\right)/h + \boldsymbol{b}_2^{*\top}\left(\boldsymbol{Z}_i^{(2)} - \boldsymbol{z}^{(2)}\right)/h \mid \boldsymbol{Z}_i\right)\right] \mid \boldsymbol{Z}_i^{(1)}\right\} \\ &= \frac{1}{h^{d_2}}\mathcal{K}^2_{\mathbf{H}_{d_1}}(\boldsymbol{Z}_i^{(1)} - \boldsymbol{z}^{(1)}) \int \mathcal{K}^2(\boldsymbol{t}_2)F\left(a^* + \boldsymbol{b}_1^{*\top}\left(\boldsymbol{Z}_i^{(1)} - \boldsymbol{z}^{(1)}\right)/h + \boldsymbol{b}_2^{*\top}\boldsymbol{t}_2 \mid \boldsymbol{Z}_i\right. \\ &\qquad \left.= \left(\boldsymbol{Z}_i^{(1)\top}, \boldsymbol{z}^{(2)\top} + h\boldsymbol{t}_2^\top\right)^\top\right) \\ &\qquad \times\left[1 - F\left(a^* + \boldsymbol{b}_1^{*\top}\left(\boldsymbol{Z}_i^{(1)} - \boldsymbol{z}^{(1)}\right)/h + \boldsymbol{b}_2^{*\top}\boldsymbol{t}_2 \mid \boldsymbol{Z}_i = \left(\boldsymbol{Z}_i^{(1)\top}, \boldsymbol{z}^{(2)\top} + h\boldsymbol{t}_2^\top\right)^\top\right)\right] \\ &\qquad \times p\left(\boldsymbol{z}^{(2)} + h\boldsymbol{t}_2 \mid \boldsymbol{Z}_i^{(1)}\right)\mathrm{d}\boldsymbol{t}_2. \end{aligned} \tag{A5}$$

Let $\gamma = a^* + \boldsymbol{b}_1^{*\top}(\boldsymbol{Z}_i^{(1)} - \boldsymbol{z}^{(1)})/h$ and note that $\boldsymbol{b} = h\boldsymbol{\beta}$. Applying the Taylor expansion to $F(\cdot \mid \cdot)$ and $f(\cdot \mid \cdot)$ in (A5) yields

$$\begin{aligned} I_1 &= \frac{1}{h^{d_2}}\mathcal{K}^2_{\mathbf{H}_{d_1}}(\boldsymbol{Z}_i^{(1)} - \boldsymbol{z}^{(1)}) \\ &\quad \times \int \mathcal{K}^2(\boldsymbol{t}_2)\left[F\left(\gamma \mid \boldsymbol{Z}_i = (\boldsymbol{Z}_i^{(1)\top}, \boldsymbol{z}^{(2)\top})^\top\right) + f\left(\gamma \mid \boldsymbol{Z}_i = (\boldsymbol{Z}_i^{(1)\top}, \boldsymbol{z}^{(2)\top})^\top\right)\boldsymbol{t}_2^\top\boldsymbol{\beta}_2^* h + o(h)\right] \\ &\quad \times\left[1 - F\left(\gamma \mid \boldsymbol{Z}_i = (\boldsymbol{Z}_i^{(1)\top}, \boldsymbol{z}^{(2)\top})^\top\right) - f\left(\gamma \mid \boldsymbol{Z}_i = (\boldsymbol{Z}_i^{(1)\top}, \boldsymbol{z}^{(2)\top})^\top\right)\boldsymbol{t}_2^\top\boldsymbol{\beta}_2^* h + o(h)\right] \\ &\quad \times\left[p(\boldsymbol{z}^{(2)} \mid \boldsymbol{Z}_i^{(1)}) + h\nabla_p^\top(\boldsymbol{z}^{(2)} \mid \boldsymbol{Z}_i^{(1)})\boldsymbol{t}_2 + o(h)\right]\mathrm{d}\boldsymbol{t}_2 \\ &= \frac{1}{h^{d_2}}\mathcal{K}^2_{\mathbf{H}_{d_1}}(\boldsymbol{Z}_i^{(1)} - \boldsymbol{z}^{(1)})p(\boldsymbol{z}^{(2)} \mid \boldsymbol{Z}_i^{(1)})F\left(\gamma \mid \boldsymbol{Z}_i = (\boldsymbol{Z}_i^{(1)\top}, \boldsymbol{z}^{(2)\top})^\top\right) \\ &\quad \times\left[1 - F\left(\gamma \mid \boldsymbol{Z}_i = (\boldsymbol{Z}_i^{(1)\top}, \boldsymbol{z}^{(2)\top})^\top\right)\right]\|\mathcal{K}_{d_2}\|_2^2, \end{aligned} \tag{A6}$$

where $\|\mathcal{K}_{d_2}\|_2^2 = \int \mathcal{K}_{d_2}^2(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}$ with $\mathcal{K}_{d_2}(\cdot)$ being the d_2-dimensional kernel function corresponding to $\boldsymbol{Z}^{(2)}$. Similarly, we have

$$I_2 = \frac{1}{h^{d_2}}\mathcal{K}_{\mathbf{H}_{d_1}}^2(\boldsymbol{Z}_i^{(1)} - \boldsymbol{z}^{(1)})p(\boldsymbol{z}^{(2)} \mid \boldsymbol{Z}_i^{(1)})\left[\tau - F\left(\gamma \mid \boldsymbol{Z}_i = (\boldsymbol{Z}_i^{(1)\top}, \boldsymbol{z}^{(2)\top})^\top\right)\right]^2 \|\mathcal{K}_{d_2}\|_2^2$$
$$- \mathcal{K}_{\mathbf{H}_{d_1}}^2(\boldsymbol{Z}_i^{(1)} - \boldsymbol{z}^{(1)})p^2(\boldsymbol{z}^{(2)} \mid \boldsymbol{Z}_i^{(2)})\left[\tau - F\left(\gamma \mid \boldsymbol{Z}_i = (\boldsymbol{Z}_i^{(1)\top}, \boldsymbol{z}^{(2)\top})^\top\right)\right]^2. \tag{A7}$$

Substituting (A6) and (A7) into (A4) yields

$$\sigma_{g_z}^2(\boldsymbol{X}_i^o) = \frac{1}{h^{d_2}}\mathcal{K}_{\mathbf{H}_{d_1}}^2(\boldsymbol{Z}_i^{(1)} - \boldsymbol{z}^{(1)})p(\boldsymbol{z}^{(2)} \mid \boldsymbol{Z}_i^{(1)})\left\{F\left(\gamma \mid \boldsymbol{Z}_i = (\boldsymbol{Z}_i^{(1)\top}, \boldsymbol{z}^{(2)\top})^\top\right)\right.$$
$$\times\left[1 - F\left(\gamma \mid \boldsymbol{Z}_i = (\boldsymbol{Z}_i^{(1)\top}, \boldsymbol{z}^{(2)\top})^\top\right)\right] + \left.\left[\tau - F\left(\gamma \mid \boldsymbol{Z}_i = (\boldsymbol{Z}_i^{(1)\top}, \boldsymbol{z}^{(2)\top})^\top\right)\right]^2\right\}\|\mathcal{K}_{d_2}\|_2^2$$
$$- \mathcal{K}_{\mathbf{H}_{d_1}}^2(\boldsymbol{Z}_i^{(1)} - \boldsymbol{z}^{(1)})p^2(\boldsymbol{z}^{(2)} \mid \boldsymbol{Z}_i^{(1)})\left[\tau - F\left(\gamma \mid \boldsymbol{Z}_i = (\boldsymbol{Z}_i^{(1)\top}, \boldsymbol{z}^{(2)\top})^\top\right)\right]^2.$$

Therefore,

$$E\left[\frac{\sigma_{g_z}^2(\boldsymbol{X}_i^o)}{\pi_i}\right]$$
$$= \frac{1}{h^{d_2+d_1}}\int \frac{\mathcal{K}^2(\boldsymbol{t}_1)p(\boldsymbol{z}^{(2)} \mid \boldsymbol{z}^{(1)} + \boldsymbol{t}_1 h)}{\pi(\boldsymbol{z}^{(1)} + \boldsymbol{t}_1 h)}\left\{F\left(a^* + h\boldsymbol{t}_1^\top\boldsymbol{\beta}_1^* \mid \boldsymbol{Z}_i = (\boldsymbol{z}^{(1)\top} + h\boldsymbol{t}_1^\top, \boldsymbol{z}^{(2)\top})^\top\right)\right.$$
$$\times\left[1 - F\left(a^* + h\boldsymbol{t}_1^\top\boldsymbol{\beta}_1^* \mid \boldsymbol{Z}_i = (\boldsymbol{z}^{(1)\top} + h\boldsymbol{t}_1^\top, \boldsymbol{z}^{(2)\top})^\top\right)\right]$$
$$\left.+\left[\tau - F\left(a^* + h\boldsymbol{t}_1^\top\boldsymbol{\beta}_1^* \mid \boldsymbol{Z}_i = (\boldsymbol{z}^{(1)\top} + h\boldsymbol{t}_1^\top, \boldsymbol{z}^{(2)\top})^\top\right)\right]^2\right\}\|\mathcal{K}_{d_2}\|_2^2 p(\boldsymbol{z}^{(1)} + \boldsymbol{t}_1 h)\,\mathrm{d}\boldsymbol{t}_1$$
$$- \frac{1}{h^{d_1}}\int \frac{\mathcal{K}^2(\boldsymbol{t}_1)p^2(\boldsymbol{z}^{(2)} \mid \boldsymbol{z}^{(1)} + \boldsymbol{t}_1 h)}{\pi(\boldsymbol{z}^{(1)} + \boldsymbol{t}_1 h)}\left[\tau - F\left(a^* + h\boldsymbol{t}_1^\top\boldsymbol{\beta}_1^* \mid \boldsymbol{Z}_i = (\boldsymbol{z}^{(1)\top} + h\boldsymbol{t}_1^\top, \boldsymbol{z}^{(2)\top})^\top\right)\right]^2$$
$$\times p(\boldsymbol{z}^{(1)} + \boldsymbol{t}_1 h)\,\mathrm{d}\boldsymbol{t}_1. \tag{A8}$$

Obviously, when $d_2 > 0$, the first integral is the leading term in (A8). Applying Taylor expansion and omitting the second term, we obtain

$$E\left[\frac{\sigma_{g_z}^2(\boldsymbol{X}_i^o)}{\pi_i}\right] = \frac{1}{h^{d_2+d_1}}\frac{\tau(1-\tau)p(\boldsymbol{z})}{\pi(\boldsymbol{z}^{(1)})}\|\mathcal{K}\|_2^2.$$

Simple calculations show that $\mathrm{Var}[m_{\boldsymbol{z}}^{\sharp(1)}(\boldsymbol{X}_i^o, \boldsymbol{r}^*)]$ in (A3) is $O(h^{-d_1})$, which is negligible compared with $E[\sigma_{g_z}^2(\boldsymbol{X}_i^o)/\pi_i]$ when $d_2 > 0$. It should be noted that $d_1 + d_2 = d$ and the dominant term of $\mathrm{Var}\{\delta_i g_{\boldsymbol{z}}^{\sharp(1)}(Y_i, \boldsymbol{Z}_i, \boldsymbol{r}^*)/\pi_i + (1 - \delta_i/\pi_i)m_{\boldsymbol{z}}^{\sharp(1)}(\boldsymbol{X}_i^o, \boldsymbol{r}^*)\}$ in Case I is $h^{-d}\tau(1-\tau)p(\boldsymbol{z})\|\mathcal{K}\|_2^2/\pi(\boldsymbol{z}^{(1)})$.

For Case II, using similar arguments above, we have

$$E\left[\frac{\sigma_{g_z}^2(\boldsymbol{X}_i^o)}{\pi_i}\right] = E\left[\frac{\sigma_{g_z}^2(Y_i, \boldsymbol{Z}_i^{(1)})}{\pi(Y_i, \boldsymbol{Z}_i^{(1)})}\right]$$
$$= \frac{p(\boldsymbol{z})}{h^d}\left\{\tau^2\int \mathcal{K}^2(\boldsymbol{t})\left[\int \frac{f(y \mid \boldsymbol{z})}{\pi(y, \boldsymbol{z}^{(1)})}\,\mathrm{d}y\right]\mathrm{d}\boldsymbol{t}\right.$$
$$\left.+(1-2\tau)\int \mathcal{K}^2(\boldsymbol{t})\left[\int_{\mathcal{D}} \frac{f(y \mid \boldsymbol{z})}{\pi(y, \boldsymbol{z}^{(1)})}\,\mathrm{d}y\right]\mathrm{d}\boldsymbol{t}\right\} + O\left(\frac{1}{h^{d_1-1}}\right),$$

where $\mathcal{D} = \{y : y - \theta(\boldsymbol{z}) - h\boldsymbol{t}^\top\nabla_\theta(\boldsymbol{z}) < 0\}$.

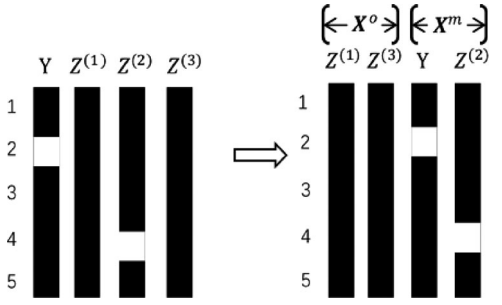

Figure A1. An illustration of the notations introduced in this article.

Table A1. The performance of MM algorithm under different settings of (κ, ε).

τ	κ	ε	K	$\hat{\theta}_1$	$\hat{\theta}_2$
0.1	10^{-3}	10^{-1}	17.34	0.7319	1.7369
		10^{-3}	7.72	1.0283	2.0299
		10^{-5}	4.74	1.0582	2.0634
		10^{-7}	4.99	1.0579	2.0612
	10^{-5}	10^{-1}	37.56	0.7255	1.7309
		10^{-3}	32.98	1.0013	1.9986
		10^{-5}	38.53	1.0054	2.0022
		10^{-7}	35.81	1.0047	2.0011
	10^{-7}	10^{-1}	57.83	0.7254	1.7309
		10^{-3}	70.11	0.9997	1.9965
		10^{-5}	76.63	1.0015	1.9985
		10^{-7}	71.16	1.0021	1.9987
0.9	10^{-3}	10^{-1}	16.24	1.2583	2.2673
		10^{-3}	7.94	0.9682	1.9748
		10^{-5}	4.95	0.9364	1.9429
		10^{-7}	4.76	0.9339	1.9424
	10^{-5}	10^{-1}	36.51	1.2663	2.2753
		10^{-3}	33.40	0.9957	2.0053
		10^{-5}	34.27	0.9911	2.0006
		10^{-7}	34.87	0.9906	1.9998
	10^{-7}	10^{-1}	56.65	1.2663	2.2753
		10^{-3}	80.99	0.9971	2.0071
		10^{-5}	83.10	0.9949	2.0053
		10^{-7}	73.09	0.9949	2.0052

In summary, $\mathrm{Var}\{\delta_i g_{\boldsymbol{z}}^{\sharp(1)}(Y_i, \boldsymbol{Z}_i, \boldsymbol{r}^*)/\pi_i + (1-\delta_i/\pi_i) m_{\boldsymbol{z}}^{\sharp(1)}(\boldsymbol{X}_i^o, \boldsymbol{r}^*)\}$ is dominated by $h^{-d} p(\boldsymbol{z})$ $c(\mathcal{K})$, and

$$c(\mathcal{K}) = \begin{cases} \tau(1-\tau)\|\mathcal{K}\|_2^2/\pi(\boldsymbol{z}^{(1)}), & \text{for Case I;} \\ \tau^2 \int \mathcal{K}^2(\boldsymbol{t}) \left[\int \dfrac{f(y \mid \boldsymbol{z})}{\pi(y, \boldsymbol{z}^{(1)})} \, \mathrm{d}y \right] \mathrm{d}\boldsymbol{t} + (1-2\tau) \int \mathcal{K}^2(\boldsymbol{t}) \left[\int_{\mathcal{D}} \dfrac{f(y \mid \boldsymbol{z})}{\pi(y, \boldsymbol{z}^{(1)})} \, \mathrm{d}y \right] \mathrm{d}\boldsymbol{t}, & \text{for Case II.} \end{cases}$$

■