entropy

MDPI

Article

Robust-BD Estimation and Inference for General Partially Linear Models

Chunming Zhang * and Zhengjun Zhang

Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA; zjz@stat.wisc.edu
* Correspondence: cmzhang@stat.wisc.edu

Received: 10 October 2017; Accepted: 16 November 2017; Published: 20 November 2017

Abstract: The classical quadratic loss for the partially linear model (PLM) and the likelihood function for the generalized PLM are not resistant to outliers. This inspires us to propose a class of "robust-Bregman divergence (BD)" estimators of both the parametric and nonparametric components in the general partially linear model (GPLM), which allows the distribution of the response variable to be partially specified, without being fully known. Using the local-polynomial function estimation method, we propose a computationally-efficient procedure for obtaining "robust-BD" estimators and establish the consistency and asymptotic normality of the "robust-BD" estimator of the parametric component $\boldsymbol{\beta}_{\mathrm{o}}$. For inference procedures of $\boldsymbol{\beta}_{\mathrm{o}}$ in the GPLM, we show that the Wald-type test statistic W_n constructed from the "robust-BD" estimators is asymptotically distribution free under the null, whereas the likelihood ratio-type test statistic Λ_n is not. This provides an insight into the distinction from the asymptotic equivalence (Fan and Huang 2005) between W_n and Λ_n in the PLM constructed from profile least-squares estimators using the non-robust quadratic loss. Numerical examples illustrate the computational effectiveness of the proposed "robust-BD" estimators and robust Wald-type test in the appearance of outlying observations.

Keywords: Bregman divergence; generalized linear model; local-polynomial regression; model check; nonparametric test; quasi-likelihood; semiparametric model; Wald statistic

1. Introduction

Semiparametric models, such as the partially linear model (PLM) and generalized PLM, play an important role in statistics, biostatistics, economics and engineering studies [1–5]. For the response variable Y and covariates $(\boldsymbol{X}, T)$, where $\boldsymbol{X} = (X_1, \ldots, X_d)^T \in \mathbb{R}^d$ and $T \in \mathcal{T} \subseteq \mathbb{R}^D$, the PLM, which is widely used for continuous responses Y, describes the model structure according to:

$$Y = \boldsymbol{X}^T\boldsymbol{\beta}_{\mathrm{o}} + \eta^{\mathrm{o}}(T) + \epsilon, \qquad \mathrm{E}(\epsilon \mid \boldsymbol{X}, T) = 0, \tag{1}$$

where $\boldsymbol{\beta}_{\mathrm{o}} = (\beta_{1;\mathrm{o}}, \ldots, \beta_{d;\mathrm{o}})^T$ is a vector of unknown parameters and $\eta^{\mathrm{o}}(\cdot)$ is an unknown smooth function; the generalized PLM, which is more suited to discrete responses Y and extends the generalized linear model [6], assumes:

$$m(\boldsymbol{x}, t) = \mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}, T = t) = F^{-1}(\boldsymbol{x}^T\boldsymbol{\beta}_{\mathrm{o}} + \eta^{\mathrm{o}}(t)), \tag{2}$$

$$Y \mid (\boldsymbol{X}, T) \sim \text{exponential family of distributions}, \tag{3}$$

where F is a known link function. Typically, the parametric component $\boldsymbol{\beta}_{\mathrm{o}}$ is of primary interest, while the nonparametric component $\eta^{\mathrm{o}}(\cdot)$ serves as a nuisance function. For illustration clarity, this paper focuses on $D = 1$. An important application of PLM to brain fMRI data was given in [7] for detecting activated brain voxels in response to external stimuli. There, $\boldsymbol{\beta}_{\mathrm{o}}$ corresponds to the part of hemodynamic response values, which is the object of primary interest to neuroscientists; $\eta^{\mathrm{o}}(\cdot)$ is the

slowly drifting baseline of time. Determining whether a voxel is activated or not can be formulated as testing for the linear form of hypotheses,

$$H_0 : \mathrm{A}\boldsymbol{\beta}_{\mathrm{o}} = \boldsymbol{g}_0 \qquad \text{versus} \qquad H_1 : \mathrm{A}\boldsymbol{\beta}_{\mathrm{o}} \neq \boldsymbol{g}_0, \tag{4}$$

where A is a given $\mathsf{k} \times d$ full row rank matrix and $\boldsymbol{g}_0$ is a known $\mathsf{k} \times 1$ vector.

Estimation of the parametric and nonparametric components of PLM and generalized PLM has received much attention in the literature. On the other hand, the existing work has some limitations: (i) The generalized PLM assumes that $Y \mid (\boldsymbol{X}, T)$ follows the distribution in (3), so that the likelihood function is fully available. From the practical viewpoint, results from the generalized PLM are not applicable to situations where the distribution of $Y \mid (\boldsymbol{X}, T)$ either departs from (3) or is incompletely known. (ii) Some commonly-used error measures, such as the quadratic loss in PLM for Gaussian-type responses (see for example [7,8]) and the (negative) likelihood function used in the generalized PLM, are not resistant to outliers. The work in [9] studied robust inference based on the kernel regression method for the generalized PLM with a canonical link, based on either the (negative) likelihood or (negative) quasi-likelihood as the error measure, and illustrated numerical examples with the dimension $d = 1$. However, the quasi-likelihood is not suitable for the exponential loss function (defined in Section 2.1), commonly used in machine learning and data mining. (iii) The work in [8] developed the inference of (4) for PLM, via the classical quadratic loss as the error measure, and demonstrated that the asymptotic distributions of the likelihood ratio-type statistic and Wald statistic under the null of (4) are both χ^2_{k}. It remains unknown whether this conclusion holds when the tests are constructed based on robust estimators.

Without completely specifying the distribution of $Y \mid (\boldsymbol{X}, T)$, we assume:

$$\mathrm{var}(Y \mid \boldsymbol{X} = \boldsymbol{x}, T = t) = V(m(\boldsymbol{x}, t)), \tag{5}$$

with a known functional form of $V(\cdot)$. We refer to a model specified by (2) and (5) as the "general partially linear model" (GPLM). This paper aims to develop robust estimation of GPLM and robust inference of $\boldsymbol{\beta}_{\mathrm{o}}$, allowing the distribution of $Y \mid (\boldsymbol{X}, T)$ to be partially specified. To introduce robust estimation, we adopt a broader class of robust error measures, called "robust-Bregman divergence (BD)" developed in [10], for a GLM, in which BD includes the quadratic loss, the (negative) quasi-likelihood, the exponential loss and many other commonly-used error measures as special cases. We propose the "robust-BD estimators" for both the parametric and nonparametric components of the GPLM. Distinct from the explicit-form estimators for PLM using the classical quadratic loss (see [8]), the "robust-BD estimators" for GPLM do not have closed-form expressions, which makes the theoretical derivation challenging. Moreover, the robust-BD estimators, as numerical solutions to non-linear optimization problems, pose key implementation challenges. Our major contributions are given below.

- The robust fitting of the nonparametric component $\eta^{\mathrm{o}}(\cdot)$ is formulated using the local-polynomial regression technique [11]. See Section 2.3.
- We develop a coordinate descent algorithm for the robust-BD estimator of $\boldsymbol{\beta}_{\mathrm{o}}$, which is computationally efficient particularly when the dimension d is large. See Section 3.
- Theorems 1 and 2 demonstrate that under the GPLM, the consistency and asymptotic normality of the proposed robust-BD estimator for $\boldsymbol{\beta}_{\mathrm{o}}$ are achieved. See Section 4.
- For robust inference of $\boldsymbol{\beta}_{\mathrm{o}}$, we propose a robust version of the Wald-type test statistic W_n, based on the robust-BD estimators, and justify its validity in Theorems 3–5. It is shown to be asymptotically χ^2 (central) under the null, thus distribution free, and χ^2 (noncentral) under the contiguous alternatives. Hence, this result, when applied to the exponential loss, as well as other loss functions in the wider class of BD, is practically feasible. See Section 5.1.

- For robust inference of $\boldsymbol{\beta}_o$, we re-examine the likelihood ratio-type test statistic Λ_n, constructed by replacing the negative log-likelihood with the robust-BD. Our Theorem 6 reveals that the asymptotic null distribution of Λ_n is generally not χ^2, but a linear combination of independent χ^2 variables, with weights relying on unknown quantities. Even in the particular case of using the classical-BD, the limit distribution is not invariant with re-scaling the generating function of the BD. Moreover, the limit null distribution of Λ_n (in either the non-robust or robust version) using the exponential loss, which does not belong to the (negative) quasi-likelihood, but falls in BD, is always a weighted χ^2, thus limiting its use in practical applications. See Section 5.2.

Simulation studies in Section 6 demonstrate that the proposed class of robust-BD estimators and robust Wald-type test either compare well with or perform better than the classical non-robust counterparts: the former is less sensitive to outliers than the latter, and both perform comparably well for non-contaminated cases. Section 7 illustrates some real data applications. Section 8 ends the paper with brief discussions. Details of technical derivations are relegated to Appendix A.

2. Robust-BD and Robust-BD Estimators

This section starts with a brief review of BD in Section 2.1 and "robust-BD" in Section 2.2, followed by the proposed "robust-BD" estimators of $\eta^o(\cdot)$ and $\boldsymbol{\beta}_o$ in Sections 2.3 and 2.4.

2.1. Classical-BD

To broaden the scope of robust estimation and inference, we consider a class of error measures motivated from the Bregman divergence (BD). For a given concave q-function, [12] defined a bivariate function,

$$Q_q(\nu, \mu) = -q(\nu) + q(\mu) + (\nu - \mu)q'(\mu). \tag{6}$$

We call Q_q the BD and call q the generating q-function of the BD. For example, a function $q(\mu) = a\mu - \mu^2$ for some constant a yields the quadratic loss $Q_q(Y, \mu) = (Y - \mu)^2$. For a binary response variable Y, $q(\mu) = \min\{\mu, (1-\mu)\}$ gives the misclassification loss $Q_q(Y, \mu) = \mathrm{I}\{Y \neq \mathrm{I}(\mu > 1/2)\}$, where $\mathrm{I}(\cdot)$ is an indicator function; $q(\mu) = -2\{\mu \log(\mu) + (1-\mu)\log(1-\mu)\}$ gives the Bernoulli deviance loss log-likelihood $Q_q(Y, \mu) = -2\{Y \log(\mu) + (1-Y)\log(1-\mu)\}$; $q(\mu) = 2\min\{\mu, (1-\mu)\}$ results in the hinge loss $Q_q(Y, \mu) = \max\{1 - (2Y-1)\,\mathrm{sign}(\mu - 0.5), 0\}$ of the support vector machine; $q(\mu) = 2\{\mu(1-\mu)\}^{1/2}$ yields the exponential loss $Q_q(Y, \mu) = \exp[-(Y - 0.5)\log\{\mu/(1-\mu)\}]$ used in AdaBoost [13]. Moreover, [14] showed that if:

$$q(\mu) = \int_a^\mu \frac{s - \mu}{V(s)} ds, \tag{7}$$

with a finite constant a such that the integral is well defined, then $Q_q(y, \mu)$ matches the "classical (negative) quasi-likelihood" function.

2.2. Robust-BD $\rho_q(y, \mu)$

Let $r(y, \mu) = (y - \mu)/\sqrt{V(\mu)}$ denote the Pearson residual, which reduces to the standardized residual for linear models. In contrast to the "classical-BD", denoted by Q_q in (6), the "robust-BD" developed in [10] for a GLM [6], is formed by:

$$\rho_q(y, \mu) = \int_y^\mu \psi(r(y, s))\{q''(s)\sqrt{V(s)}\}ds - G(\mu), \tag{8}$$

where $\psi(r)$ is chosen to be a bounded, odd function, such as the Huber ψ-function [15], $\psi(r) = r\min(1, c/|r|)$, and the bias-correction term, $G(\mu)$, entails the Fisher consistency of the parameter estimator and satisfies:

$$G'(\mu) = G_1'(\mu)\{q''(\mu)\sqrt{V(\mu)}\},$$

with

$$G_1'(m(\boldsymbol{x},t)) = \mathrm{E}\{\psi(r(Y,m(\boldsymbol{x},t))) \mid \boldsymbol{X}=\boldsymbol{x}, T=t\}. \tag{9}$$

We make the following discussions regarding features of the "robust-BD". To facilitate the discussion, we first introduce some necessary notation. Assume that the quantities:

$$\mathrm{p}_j(y;\theta) = \frac{\partial^j}{\partial\theta^j}\rho_q(y,F^{-1}(\theta)), \qquad j=0,1,\ldots, \tag{10}$$

exist finitely up to any order required. Then, we have the following expressions,

$$\begin{aligned}
\mathrm{p}_1(y;\theta) &= \{\psi(r(y,\mu)) - G_1'(\mu)\}\{q''(\mu)\sqrt{V(\mu)}\}/F'(\mu),\\
\mathrm{p}_2(y;\theta) &= A_0(y,\mu) + \{\psi(r(y,\mu)) - G_1'(\mu)\}A_1(\mu),\\
\mathrm{p}_3(y;\theta) &= A_2(y,\mu) + \{\psi(r(y,\mu)) - G_1'(\mu)\}A_1'(\mu)/F'(\mu),
\end{aligned} \tag{11}$$

where $\mu = F^{-1}(\theta)$,

$$A_0(y,\mu) = -\Big[\psi'(r(y,\mu))\Big\{1+\frac{y-\mu}{\sqrt{V(\mu)}}\times\frac{V'(\mu)}{2\sqrt{V(\mu)}}\Big\} + G_1''(\mu)\sqrt{V(\mu)}\Big]\frac{q''(\mu)}{\{F'(\mu)\}^2},$$

$A_1(\mu) = [\{q^{(3)}(\mu)\sqrt{V(\mu)} + 2^{-1}q''(\mu)V'(\mu)/\sqrt{V(\mu)}\}F'(\mu) - q''(\mu)\sqrt{V(\mu)}F''(\mu)]/\{F'(\mu)\}^3$ and $A_2(y,\mu) = [\partial A_0(y,\mu)/\partial\mu + \partial\{\psi(r(y,\mu)) - G_1'(\mu)\}/\partial\mu\, A_1(\mu)]/F'(\mu)$. Particularly, $\mathrm{p}_1(y;\theta)$ contains $\psi(r)$; $\mathrm{p}_2(y;\theta)$ contains $\psi(r)$, $\psi'(r)$ and $\psi'(r)r$; $\mathrm{p}_3(y;\theta)$ contains $\psi(r)$, $\psi'(r)$, $\psi'(r)r$, $\psi''(r)$, $\psi''(r)r$, and $\psi''(r)r^2$, where $r = r(y,\mu) = (y-\mu)/\sqrt{V(\mu)}$ denotes the Pearson residual. Accordingly, $\{\mathrm{p}_j(y;\theta) : j=1,2,3\}$ depend on y through $\psi(r)$ and its derivatives coupled with r. Then, we observe from (9) and (11) that:

$$\mathrm{E}\{\mathrm{p}_1(Y;\boldsymbol{X}^T\boldsymbol{\beta}_{\mathrm{o}} + \eta^{\mathrm{o}}(T)) \mid \boldsymbol{X}, T\} = 0. \tag{12}$$

In the particular choice of $\psi(r) = r$, it is clearly noticed from (9) that $G_1'(\cdot) = 0$, and thus, $G'(\cdot) = 0$. In such a case, the proposed "robust-BD" $\rho_q(y,\mu)$ reduces to the "classical-BD" $Q_q(y,\mu)$.

2.3. Local-Polynomial Robust-BD Estimator of $\eta^{\mathrm{o}}(\cdot)$

Let $\{(Y_i,\boldsymbol{X}_i,T_i)\}_{i=1}^n$ be i.i.d. observations of $(Y,\boldsymbol{X},T)$ captured by the GPLM in (2) and (5), where the dimension $d \geq 1$ is a finite integer. From (2), it is directly observed that if the true value of $\boldsymbol{\beta}_{\mathrm{o}}$ is known, then estimating $\eta^{\mathrm{o}}(\cdot)$ becomes estimating a nonparametric function; conversely, if the actual form of $\eta^{\mathrm{o}}(\cdot)$ is available, then estimating $\boldsymbol{\beta}_{\mathrm{o}}$ amounts to estimating a vector parameter.

To motivate the estimation of $\eta^{\mathrm{o}}(\cdot)$ at a fitting point t, a proper way to characterize $\eta^{\mathrm{o}}(t)$ is desired. For any given value of $\boldsymbol{\beta}$, define:

$$S(a;t,\boldsymbol{\beta}) = \mathrm{E}\{\rho_q(Y,F^{-1}(\boldsymbol{X}^T\boldsymbol{\beta}+a))\, w_1(\boldsymbol{X}) \mid T=t\}, \tag{13}$$

where a is a scalar, $\rho_q(y,\mu)$ is the "robust-BD" defined in (8), which aims to guard against outlying observations in the response space of Y, and $w_1(\cdot) \geq 0$ is a given bounded weight function that downweights high leverage points in the covariate space of $\boldsymbol{X}$. See Sections 6 and 7 for an example of $w_1(\boldsymbol{x})$. Set:

$$\eta_\beta(t) = \arg\min_{a\in\mathbb{R}^1} S(a;t,\boldsymbol{\beta}). \tag{14}$$

Theoretically, $\eta^{\mathrm{o}}(t) = \eta_{\beta_{\mathrm{o}}}(t)$ will be assumed (in Condition A3) for obtaining asymptotically unbiased estimators of $\eta^{\mathrm{o}}(\cdot)$. Such property indeed holds, for example, when a classical quadratic loss combined with an identity link is used in (14). Thus, we call $\eta_\beta(\cdot)$ the "surrogate function" for $\eta^{\mathrm{o}}(\cdot)$.

The characterization of the surrogate function $\eta_\beta(t)$ in (14) enables us to develop its robust-BD estimator $\widehat{\eta}_\beta(t)$ based on nonparametric function estimation. Assume that

$\eta^{\mathrm{o}}(\cdot)$ is $(p+1)$-times continuously differentiable at the fitting point t. Denote by $\mathbf{a}_{\mathrm{o}}(t) = (\eta^{\mathrm{o}}(t), (\eta^{\mathrm{o}})^{(1)}(t), \ldots, (\eta^{\mathrm{o}})^{(p)}(t)/p!)^T \in \mathbb{R}^{p+1}$ the vector consisting of $\eta^{\mathrm{o}}(t)$ along with its (re-scaled) derivatives. For observed covariates T_i close to the point t, the Taylor expansion implies that:

$$\begin{aligned} \eta^{\mathrm{o}}(T_i) &\approx \eta^{\mathrm{o}}(t) + (T_i - t)(\eta^{\mathrm{o}})^{(1)}(t) + \cdots + (T_i - t)^p (\eta^{\mathrm{o}})^{(p)}(t)/p! \\ &= \mathbf{t}_i(t)^T \mathbf{a}_{\mathrm{o}}(t), \end{aligned} \tag{15}$$

where $\mathbf{t}_i(t) = (1, (T_i - t), \ldots, (T_i - t)^p)^T$. For any given value of $\boldsymbol{\beta}$, let $\widehat{\boldsymbol{a}}(t;\boldsymbol{\beta}) = (\widehat{a}_0(t;\boldsymbol{\beta}), \widehat{a}_1(t;\boldsymbol{\beta}), \ldots, \widehat{a}_p(t;\boldsymbol{\beta}))^T$ be the minimizer of the criterion function,

$$S_n(\boldsymbol{a}; t, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \rho_q(Y_i, F^{-1}(\boldsymbol{X}_i^T\boldsymbol{\beta} + \mathbf{t}_i(t)^T\boldsymbol{a}))\, w_1(\boldsymbol{X}_i)\, K_h(T_i - t), \tag{16}$$

with respect to $\boldsymbol{a} \in \mathbb{R}^{p+1}$, where $K_h(\cdot) = K(\cdot/h)/h$ is re-scaled from a kernel function K and $h > 0$ is termed a bandwidth parameter. The first entry of $\widehat{\boldsymbol{a}}(t;\boldsymbol{\beta})$ supplies the local-polynomial robust-BD estimator $\widehat{\eta}_\beta(t)$ of $\eta_\beta(t)$, i.e.,

$$\widehat{\eta}_\beta(t) = \boldsymbol{e}_{1,p+1}^T \Big\{ \arg\min_{\boldsymbol{a}\in\mathbb{R}^{p+1}} S_n(\boldsymbol{a}; t, \boldsymbol{\beta}) \Big\}, \tag{17}$$

where $\boldsymbol{e}_{j,p+1}$ denotes the j-th column of a $(p+1)\times(p+1)$ identity matrix.

It is noted that the reliance of $\widehat{\eta}_\beta(t)$ on $\boldsymbol{\beta}$ does not guarantee its consistency to $\eta^{\mathrm{o}}(t)$. Nonetheless, it is anticipated from the uniform consistency of $\widehat{\eta}_{\widehat{\beta}}$ in Lemma 1 that $\widehat{\eta}_{\widehat{\beta}}(t)$ will offer a valid estimator of $\eta^{\mathrm{o}}(t)$, provided that $\widehat{\boldsymbol{\beta}}$ consistently estimates $\boldsymbol{\beta}_{\mathrm{o}}$. Section 2.4 will discuss our proposed robust-BD estimator $\widehat{\boldsymbol{\beta}}$. Furthermore, Lemma 1 will assume (in Condition A1) that $\eta_\beta(t)$ is the unique minimizer of $S(a; t, \boldsymbol{\beta})$ with respect to a.

Remark 1. *The case of using the "kernel estimation", or locally-constant estimation, corresponds to the choice of degree $p = 0$ in (15). In that case, the criterion function in (16) and the estimator in (17) reduce to:*

$$S_n(a; t, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \rho_q(Y_i, F^{-1}(\boldsymbol{X}_i^T\boldsymbol{\beta} + a))\, w_1(\boldsymbol{X}_i)\, K_h(T_i - t), \tag{18}$$

$$\widehat{\eta}_\beta(t) = \arg\min_{a\in\mathbb{R}^1} S_n(a; t, \boldsymbol{\beta}), \tag{19}$$

respectively.

2.4. Robust-BD Estimator of β_{o}

For any given value of $\boldsymbol{\beta}$, define:

$$J(\boldsymbol{\beta}, \eta_\beta) = \mathrm{E}\{\rho_q(Y, F^{-1}(\boldsymbol{X}^T\boldsymbol{\beta} + \eta_\beta(T)))\, w_2(\boldsymbol{X})\}, \tag{20}$$

where $\eta_\beta(\cdot)$ is as defined in (14) and $w_2(\cdot)$ plays the same role as $w_1(\cdot)$ in (13). Theoretically, it is anticipated that:

$$\boldsymbol{\beta}_{\mathrm{o}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} J(\boldsymbol{\beta}, \eta_\beta), \tag{21}$$

which holds for example in the case where a classical quadratic loss combined with an identity link is used. To estimate $\boldsymbol{\beta}_{\mathrm{o}}$, it is natural to replace (20) by its sample-based criterion,

$$J_n(\boldsymbol{\beta}, \widehat{\eta}_\beta) = \frac{1}{n}\sum_{i=1}^{n} \rho_q(Y_i, F^{-1}(\boldsymbol{X}_i^T\boldsymbol{\beta} + \widehat{\eta}_\beta(T_i)))\, w_2(\boldsymbol{X}_i), \tag{22}$$

where $\widehat{\eta}_{\beta}(\cdot)$ is as defined in (17). Hence, a parametric estimator of $\boldsymbol{\beta}_{\mathrm{o}}$ is provided by:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} J_n(\boldsymbol{\beta}, \widehat{\eta}_{\beta}). \tag{23}$$

Finally, the estimator of $\eta^{\mathrm{o}}(\cdot)$ is given by:

$$\widehat{\eta}(\cdot) = \widehat{\eta}_{\widehat{\beta}}(\cdot).$$

To achieve asymptotic normality of $\widehat{\boldsymbol{\beta}}$, Theorem 2 assumes (in Condition A2) that $\boldsymbol{\beta}_{\mathrm{o}}$ is the unique minimizer in (21), a standard condition for consistent M-estimators [16].

As a comparison, it is seen that $w_1(\cdot)$ in (16) is used to robustify covariates $\boldsymbol{X}_i$ in estimating $\eta^{\mathrm{o}}(\cdot)$, $w_2(\cdot)$ in (22) is used to robustify covariates $\boldsymbol{X}_i$ in estimating $\boldsymbol{\beta}_{\mathrm{o}}$ and $\rho_q(\cdot,\cdot)$ serves to robustify the responses Y_i in both estimating procedures.

3. Two-Step Iterative Algorithm for Robust-BD Estimation

In a special case of using the classical quadratic loss combined with an identity link function, the robust-BD estimators for parametric and nonparametric components have explicit expressions,

$$\widehat{\boldsymbol{\beta}} = (\widetilde{\mathbf{X}}^T\mathbf{w}_2\widetilde{\mathbf{X}})^{-1}(\widetilde{\mathbf{X}}^T\mathbf{w}_2\widetilde{\boldsymbol{y}}), \qquad (\widehat{\eta}(T_1),\ldots,\widehat{\eta}(T_n))^T = \boldsymbol{S}_h(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}), \tag{24}$$

where $\mathbf{w}_2 = \mathrm{diag}(w_2(\boldsymbol{X}_1),\ldots,w_2(\boldsymbol{X}_n))$, $\widetilde{\boldsymbol{y}} = (\mathbf{I} - \boldsymbol{S}_h)\boldsymbol{y}$, $\widetilde{\mathbf{X}} = (\mathbf{I} - \boldsymbol{S}_h)\mathbf{X}$, with $\mathbf{I}$ being an identity matrix, $\boldsymbol{y} = (Y_1,\ldots,Y_n)^T$, $\mathbf{X} = (\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)^T$ the design matrix,

$$\boldsymbol{S}_h = \begin{pmatrix} \boldsymbol{e}_{1,p+1}^T[\{\mathbf{T}(T_1)\}^T\mathbf{W}_{w_1;K}(T_1)\mathbf{T}(T_1)]^{-1}\{\mathbf{T}(T_1)\}^T\mathbf{W}_{w_1;K}(T_1) \\ \vdots \\ \boldsymbol{e}_{1,p+1}^T[\{\mathbf{T}(T_n)\}^T\mathbf{W}_{w_1;K}(T_n)\mathbf{T}(T_n)]^{-1}\{\mathbf{T}(T_n)\}^T\mathbf{W}_{w_1;K}(T_n) \end{pmatrix},$$

and:

$$\mathbf{T}(t) = (\mathbf{t}_1(t),\ldots,\mathbf{t}_n(t))^T, \quad \mathbf{W}_{w_1;K}(t) = \mathrm{diag}\{w_1(\boldsymbol{X}_i)\,K_h(T_i - t) : i = 1,\ldots,n\}.$$

When $w_1(\boldsymbol{x}) = w_2(\boldsymbol{x}) \equiv 1$, (24) reduces to the "profile least-squares estimators" of [8].

In other cases, robust-BD estimators from (17) and (23) do not have closed-form expressions and need to be solved numerically, which are computationally challenging and intensive. We now discuss a two-step robust proposal for iteratively estimating $\boldsymbol{\beta}_{\mathrm{o}}$ and $\eta^{\mathrm{o}}(\cdot)$. Let $\widehat{\boldsymbol{\beta}}^{[k-1]}$ and $\{\widehat{\eta}^{[k-1]}(T_i)\}_{i=1}^n$ denote the estimates in the $(k-1)$-th iteration, where $\widehat{\eta}^{[k-1]}(\cdot) = \widehat{\eta}_{\widehat{\boldsymbol{\beta}}^{[k-1]}}(\cdot)$. The k-th iteration consists of two steps below.

Step 1: Instead of solving (23) directly, we propose to solve a surrogate optimization problem, $\widehat{\boldsymbol{\beta}}^{[k]} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} J_n(\boldsymbol{\beta}, \widehat{\eta}^{[k-1]})$. This minimizer approximates $\widehat{\boldsymbol{\beta}}$.
Step 2: Obtain $\widehat{\eta}^{[k]}(T_i) = \widehat{\eta}_{\widehat{\boldsymbol{\beta}}^{[k]}}(T_i)$, $i = 1,\ldots,n$, where $\widehat{\eta}_{\beta}(t)$ is defined in (17).

The algorithm terminates provided that $\|\widehat{\boldsymbol{\beta}}^{[k]} - \widehat{\boldsymbol{\beta}}^{[k-1]}\|$ is below some pre-specified threshold value, and all $\{\widehat{\eta}^{[k]}(T_i)\}_{i=1}^n$ stabilize.

3.1. Step 1

For the above two-step algorithm, we first elaborate on the procedure of acquiring $\widehat{\boldsymbol{\beta}}^{[k]}$ in Step 1, by extending the coordinate descent (CD) iterative algorithm [17] designed for penalized estimation to our current robust-BD estimation, which is computationally efficient. For any given value of η,

by Taylor expansion, around some initial estimate $\boldsymbol{\beta}^*$ (for example, $\widehat{\boldsymbol{\beta}}^{[k-1]}$), we obtain the weighted quadratic approximation,

$$\rho_q(Y_i, F^{-1}(\boldsymbol{X}_i^T\boldsymbol{\beta}+\eta)) \approx \frac{1}{2}s_i^{\mathrm{I}}(Z_i^{\mathrm{I}} - \boldsymbol{X}_i^T\boldsymbol{\beta})^2 + C_i,$$

where C_i is a constant not depending on $\boldsymbol{\beta}$,

$$\begin{array}{rcl} s_i^{\mathrm{I}} & = & \mathrm{p}_2(Y_i; \boldsymbol{X}_i^T\boldsymbol{\beta}^* + \eta), \\ Z_i^{\mathrm{I}} & = & \boldsymbol{X}_i^T\boldsymbol{\beta}^* - \mathrm{p}_1(Y_i; \boldsymbol{X}_i^T\boldsymbol{\beta}^* + \eta)/\mathrm{p}_2(Y_i; \boldsymbol{X}_i^T\boldsymbol{\beta}^* + \eta), \end{array}$$

with $\mathrm{p}_j(y;\theta)$ defined in (10). Hence,

$$\begin{array}{rcl} J_n(\boldsymbol{\beta}, \eta) & = & \dfrac{1}{n}\displaystyle\sum_{i=1}^n \rho_q(Y_i, F^{-1}(\boldsymbol{X}_i^T\boldsymbol{\beta}+\eta))\, w_2(\boldsymbol{X}_i) \\ & \approx & \dfrac{1}{2}\displaystyle\sum_{i=1}^n \left\{n^{-1}s_i^{\mathrm{I}}\, w_2(\boldsymbol{X}_i)\right\}(Z_i^{\mathrm{I}} - \boldsymbol{X}_i^T\boldsymbol{\beta})^2 + \text{constant}. \end{array}$$

Thus it suffices to conduct minimization of $\sum_{i=1}^n s_i^{\mathrm{I}} w_2(\boldsymbol{X}_i)(Z_i^{\mathrm{I}} - \boldsymbol{X}_i^T\boldsymbol{\beta})^2$ with respect to $\boldsymbol{\beta}$, using a coordinate descent (CD) updating procedure. Suppose that the current estimate is $\widehat{\boldsymbol{\beta}}^{\mathrm{old}} = (\widehat{\beta}_1^{\mathrm{old}}, \ldots, \widehat{\beta}_d^{\mathrm{old}})^T$, with the current residual vector $\widehat{\boldsymbol{r}}^{\mathrm{old}} = (\widehat{r}_1^{\mathrm{old}}, \ldots, \widehat{r}_n^{\mathrm{old}}) = \boldsymbol{z}^{\mathrm{I}} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{\mathrm{old}}$, where $\boldsymbol{z}^{\mathrm{I}} = (Z_1^{\mathrm{I}}, \ldots, Z_n^{\mathrm{I}})^T$ is the vector of pseudo responses. Adopting the Newton–Raphson algorithm, the estimate of the j-th coordinate based on the previous estimate $\widehat{\beta}_j^{\mathrm{old}}$ is updated to:

$$\widehat{\beta}_j^{\mathrm{new}} = \widehat{\beta}_j^{\mathrm{old}} + \frac{\sum_{i=1}^n \{s_i^{\mathrm{I}} w_2(\boldsymbol{X}_i)\}\widehat{r}_i^{\mathrm{old}} X_{i,j}}{\sum_{i=1}^n \{s_i^{\mathrm{I}} w_2(\boldsymbol{X}_i)\} X_{i,j}^2}.$$

As a result, the residuals due to such an update are updated to:

$$\widehat{r}_i^{\mathrm{new}} = \widehat{r}_i^{\mathrm{old}} - X_{i,j}(\widehat{\beta}_j^{\mathrm{new}} - \widehat{\beta}_j^{\mathrm{old}}), \qquad i = 1, \ldots, n.$$

Cycling through $j = 1, \ldots, d$, we obtain the estimate $\widehat{\boldsymbol{\beta}}^{\mathrm{new}} = (\widehat{\beta}_1^{\mathrm{new}}, \ldots, \widehat{\beta}_d^{\mathrm{new}})^T$. Now, we set $\eta = \widehat{\eta}^{[k-1]}$ and $\boldsymbol{\beta}^* = \widehat{\boldsymbol{\beta}}^{[k-1]}$. Iterate the process of weighted quadratic approximation followed by the CD updating, for a number of times, until the estimate $\widehat{\boldsymbol{\beta}}^{\mathrm{new}}$ stabilizes to the solution $\widehat{\boldsymbol{\beta}}^{[k]}$.

The validity of $\widehat{\boldsymbol{\beta}}^{[k]}$ in Step 1 converging to the true parameter $\boldsymbol{\beta}_{\mathrm{o}}$ is justified as follows. (i) Standard results for M-estimation [16] indicate that the minimizer of $J_n(\boldsymbol{\beta}, \eta_{\boldsymbol{\beta}_{\mathrm{o}}})$ is consistent with $\boldsymbol{\beta}_{\mathrm{o}}$. (ii) According to our Theorem 1 (ii) in Section 4.1, $\sup_{t\in\mathcal{T}} |\widehat{\eta}_{\widehat{\boldsymbol{\beta}}}(t) - \eta_{\boldsymbol{\beta}_{\mathrm{o}}}(t)| \xrightarrow{\mathrm{P}} 0$ for a compact set $\mathcal{T}$, where $\xrightarrow{\mathrm{P}}$ stands for convergence in probability. Using derivations similar to those of (A4) gives $\sup_{\boldsymbol{\beta}\in\mathcal{K}} |J_n(\boldsymbol{\beta}, \widehat{\eta}_{\widehat{\boldsymbol{\beta}}}) - J_n(\boldsymbol{\beta}, \eta_{\boldsymbol{\beta}_{\mathrm{o}}})| \xrightarrow{\mathrm{P}} 0$ for any compact set $\mathcal{K}$. Thus, minimizing $J_n(\boldsymbol{\beta}, \widehat{\eta}_{\widehat{\boldsymbol{\beta}}})$ is asymptotically equivalent to minimizing $J_n(\boldsymbol{\beta}, \eta_{\boldsymbol{\beta}_{\mathrm{o}}})$. (iii) Similarly, provided that $\widehat{\boldsymbol{\beta}}^{[k-1]}$ is close to $\widehat{\boldsymbol{\beta}}$, minimizing $J_n(\boldsymbol{\beta}, \widehat{\eta}_{\widehat{\boldsymbol{\beta}}^{[k-1]}})$ is asymptotically equivalent to minimizing $J_n(\boldsymbol{\beta}, \widehat{\eta}_{\widehat{\boldsymbol{\beta}}})$. Assembling these three results with the definition of $\widehat{\boldsymbol{\beta}}^{[k]}$ yields:

$$\begin{array}{rcl} \widehat{\boldsymbol{\beta}}^{[k]} & = & \arg\min_{\boldsymbol{\beta}} J_n(\boldsymbol{\beta}, \widehat{\eta}_{\widehat{\boldsymbol{\beta}}^{[k-1]}}) \\ & = & \arg\min_{\boldsymbol{\beta}} J_n(\boldsymbol{\beta}, \widehat{\eta}_{\widehat{\boldsymbol{\beta}}}) + o_{\mathrm{P}}(1) \\ & = & \arg\min_{\boldsymbol{\beta}} J_n(\boldsymbol{\beta}, \eta_{\boldsymbol{\beta}_{\mathrm{o}}}) + o_{\mathrm{P}}(1) \\ & = & \boldsymbol{\beta}_{\mathrm{o}} + o_{\mathrm{P}}(1). \end{array}$$

3.2. Step 2

In Step 2, obtaining $\widehat{\eta}_{\boldsymbol{\beta}}(t)$ for any given values of $\boldsymbol{\beta}$ and t is equivalent to minimizing $S_n(\boldsymbol{a}; t, \boldsymbol{\beta})$ in (16). Notice that the dimension $(p+1)$ of $\boldsymbol{a}$ is typically low, with degrees $p = 0$ or $p = 1$ being the most commonly used in practice. Hence, the minimizer of $S_n(\boldsymbol{a}; t, \boldsymbol{\beta})$ can be obtained by directly applying the Newton–Raphson iteration: for $k = 0, 1, \ldots$,

$$\boldsymbol{a}^{[k+1]}(t; \boldsymbol{\beta}) = \boldsymbol{a}^{[k]}(t; \boldsymbol{\beta}) - \left\{ \frac{\partial^2 S_n(\boldsymbol{a}; t, \boldsymbol{\beta})}{\partial \boldsymbol{a} \partial \boldsymbol{a}^T} \Big|_{\boldsymbol{a}=\boldsymbol{a}^{[k]}(t;\boldsymbol{\beta})} \right\}^{-1} \frac{\partial S_n(\boldsymbol{a}; t, \boldsymbol{\beta})}{\partial \boldsymbol{a}} \Big|_{\boldsymbol{a}=\boldsymbol{a}^{[k]}(t;\boldsymbol{\beta})},$$

where $\boldsymbol{a}^{[k]}(t; \boldsymbol{\beta})$ denotes the estimate in the k-th iteration, and:

$$\begin{aligned} \frac{\partial S_n(\boldsymbol{a}; t, \boldsymbol{\beta})}{\partial \boldsymbol{a}} &= \frac{1}{n} \sum_{i=1}^{n} \mathrm{p}_1(Y_i; \boldsymbol{X}_i^T \boldsymbol{\beta} + \mathbf{t}_i(t)^T \boldsymbol{a})\ \mathbf{t}_i(t)\, w_1(\boldsymbol{X}_i)\, K_h(T_i - t), \\ \frac{\partial^2 S_n(\boldsymbol{a}; t, \boldsymbol{\beta})}{\partial \boldsymbol{a} \partial \boldsymbol{a}^T} &= \frac{1}{n} \sum_{i=1}^{n} \mathrm{p}_2(Y_i; \boldsymbol{X}_i^T \boldsymbol{\beta} + \mathbf{t}_i(t)^T \boldsymbol{a})\ \mathbf{t}_i(t) \mathbf{t}_i(t)^T\, w_1(\boldsymbol{X}_i)\, K_h(T_i - t). \end{aligned}$$

The iterations terminate until the estimate $\widehat{\eta}^{[k+1]}(t) = \boldsymbol{e}_{1,p+1}^T \boldsymbol{a}^{[k+1]}(t; \boldsymbol{\beta})$ stabilizes.

Our numerical studies of the robust-BD estimation indicate that (i) the kernel regression method can be both faster and stabler than the local-linear method; (ii) to estimate the nonparametric component $\eta^{\mathrm{o}}(\cdot)$, the local-linear method outperforms the kernel method, especially at the edges of points $\{T_i\}_{i=1}^n$; (iii) for the performance of the robust estimation of $\boldsymbol{\beta}_{\mathrm{o}}$, which is of major interest, there is a relatively negligible difference between choices of using the kernel and local-linear methods in estimating nonparametric components.

4. Asymptotic Property of the Robust-BD Estimators

This section investigates the asymptotic behavior of robust-BD estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\eta}_{\widehat{\boldsymbol{\beta}}}$, under regularity conditions. The consistency of $\widehat{\boldsymbol{\beta}}$ to $\boldsymbol{\beta}_{\mathrm{o}}$ and uniform consistency of $\widehat{\eta}_{\widehat{\boldsymbol{\beta}}}$ to η^{o} are given in Theorem 1; the asymptotic normality of $\widehat{\boldsymbol{\beta}}$ is obtained in Theorem 2. For the sake of exposition, the asymptotic results will be derived using local-linear estimation with degree $p = 1$. Analogous results can be obtained for local-polynomial methods with lengthier technical details and are omitted.

We assume that $T \in \mathcal{T}$, and let $\mathcal{T}_0 \subseteq \mathcal{T}$ be a compact set. For any continuous function $v : \mathcal{T} \mapsto \mathbb{R}$, define $\|v\|_\infty = \sup_{t \in \mathcal{T}} |v(t)|$ and $\|v\|_{\mathcal{T}_0;\infty} = \sup_{t \in \mathcal{T}_0} |v(t)|$. For a matrix M, the smallest and largest eigenvalues are denoted by $\lambda_j(M)$, $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$, respectively. Let $\|M\| = \sup_{\|\boldsymbol{x}\|=1} \|M\boldsymbol{x}\| = \{\lambda_{\max}(M^T M)\}^{1/2}$ be the matrix L_2 norm. Denote by $\xrightarrow{\mathrm{P}}$ convergence in probability and $\xrightarrow{\mathcal{D}}$ convergence in distribution.

4.1. Consistency

We first present Lemma 1, which states the uniform consistency of $\widehat{\eta}_{\boldsymbol{\beta}}(\cdot)$ to the surrogate function $\eta_{\boldsymbol{\beta}}(\cdot)$. Theorem 1 gives the consistency of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\eta}_{\widehat{\boldsymbol{\beta}}}$.

Lemma 1 (For the non-parametric surrogate $\eta_{\boldsymbol{\beta}}(\cdot)$)**.** *Let $\mathcal{K} \subseteq \mathbb{R}^d$ and $\mathcal{T}_0 \subseteq \mathcal{T}$ be compact sets. Assume Condition* A1 *and Condition* B *in the Appendix. If $n \to \infty$, $h \to 0$, $nh \to \infty$, $\log(1/h)/(nh) \to 0$, then $\sup_{\boldsymbol{\beta} \in \mathcal{K}} \|\widehat{\eta}_{\boldsymbol{\beta}} - \eta_{\boldsymbol{\beta}}\|_{\mathcal{T}_0;\infty} \xrightarrow{\mathrm{P}} 0$.*

Theorem 1 (For $\boldsymbol{\beta}_{\mathrm{o}}$ and $\eta^{\mathrm{o}}(\cdot)$)**.** *Assume conditions in Lemma 1.*

(i) *If there exists a compact set $\mathcal{K}_1$ such that $\lim_{n\to\infty} \mathrm{P}(\widehat{\boldsymbol{\beta}} \in \mathcal{K}_1) = 1$ and Condition* A2 *holds, then $\widehat{\boldsymbol{\beta}} \xrightarrow{\mathrm{P}} \boldsymbol{\beta}_{\mathrm{o}}$.*

(ii) *Moreover, if Condition* A3 *holds, then* $\|\widehat{\eta}_{\widehat{\beta}} - \eta^{\mathrm{o}}\|_{\mathcal{T}_0;\infty} \xrightarrow{\mathrm{P}} 0$.

4.2. Asymptotic Normality

The asymptotic normality of $\widehat{\boldsymbol{\beta}}$ is provided in Theorem 2.

Theorem 2 (For the parametric part $\boldsymbol{\beta}_{\mathrm{o}}$)**.** *Assume Conditions A and Condition B in the Appendix. If* $n \to \infty$, $nh^4 \to 0$ *and* $\log(1/h)/(nh^2) \to 0$, *then:*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathrm{o}}) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{H}_0^{-1}\Omega_0^*\mathbf{H}_0^{-1}),$$

where:

$$\mathbf{H}_0 = \mathrm{E}\Big[\mathrm{p}_2(Y; \boldsymbol{X}^T\boldsymbol{\beta}_{\mathrm{o}} + \eta^{\mathrm{o}}(T))\Big\{\boldsymbol{X} + \frac{\partial \eta_{\boldsymbol{\beta}}(T)}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_{\mathrm{o}}}\Big\}\Big\{\boldsymbol{X} + \frac{\partial \eta_{\boldsymbol{\beta}}(T)}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_{\mathrm{o}}}\Big\}^T w_2(\boldsymbol{X})\Big], \tag{25}$$

and:

$$\begin{aligned}\Omega_0^* = \mathrm{E}\Big(&\mathrm{p}_1^2(Y; \boldsymbol{X}^T\boldsymbol{\beta}_{\mathrm{o}} + \eta^{\mathrm{o}}(T))\Big[\Big\{\boldsymbol{X} + \frac{\partial \eta_{\boldsymbol{\beta}}(T)}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_{\mathrm{o}}}\Big\}w_2(\boldsymbol{X}) - \frac{\boldsymbol{\gamma}(T)}{g_2(T; T, \boldsymbol{\beta}_{\mathrm{o}})}w_1(\boldsymbol{X})\Big] \\ &\times\Big[\Big\{\boldsymbol{X} + \frac{\partial \eta_{\boldsymbol{\beta}}(T)}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_{\mathrm{o}}}\Big\}w_2(\boldsymbol{X}) - \frac{\boldsymbol{\gamma}(T)}{g_2(T; T, \boldsymbol{\beta}_{\mathrm{o}})}w_1(\boldsymbol{X})\Big]^T\Big)\end{aligned} \tag{26}$$

with:

$$\begin{aligned}\boldsymbol{\gamma}(t) &= \mathrm{E}\Big[\mathrm{p}_2(Y; \boldsymbol{X}^T\boldsymbol{\beta}_{\mathrm{o}} + \eta^{\mathrm{o}}(t))\Big\{\boldsymbol{X} + \frac{\partial \eta_{\boldsymbol{\beta}}(t)}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_{\mathrm{o}}}\Big\}w_2(\boldsymbol{X})\Big|T = t\Big], \\ g_2(t; t, \boldsymbol{\beta}) &= \mathrm{E}\{\mathrm{p}_2(Y; \boldsymbol{X}^T\boldsymbol{\beta} + \eta_{\boldsymbol{\beta}}(t))w_1(\boldsymbol{X}) \mid T = t\}.\end{aligned}$$

From Condition A1, (13) and (14), we can show that if $w_1(\cdot) \equiv Cw_2(\cdot)$ for some constant $C \in (0, \infty)$, then $\boldsymbol{\gamma}(t) = \mathbf{0}$. In that case, $\Omega_0^* = \Omega_0$, where:

$$\Omega_0 = \mathrm{E}\Big[\mathrm{p}_1^2(Y; \boldsymbol{X}^T\boldsymbol{\beta}_{\mathrm{o}} + \eta^{\mathrm{o}}(T))\Big\{\boldsymbol{X} + \frac{\partial \eta_{\boldsymbol{\beta}}(T)}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_{\mathrm{o}}}\Big\}\Big\{\boldsymbol{X} + \frac{\partial \eta_{\boldsymbol{\beta}}(T)}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_{\mathrm{o}}}\Big\}^T w_2^2(\boldsymbol{X})\Big]. \tag{27}$$

Consider the conventional PLM in (1), estimated using the classical quadratic loss, identity link and $w_1(\cdot) = w_2(\cdot) \equiv 1$. If $\mathrm{var}(\epsilon \mid \boldsymbol{X}, T) \equiv \sigma^2$, then $\mathbf{H}_0^{-1}\Omega_0\mathbf{H}_0^{-1} = \sigma^2[\mathrm{E}\{\mathrm{var}(\boldsymbol{X} \mid T)\}]^{-1}$, and thus, the result of Theorem 2 agrees with that in [18].

Remark 2. *Theorem 2 implies the root-n convergence rate of* $\widehat{\boldsymbol{\beta}}$. *This differs from* $\widehat{\eta}_{\widehat{\beta}}(t)$, *which converges at some rate incorporating both the sample size n and the bandwidth h, as seen in the proofs of Lemma 1 and Theorem 2.*

5. Robust Inference for $\boldsymbol{\beta}_{\mathbf{o}}$ Based on BD

In many statistical applications, we will check whether or not a subset of explanatory variables used is statistically significant. Specific examples include:

$$\begin{aligned}&H_0 : \beta_{j;\mathrm{o}} = 0, \qquad \text{for } j = j_0, \\ &H_0 : \beta_{j;\mathrm{o}} = 0, \qquad \text{for } j = j_1, \ldots, j_2.\end{aligned}$$

These forms of linear hypotheses for β_{o} can be more generally formulated as: (4).

5.1. Wald-Type Test W_n

We propose a robust version of the Wald-type test statistic,

$$W_n = n(\mathrm{A}\widehat{\boldsymbol{\beta}} - \boldsymbol{g}_0)^T(\mathrm{A}\widehat{\mathbf{H}}_0^{-1}\widehat{\Omega}_0^*\widehat{\mathbf{H}}_0^{-1}\mathrm{A}^T)^{-1}(\mathrm{A}\widehat{\boldsymbol{\beta}} - \boldsymbol{g}_0), \tag{28}$$

based on the robust-BD estimator $\widehat{\boldsymbol{\beta}}$ proposed in Section 2.4, where $\widehat{\Omega}_0^*$ and $\widehat{\mathbf{H}}_0$ are estimates of Ω_0^* and $\mathbf{H}_0$ satisfying $\widehat{\mathbf{H}}_0^{-1}\widehat{\Omega}_0^*\widehat{\mathbf{H}}_0^{-1} \xrightarrow{\mathrm{P}} \mathbf{H}_0^{-1}\Omega_0^*\mathbf{H}_0^{-1}$. For example,

$$\widehat{\mathbf{H}}_0 = \frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_2(Y_i; \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}} + \widehat{\eta}_{\widehat{\beta}}(T_i))\Big\{\boldsymbol{X}_i + \frac{\partial\widehat{\eta}_{\beta}(T_i)}{\partial\boldsymbol{\beta}}\Big|_{\beta=\widehat{\beta}}\Big\}\Big\{\boldsymbol{X}_i + \frac{\partial\widehat{\eta}_{\beta}(T_i)}{\partial\boldsymbol{\beta}}\Big|_{\beta=\widehat{\beta}}\Big\}^T w_2(\boldsymbol{X}_i),$$

and:

$$\begin{aligned}\widehat{\Omega}_0^* &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_1^2(Y_i; \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}} + \widehat{\eta}_{\widehat{\beta}}(T_i))\Big[\Big\{\boldsymbol{X}_i + \frac{\partial\widehat{\eta}_{\beta}(T_i)}{\partial\boldsymbol{\beta}}\Big|_{\beta=\widehat{\beta}}\Big\}w_2(\boldsymbol{X}_i) - \frac{\widehat{\boldsymbol{\gamma}}(T_i)}{\widehat{g}_2(T_i; T_i, \widehat{\boldsymbol{\beta}})}w_1(\boldsymbol{X}_i)\Big] \\ &\quad \times\Big[\Big\{\boldsymbol{X}_i + \frac{\partial\widehat{\eta}_{\beta}(T_i)}{\partial\boldsymbol{\beta}}\Big|_{\beta=\widehat{\beta}}\Big\}w_2(\boldsymbol{X}_i) - \frac{\widehat{\boldsymbol{\gamma}}(T_i)}{\widehat{g}_2(T_i; T_i, \widehat{\boldsymbol{\beta}})}w_1(\boldsymbol{X}_i)\Big]^T,\end{aligned}$$

fulfill the requirement, where:

$$\begin{aligned}\frac{\partial\widehat{\eta}_{\beta}(t)}{\partial\boldsymbol{\beta}} &= -\frac{\sum_{k=1}^{n} \mathrm{p}_2(Y_k; \boldsymbol{X}_k^T\boldsymbol{\beta} + \widehat{\eta}_{\beta}(t))\boldsymbol{X}_k\, w_1(\boldsymbol{X}_k)\, K_h(T_k - t)}{\sum_{k=1}^{n} \mathrm{p}_2(Y_k; \boldsymbol{X}_k^T\boldsymbol{\beta} + \widehat{\eta}_{\beta}(t))\, w_1(\boldsymbol{X}_k)\, K_h(T_k - t)}, \\ \widehat{\boldsymbol{\gamma}}(t) &= \frac{1}{n}\sum_{k=1}^{n} \mathrm{p}_2(Y_k; \boldsymbol{X}_k^T\widehat{\boldsymbol{\beta}} + \widehat{\eta}_{\widehat{\beta}}(t))\Big\{\boldsymbol{X}_k + \frac{\partial\widehat{\eta}_{\beta}(t)}{\partial\boldsymbol{\beta}}\Big|_{\beta=\widehat{\beta}}\Big\}w_2(\boldsymbol{X}_k)\, K_h(T_k - t), \\ \widehat{g}_2(t; t, \boldsymbol{\beta}) &= \frac{1}{n}\sum_{k=1}^{n} \mathrm{p}_2(Y_k; \boldsymbol{X}_k^T\boldsymbol{\beta} + \widehat{\eta}_{\beta}(t))w_1(\boldsymbol{X}_k)\, K_h(T_k - t).\end{aligned}$$

Again, we can verify that if $w_1(\cdot) \equiv Cw_2(\cdot)$ for some constant $C \in (0, \infty)$ and $\widehat{\eta}_{\beta}(t)$ is obtained from kernel estimation method, then $\widehat{\boldsymbol{\gamma}}(t) = \mathbf{0}$, and hence, $\widehat{\Omega}_0^* = \widehat{\Omega}_0$, where:

$$\widehat{\Omega}_0 = \frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_1^2(Y_i; \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}} + \widehat{\eta}_{\widehat{\beta}}(T_i))\Big\{\boldsymbol{X}_i + \frac{\partial\widehat{\eta}_{\beta}(T_i)}{\partial\boldsymbol{\beta}}\Big|_{\beta=\widehat{\beta}}\Big\}\Big\{\boldsymbol{X}_i + \frac{\partial\widehat{\eta}_{\beta}(T_i)}{\partial\boldsymbol{\beta}}\Big|_{\beta=\widehat{\beta}}\Big\}^T w_2^2(\boldsymbol{X}_i).$$

Theorem 3 justifies that under the null, W_n would for large n be distributed as χ_{k}^2, thus asymptotically distribution-free.

Theorem 3 (Wald-type test based on robust-BD under H_0)**.** *Assume conditions in Theorem 2, and* $\widehat{\mathbf{H}}_0^{-1}\widehat{\Omega}_0^*\widehat{\mathbf{H}}_0^{-1} \xrightarrow{\mathrm{P}} \mathbf{H}_0^{-1}\Omega_0^*\mathbf{H}_0^{-1}$ *in* (28)*. Then, under* H_0 *in* (4)*, we have that:*

$$W_n \xrightarrow{\mathcal{D}} \chi_k^2.$$

Theorem 4 indicates that W_n has a non-trivial local power detecting contiguous alternatives approaching the null at the rate $n^{-1/2}$:

$$H_{1n} : \mathrm{A}\boldsymbol{\beta}_{\mathrm{o}} - \boldsymbol{g}_0 = \boldsymbol{c}/\sqrt{n}\,\{1 + o(1)\}, \tag{29}$$

where $\boldsymbol{c} = (c_1, \ldots, c_{\mathrm{k}})^T \neq \mathbf{0}$.

Theorem 4 (Wald-type test based on robust-BD under H_{1n})**.** *Assume conditions in Theorem 2, and* $\widehat{\mathbf{H}}_0^{-1}\widehat{\Omega}_0^*\widehat{\mathbf{H}}_0^{-1} \xrightarrow{\mathrm{P}} \mathbf{H}_0^{-1}\Omega_0^*\mathbf{H}_0^{-1}$ *in* (28)*. Then, under* H_{1n} *in* (29)*,* $W_n \xrightarrow{\mathcal{D}} \chi_k^2(\tau^2)$*, where* $\tau^2 = \boldsymbol{c}^T(\mathrm{A}\mathbf{H}_0^{-1}\Omega_0^*\mathbf{H}_0^{-1}\mathrm{A}^T)^{-1}\boldsymbol{c} > 0$.

To appreciate the discriminating power of W_n in assessing the significance, the asymptotic power is analyzed. Theorem 5 manifests that under the fixed alternative H_1, $W_n \xrightarrow{\mathrm{P}} +\infty$ at the rate n. Thus, W_n has the power approaching to one against fixed alternatives.

Theorem 5 (Wald-type test based on robust-BD under H_1)**.** *Assume conditions in Theorem 2, and* $\widehat{\mathbf{H}}_0^{-1}\widehat{\Omega}_0^*\widehat{\mathbf{H}}_0^{-1} \xrightarrow{\mathrm{P}} \mathbf{H}_0^{-1}\Omega_0^*\mathbf{H}_0^{-1}$ *in (28). Then, under* H_1 *in (4),* $n^{-1}W_n \geq \lambda_{\max}^{-1}(\mathrm{A}\mathbf{H}_0^{-1}\Omega_0^*\mathbf{H}_0^{-1}\mathrm{A}^T)\|\mathrm{A}\boldsymbol{\beta}_{\mathrm{o}} - \boldsymbol{g}_0\|^2 + o_{\mathrm{P}}(1)$.

For the conventional PLM in (1) estimated using the non-robust quadratic loss, [8] showed the asymptotic equivalence between the Wald-type test and likelihood ratio-type test. Our results in the next Section 5.2 reveal that such equivalence is violated when estimators are obtained using the robust loss functions.

5.2. Likelihood Ratio-Type Test Λ_n

This section explores the degree to which the likelihood ratio-type test is extended to the "robust-BD" for testing the null hypothesis in (4) for the GPLM. The robust-BD test statistic is:

$$\Lambda_n = 2n\Big\{\min_{\boldsymbol{\beta}\in\mathbb{R}^d:\mathrm{A}\boldsymbol{\beta}=\boldsymbol{g}_0} J_n(\boldsymbol{\beta},\widehat{\eta}_\beta) - J_n(\widehat{\boldsymbol{\beta}},\widehat{\eta}_{\widehat{\beta}})\Big\}, \tag{30}$$

where $\widehat{\boldsymbol{\beta}}$ is the robust-BD estimator for $\boldsymbol{\beta}_{\mathrm{o}}$ developed in Section 2.4.

Theorem 6 indicates that the limit distribution of Λ_n under H_0 is a linear combination of independent chi-squared variables, with weights relying on some unknown quantities, thus not distribution free.

Theorem 6 (Likelihood ratio-type test based on robust-BD under H_0)**.** *Assume conditions in Theorem 2.*

(i) *Under* H_0 *in (4), we obtain:*

$$\Lambda_n \xrightarrow{\mathcal{D}} \sum_{j=1}^{k}\lambda_j\{(\mathrm{A}\mathbf{H}_0^{-1}\mathrm{A}^T)^{-1}(\mathrm{A}\mathbf{V}_0\mathrm{A}^T)\}Z_j^2,$$

where $\mathbf{V}_0 = \mathbf{H}_0^{-1}\Omega_0^*\mathbf{H}_0^{-1}$ *and* $\{Z_j\}_{j=1}^k \overset{\text{i.i.d.}}{\sim} N(0,1)$.

(ii) *Moreover, if* $\psi(r) = r$, $w_1(\boldsymbol{x}) = w_2(\boldsymbol{x}) \equiv 1$, *and the generating q-function of BD satisfies:*

$$q''(m(\boldsymbol{x},t)) = -\frac{C}{V(m(\boldsymbol{x},t))}, \quad \textit{for a constant } C > 0, \tag{31}$$

then under H_0 *in (4), we have that* $\Lambda_n/C \xrightarrow{\mathcal{D}} \chi_k^2$.

Theorem 7 states that Λ_n has non-trivial local power for identifying contiguous alternatives approaching the null at rate $n^{-1/2}$ and that $\Lambda_n \xrightarrow{\mathrm{P}} +\infty$ at the rate n under H_1, thus having the power approaching to one against fixed alternatives.

Theorem 7 (Likelihood ratio-type test based on robust-BD under H_{1n} and H_1)**.** *Assume conditions in Theorem 2. Let* $\mathbf{V}_0 = \mathbf{H}_0^{-1}\Omega_0^*\mathbf{H}_0^{-1}$ *and* $\lambda_j = \lambda_j\{(\mathrm{A}\mathbf{H}_0^{-1}\mathrm{A}^T)^{-1}(\mathrm{A}\mathbf{V}_0\mathrm{A}^T)\}$, $j = 1,\ldots,k$.

(i) *Under* H_{1n} *in (29),* $\Lambda_n \xrightarrow{\mathcal{D}} \sum_{j=1}^k(\sqrt{\lambda_j}Z_j + \boldsymbol{e}_{j,k}^T S\boldsymbol{c})^2$, *where* $\{Z_j\}_{j=1}^k \overset{\text{i.i.d.}}{\sim} N(0,1)$, *and* S *is a matrix satisfying* $S^TS = (\mathrm{A}\mathbf{H}_0^{-1}\mathrm{A}^T)^{-1}$ *and* $S(\mathrm{A}\mathbf{V}_0\mathrm{A}^T)S^T = \mathrm{diag}(\lambda_1,\ldots,\lambda_k)$.

(ii) *Under* H_1 *in (4),* $n^{-1}\Lambda_n \geq c\|\mathrm{A}\boldsymbol{\beta}_{\mathrm{o}} - \boldsymbol{g}_0\|^2 + o_{\mathrm{P}}(1)$ *for a constant* $c > 0$.

5.3. Comparison between W_n and Λ_n

In summary, the test W_n has some advantages over the test Λ_n. First, the asymptotic null distribution of W_n is distribution-free, whereas the asymptotic null distribution of Λ_n in general depends on unknown quantities. Second, W_n is invariant with re-scaling the generating q-function of the BD, but Λ_n is not. Third, the computational expense of W_n is much more reduced than that of Λ_n, partly because the integration operations for ρ_q are involved in Λ_n, but not in W_n, and partly because Λ_n requires both unrestricted and restricted parameter estimates, while W_n is useful in cases where restricted parameter estimates are difficult to compute. Thus, W_n will be focused on in numerical studies of Section 6.

6. Simulation Study

We conduct simulation evaluations of the performance of robust-BD estimation methods for general partially linear models. We use the Huber ψ-function $\psi(\cdot)$ with $c = 1.345$. The weight functions are chosen to be $w_1(\boldsymbol{x}) = w_2(\boldsymbol{x}) = 1/\{1 + \sum_{j=1}^{d}(\frac{x_j - m_j}{s_j})^2\}^{1/2}$, where $\boldsymbol{x} = (x_1, \ldots, x_d)^T$, m_j and s_j denote the sample median and sample median absolute deviation of $\{X_{i,j} : i = 1, \ldots, n\}$ respectively, $j = 1, \ldots, d$. As a comparison, the classical non-robust estimation counterparts correspond to using $\psi(r) = r$ and $w_1(\boldsymbol{x}) = w_2(\boldsymbol{x}) \equiv 1$. Throughout the numerical work, the Epanechnikov kernel function $K(t) = 0.75\max(1 - t^2, 0)$ is used. All these choices (among many others) are for feasibility; the issues on the trade-off between robustness and efficiency are not pursued further in the paper.

The following setup is used in the simulation studies. The sample size is $n = 200$, and the number of replications is 500. (Incorporating a nonparametric component in the GPLM desires a larger n when the number of covariates increases for better numerical performance.) Local-linear robust-BD estimation is illustrated with the bandwidth parameter h to be 20% of the interval length of the variable T. Results using other data-driven choices of h are similar and are omitted.

6.1. Bernoulli Responses

We generate observations $\{(\boldsymbol{X}_i, T_i, Y_i)\}_{i=1}^n$ randomly from the model,

$$Y \mid (\boldsymbol{X}, T) \sim \text{Bernoulli}(m(\boldsymbol{X}, T)), \quad \boldsymbol{X} \sim N(\boldsymbol{0}, \Sigma), \quad T \sim \text{Uniform}(0, 1),$$

where $\Sigma = (\sigma_{jk})$ with $\sigma_{jk} = 0.2^{|j-k|}$, and $\boldsymbol{X}$ is independent of T. The link function is $\text{logit}\{m(\boldsymbol{x}, t)\} = \boldsymbol{x}^T\boldsymbol{\beta}_o + \eta^o(t)$, where $\boldsymbol{\beta}_o = (2, 2, 0, 0)^T$ and $\eta^o(t) = 2\sin\{\pi(1 + 2t)\}$. Both the deviance and exponential loss functions are employed as the BD.

For each generated dataset from the true model, we create a contaminated dataset, where 10 data points $(X_{i,j}, Y_i)$ are contaminated as follows: they are replaced by $(X^*_{i,j}, Y^*_i)$, where $Y^*_i = 1 - Y_i$, $i = 1, \ldots, 5$,

$$\begin{aligned}
&X^*_{1,2} = 5\,\text{sign}(U_1 - 0.5), \quad X^*_{2,2} = 5\,\text{sign}(U_2 - 0.5), \quad X^*_{3,2} = 5\,\text{sign}(U_3 - 0.5),\\
&X^*_{4,4} = 5\,\text{sign}(U_4 - 0.5), \quad X^*_{5,1} = 5\,\text{sign}(U_5 - 0.5), \quad X^*_{6,2} = 5\,\text{sign}(U_6 - 0.5),\\
&X^*_{7,3} = 5\,\text{sign}(U_7 - 0.5), \quad X^*_{8,4} = 5\,\text{sign}(U_8 - 0.5), \quad X^*_{9,2} = 5\,\text{sign}(U_9 - 0.5),\\
&X^*_{10,3} = 5\,\text{sign}(U_{10} - 0.5),
\end{aligned}$$

with $\{U_i\} \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0, 1)$.

Figures 1 and 2 compare the boxplots of $(\widehat{\beta}_j - \beta_{j;o})$, $j = 1, \ldots, d$, based on the non-robust and robust-BD estimates, where the deviance loss and exponential loss are used as the BD in the top and bottom panels respectively. As seen from Figure 1 in the absence of contamination, both non-robust and robust methods perform comparably well. Besides, the bias in non-robust methods using the exponential loss (with $\text{p}_2(y;\theta)$ unbounded) is larger than that of the deviance loss (with $\text{p}_2(y;\theta)$ bounded). In the presence of contamination, Figure 2 reveals that the robust method is more effective in decreasing the estimation bias without excessively increasing the estimation variance.

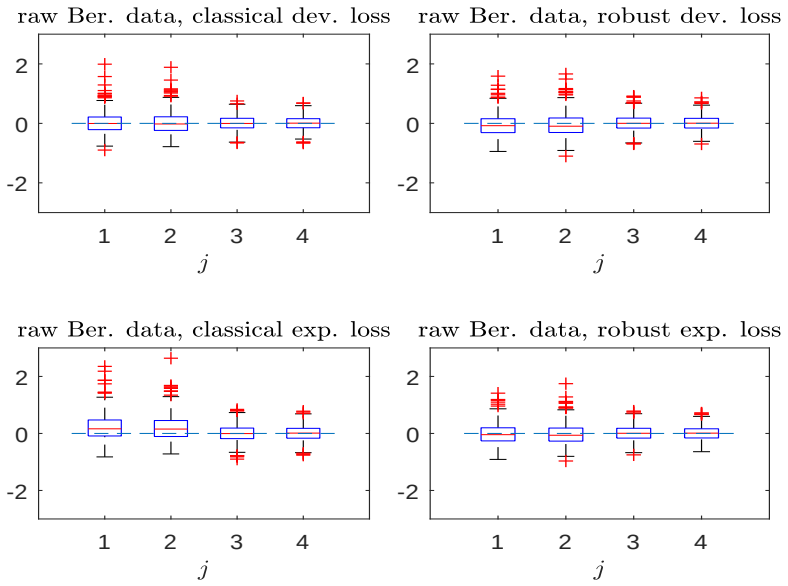

Figure 1. Simulated Bernoulli response data without contamination. Boxplots of $(\widehat{\beta}_j - \beta_{j;\mathrm{o}})$, $j = 1, \ldots, d$ (from left to right). (**Left panels**): non-robust method; (**right panels**): robust method.

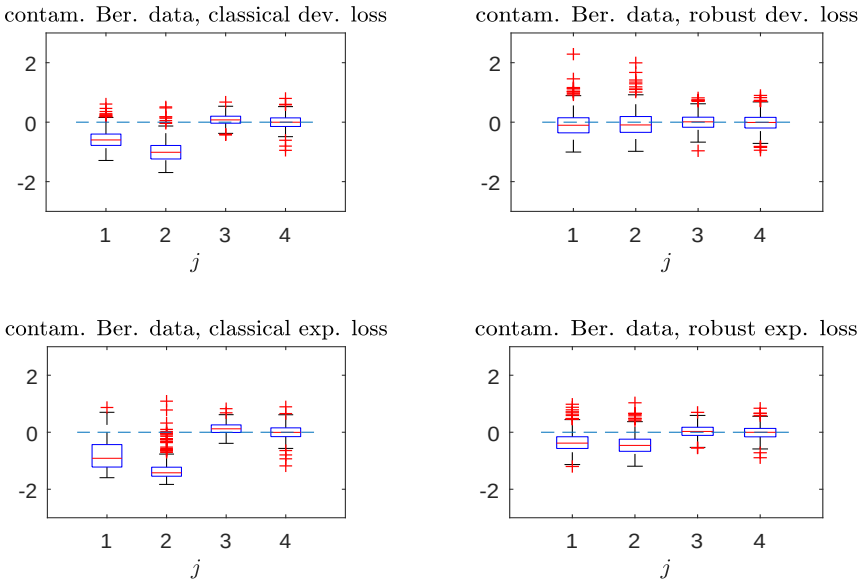

Figure 2. Simulated Bernoulli response data with contamination. The captions are identical to those in Figure 1.

For each replication, we calculate $\mathrm{MSE}(\widehat{\eta}) = n^{-1} \sum_{i=1}^{n} \{\widehat{\eta}_{\widehat{\beta}}(t_i) - \eta^{\mathrm{o}}(t_i)\}^2$. Figures 3 and 4 compare the plots of $\widehat{\eta}_{\widehat{\beta}}(t)$ from typical samples, using non-robust and robust-BD estimates, where the deviance loss and exponential loss are used as the BD in the top and bottom panels, respectively. There, the typical sample in each panel is selected in a way such that its MSE value corresponds to the 50-th percentile among the MSE-ranked values from 500 replications. These fitted curves reveal little difference between using the robust and non-robust methods, in the absence of contamination. For contaminated cases, robust estimates perform slightly better than non-robust estimates. Moreover, the boundary bias issue arising from the curve estimates at the edges using the local constant method can be ameliorated by using the local-linear method.

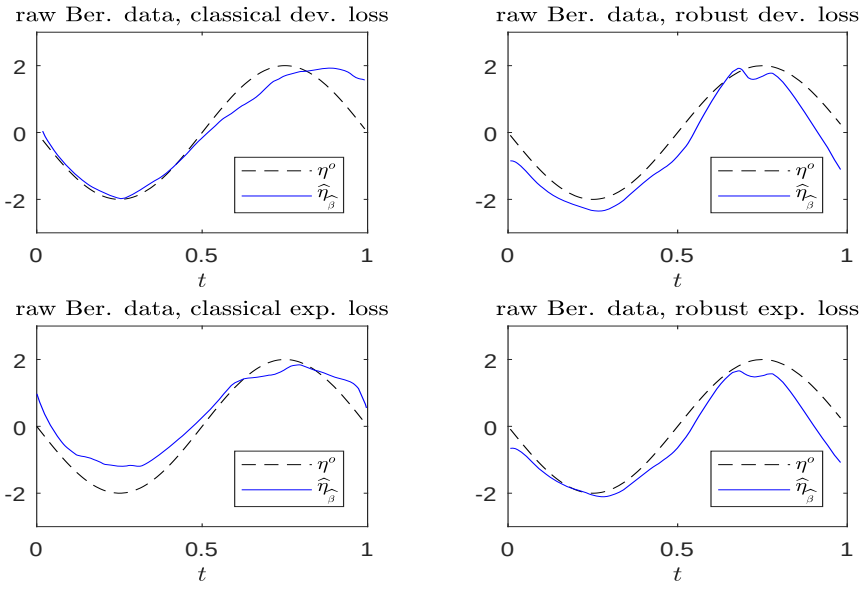

Figure 3. Simulated Bernoulli response data without contamination. Plots of $\eta^{\mathrm{o}}(t)$ and $\widehat{\eta}_{\widehat{\beta}}(t)$. (**Left panels**): non-robust method; (**right panels**): robust method.

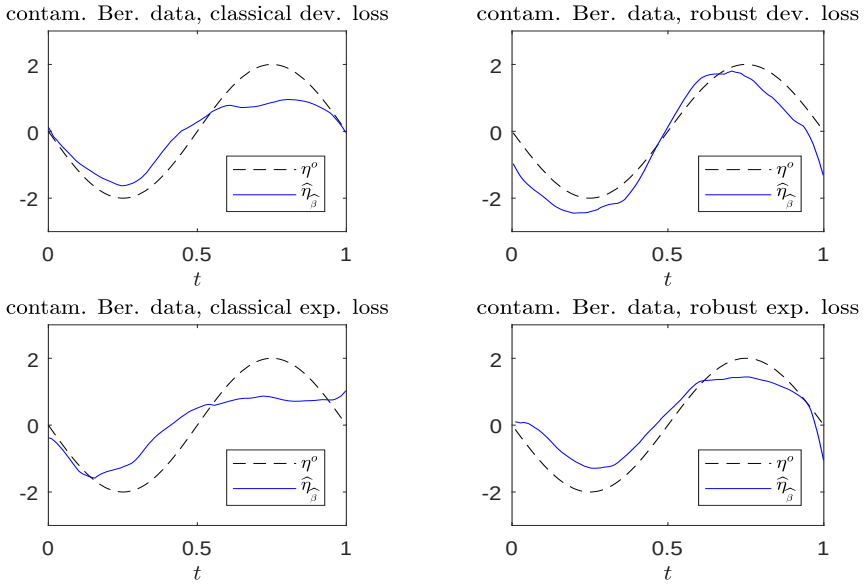

Figure 4. Simulated Bernoulli response data with contamination. Plots of $\eta^{\mathrm{o}}(t)$ and $\widehat{\eta}_{\widehat{\beta}}(t)$. (**Left panels**): non-robust method; (**right panels**): robust method.

6.2. Gaussian Responses

We generate independent observations $\{(\boldsymbol{X}_i, T_i, Y_i)\}_{i=1}^n$ from $(\boldsymbol{X}, T, Y)$ satisfying:

$$Y \mid (\boldsymbol{X}, T) \sim N(m(\boldsymbol{X}, T), \sigma^2), \quad (\boldsymbol{X}, \Phi^{-1}(T)) \sim N(\boldsymbol{0}, \Sigma),$$

where $\sigma = 1$, $\Sigma = (\sigma_{jk})$ with $\sigma_{jk} = 0.2^{|j-k|}$, Φ denotes the CDF of the standard normal distribution. The link function is $m(\boldsymbol{x}, t) = \boldsymbol{x}^T\boldsymbol{\beta}_{\mathrm{o}} + \eta^{\mathrm{o}}(t)$, where $\boldsymbol{\beta}_{\mathrm{o}} = (2, -2, 1, -1, 0, 0)^T$ and $\eta^{\mathrm{o}}(t) = 2\sin\{\pi(1+2t)\}$. The quadratic loss is utilized as the BD.

For each dataset simulated from the true model, a contaminated data-set is created, where 10 data points $(X_{i,j}, Y_i)$ are subject to contamination. They are replaced by $(X^*_{i,j}, Y^*_i)$, where $Y^*_i = Y_i \, \mathrm{I}\{|Y_i - m(\boldsymbol{X}_i, T_i)|/\sigma > 2\} + 15\,\mathrm{I}\{|Y_i - m(\boldsymbol{X}_i, T_i)|/\sigma \le 2\}$, $i = 1, \ldots, 10$,

$$\begin{array}{lll} X^*_{1,2} = 5\,\mathrm{sign}(U_1 - 0.5), & X^*_{2,2} = 5\,\mathrm{sign}(U_2 - 0.5), & X^*_{3,2} = 5\,\mathrm{sign}(U_3 - 0.5), \\ X^*_{4,4} = 5\,\mathrm{sign}(U_4 - 0.5), & X^*_{5,6} = 5\,\mathrm{sign}(U_5 - 0.5), & X^*_{6,1} = 5\,\mathrm{sign}(U_6 - 0.5), \\ X^*_{7,2} = 5\,\mathrm{sign}(U_7 - 0.5), & X^*_{8,3} = 5\,\mathrm{sign}(U_8 - 0.5), & X^*_{9,4} = 5\,\mathrm{sign}(U_9 - 0.5), \\ X^*_{10,5} = 5\,\mathrm{sign}(U_{10} - 0.5), & & \end{array}$$

with $\{U_i\} \overset{\text{i.i.d.}}{\sim} \mathrm{Uniform}(0, 1)$.

Figures 5 and 6 compare the boxplots of $(\widehat{\beta}_j - \beta_{j;\mathrm{o}})$, $j = 1, \ldots, d$, on the top panels, and plots of $\widehat{\eta}_{\widehat{\beta}}(t)$ from typical samples, on the bottom panels, using the non-robust and robust-BD estimates. The typical samples are selected similar to those in Section 6.1. The simulation results in Figure 5 indicate that the robust method performs, as well as the non-robust method for estimating both the parameter vector and non-parametric curve in non-contaminated cases. Figure 6 reveals that the robust estimates are less sensitive to outliers than the non-robust counterparts. Indeed, the non-robust method yields a conceivable bias for parametric estimation, and non-parametric estimation is worse than that of the robust method.

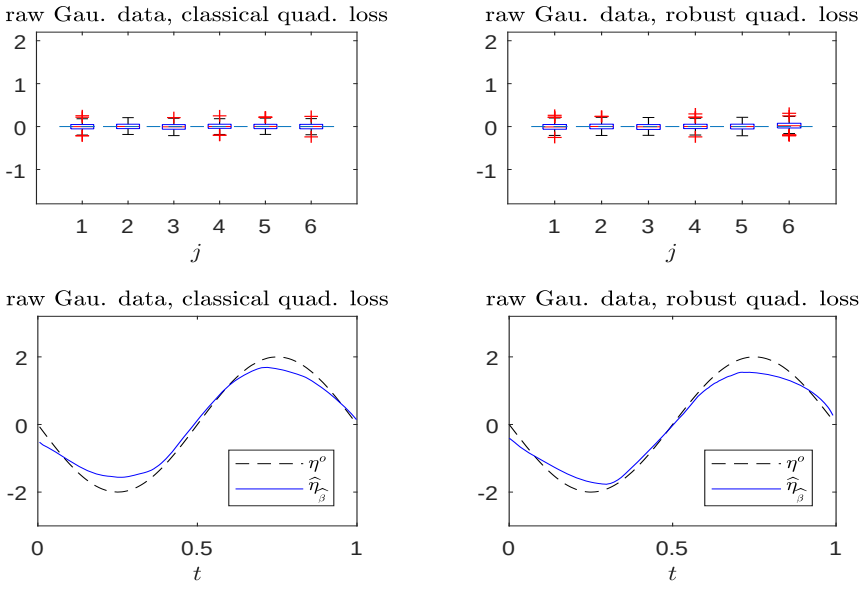

Figure 5. Simulated Gaussian response data without contamination. Top panels: boxplots of $(\widehat{\beta}_j - \beta_{j;\mathrm{o}})$, $j = 1, \ldots, d$ (from left to right). Bottom panels: plots of $\eta^{\mathrm{o}}(t)$ and $\widehat{\eta}_{\widehat{\beta}}(t)$. (**Left panels**): non-robust method; (**right panels**): robust method.

Figure 7 gives the QQ plots of the (first to 95-th) percentiles of the Wald-type statistic W_n versus those of the χ^2_2 distribution for testing the null hypothesis:

$$H_0 : \beta_{2;\mathrm{o}} = -2 \text{ and } \beta_{4;\mathrm{o}} = -1. \tag{32}$$

The plots depict that in both clean and contaminated cases, the robust W_n (in right panels) closely follows the χ^2_2 distribution, lending support to Theorem 3. On the other hand, the non-robust W_n agrees well with the χ^2_2 distribution in clean data; the presence of a small number of outlying data points severely distorts the sampling distribution of the non-robust W_n (in the bottom left panel) from the χ^2_2 distribution, yielding inaccurate levels of the test.

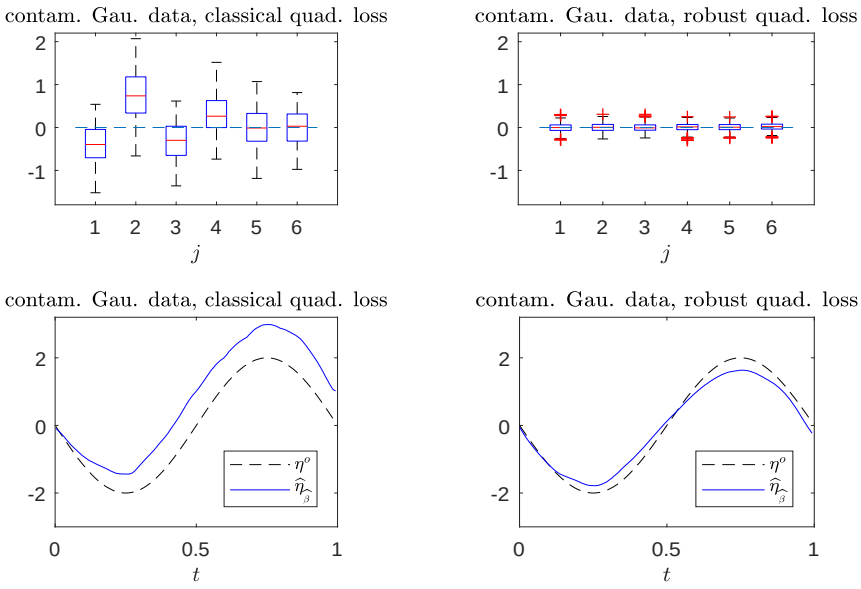

Figure 6. Simulated Gaussian response data with contamination. Top panels: boxplots of $(\widehat{\beta}_j - \beta_{j;o})$, $j = 1, \ldots, d$ (from left to right). Bottom panels: plots of $\eta^o(t)$ and $\widehat{\eta}_{\widehat{\beta}}(t)$. (**Left panels**): non-robust method; (**right panels**): robust method.

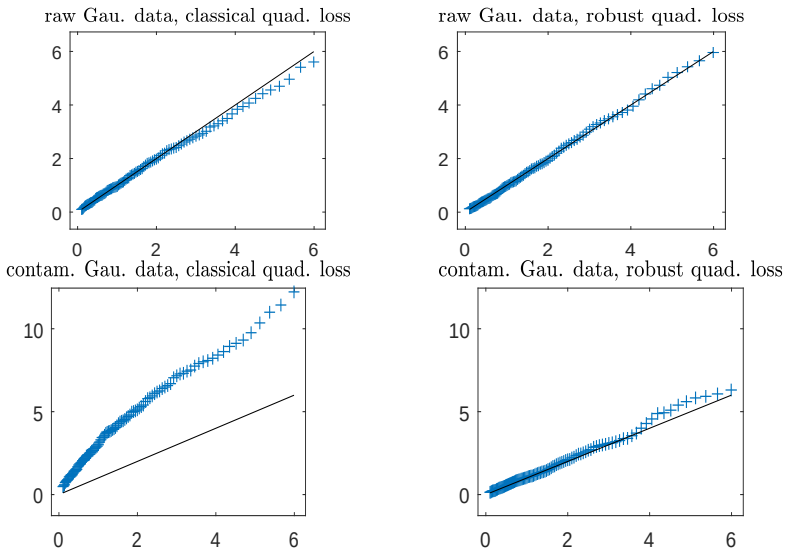

Figure 7. Simulated Gaussian response data with contamination. Empirical quantiles (on the y-axis) of the Wald-type statistics W_n versus quantiles (on the x-axis) of the χ^2 distribution. Solid line: the 45 degree reference line. (**Left panels**): non-robust method; (**right panels**): robust method.

To assess the stability of the power of the Wald-type test for testing the hypothesis (32), we evaluate the power in a sequence of alternatives with parameters $\boldsymbol{\beta}_o + \Delta \boldsymbol{c}$ for each given Δ, where $\boldsymbol{c} = \boldsymbol{\beta}_o + (1, \ldots, 1)^T$. Figure 8 plots the empirical rejection rates of the null model in the non-contaminated case and the contaminated case. The price to pay for the robust W_n is a little loss of power in the non-contaminated cases. However, under contamination, a very different behavior is observed. The observed power curve of the robust W_n is close to those attained in the non-contaminated case. On the contrary, the non-robust W_n is less informative, since its power curve is much lower than that of the robust W_n against the alternative hypotheses with $\Delta \neq 0$, but higher than the nominal level at the null hypothesis with $\Delta = 0$.

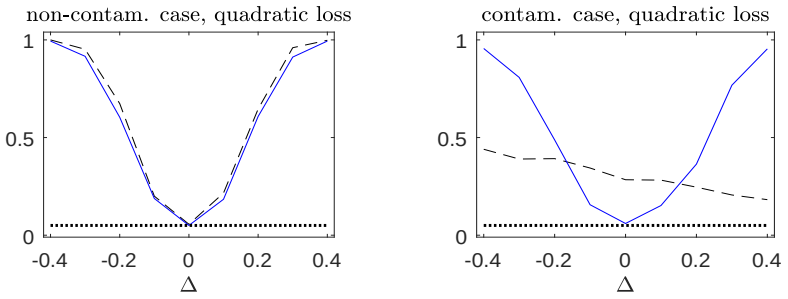

Figure 8. Observed power curves of tests for the Gaussian response data. The dashed line corresponds to the non-robust Wald-type test W_n; the solid line corresponds to the robust W_n; the dotted line indicates the 5% nominal level. (**Left panels**): non-contaminated case; (**right panels**): contaminated case.

7. Real Data Analysis

Two real datasets are analyzed. In both cases, the quadratic loss is set to be the BD, and the nonparametric function is fitted via local-linear regression method, where the bandwidth parameter is chosen to be 25% of the interval length of the variable T. Choices of the Huber ψ-function and weight functions are identical to those in Section 6.

7.1. Example 1

The dataset studied in [19] consists of 2447 observations on three variables, $\log(\mathsf{wage})$, age and education, for women. It is of interest to learn how wages change with years of age and years of education. It is anticipated to find an increasing regression function of $Y = \log(\mathsf{wage})$ in $T = \mathsf{age}$ as well as in $X_1 = \mathsf{education}$. We fit a partially linear model $Y = \eta(T) + \beta_1 X_1 + \epsilon$. Profiles of the fitted nonparametric functions $\widehat{\eta}(\cdot)$ in Figure 9 indeed exhibit the overall upward trend in age. The coefficient estimate is $\widehat{\beta}_1 = 0.0809$ with standard error 0.0042 using the non-robust method, and is $\widehat{\beta}_1 = 0.1334$ with standard error 0.0046 by means of the robust method. It is seen that robust estimates are similar to the non-robust counterparts. Our evaluation, based on both the non-robust and robust methods, supports the predicted result in theoretical and empirical literature in socio-economical studies.

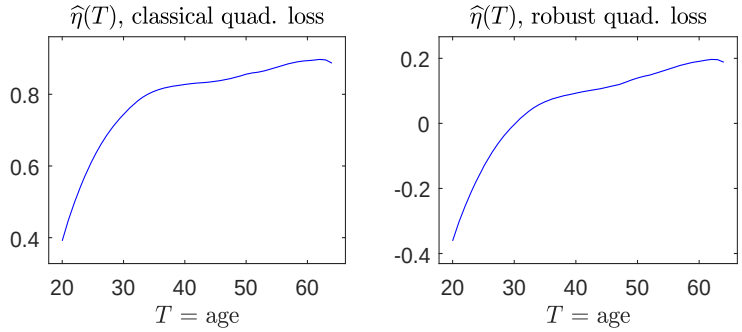

Figure 9. The dataset in [19]. (**Left panels**): estimate of $\eta(T)$ via the non-robust quadratic loss; (**right panels**): estimate of $\eta(T)$ via the robust quadratic loss.

7.2. Example 2

We analyze an employee dataset (Example 11.3 of [20]) of the Fifth National Bank of Springfield, based on year 1995 data. The bank, whose name has been changed, was charged in court with that its female employees received substantially smaller salaries than its male employees. For each of its 208 employees, the dataset consists of seven variables, EducLev (education level), JobGrade (job grade), YrHired (year that an employee was hired), YrBorn (year that an employee was born), Female

(indicator of being female), YrsPrior (years of work experience at another bank before working at the Fifth National bank), and Salary (current annual salary in thousands of dollars).

To explain variation in salary, we fit a partial linear model, $Y = \eta(T) + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$, for $Y = \log(\text{Salary})$, $T = \text{Age}$, $X_1 = \text{Female}$, $X_2 = \text{YrHired}$, $X_3 = \text{EducLev}$, $X_4 = \text{JobGrade}$ and $X_5 = \text{YrsPrior}$, where $\text{Age} = 95 - \text{YrBorn}$ is age. Table 1 presents parameter estimates and their standard errors (given within brackets), along with p-values calculated from the Wald-type test W_n. Figure 10 depicts the estimated nonparametric functions.

It is interesting to note that for this dataset, results from using the robust and non-robust methods make a difference in drawing conclusions. For example, from Table 1, the non-robust method gives the estimate of parameter β_1 for gender to be below zero, which may be interpreted as the evidence of discrimination against female employees in salary and lends support to the plaintiff. In contrast, the robust method yields $\widehat{\beta}_1 > 0$, which does not indicate that gender has an adverse effect. (A similar conclusion made from penalized-likelihood was obtained in Section 4.1 of [21]). Moreover, the estimated nonparametric functions $\widehat{\eta}(\cdot)$ obtained from non-robust and robust methods are qualitatively different: the former method does not deliver a monotone increasing pattern with Age, whereas the latter method does. Whether or not the difference was caused by outlying observations will be an interesting issue to be investigated.

Table 1. Parameter estimates and p-values for partially linear model of the dataset in [20]

Variable	Classical-BD Estimation			Robust-BD Estimation		
	Estimate (s.e.)		p-Value of W_n	Estimate (s.e.)		p-Value of W_n
Female	−0.0491	(0.0232)	0.0339	0.0530	(0.0323)	0.1010
YrHired	−0.0093	(0.0026)	0.0005	0.0359	(0.0086)	0.0000
EducLev	0.0179	(0.0079)	0.0228	−0.0133	(0.0131)	0.3103
JobGrade	0.0899	(0.0075)	0.0000	0.1672	(0.0168)	0.0000
YrsPrior	0.0033	(0.0023)	0.1528	−0.0050	(0.0061)	0.4104

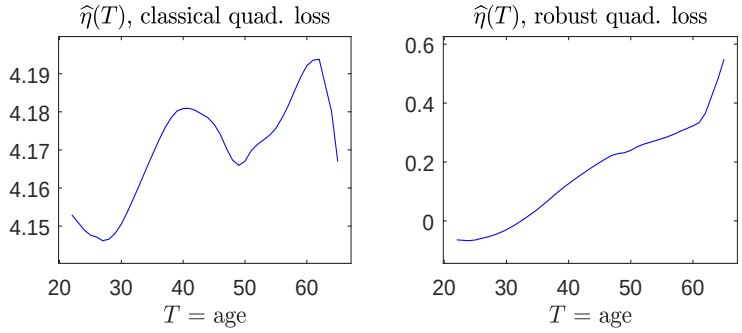

Figure 10. The dataset in [20]. (**Left panel**): estimate of $\eta(T)$ via the non-robust quadratic loss; (**right panel**): estimate of $\eta(T)$ via the robust quadratic loss.

8. Discussion

Over the past two decades, nonparametric inference procedures for testing hypotheses concerning nonparametric regression functions have been developed extensively. See [22–26] and the references therein. The work on the generalized likelihood ratio test [24] offers light into nonparametric inference, based on function estimation under nonparametric models, using the quadratic loss function as the error measure. These works do not directly deal with the robust procedure. Exploring the inference on nonparametric functions, such as $\eta^{\text{o}}(t)$ in GPLM associated with a scalar variable T and the additive structure $\sum_{d=1}^{D} \eta_d^{\text{o}}(t_d)$ as in [27] with a vector variable $\boldsymbol{T} = (T_1, \ldots, T_D)$, estimated via the "robust-BD" as the error measure, when there are possible outlying data points, will be the future work.

This paper utilizes the class BD of loss functions, the optimal choice of which depends on specific settings and criteria. For e.g., regression and classification will utilize different loss functions, and thus further study on optimality is desirable.

Some recent work on partially linear models in econometrics includes [28–30]. There, the nonparametric function is approximated via linear expansions, with the number of coefficients diverging with n. Developing inference procedures to be resistant to outliers could be of interest.

Acknowledgments: The authors thank the two referees for insightful comments and suggestions. The research is supported by the U.S. NSF Grants DMS–1712418, DMS–1505367, CMMI–1536978, DMS–1308872, the Wisconsin Alumni Research Foundation and the National Natural Science Foundation of China, grants 11690014.

Author Contributions: C.Z. conceived and designed the experiments; C.Z. analyzed the data; Z.Z. contributed to discussions and analysis tools; C.Z. wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proofs of Main Results

Throughout the proof, C represents a generic finite constant. We impose some regularity conditions, which may not be the weakest, but facilitate the technical derivations.

Notation:

For integers $j \geq 0$, $\mu_j(K) = \int u^j K(u)du$; $\boldsymbol{c}_p = (\mu_{p+1}(K), \ldots, \mu_{2p+1}(K))^T$; $\mathcal{S} = (\mu_{j+k-2}(K))_{1\leq j,k\leq p+1}$. Define: $\eta(\boldsymbol{x},t) = F(m(\boldsymbol{x},t)) = \boldsymbol{x}^T\boldsymbol{\beta}_{\mathrm{o}} + \eta^{\mathrm{o}}(t)$; $\eta_i = \eta(\boldsymbol{X}_i, T_i)$. Set $\eta_i(t;\boldsymbol{\beta}) = \boldsymbol{X}_i^T\boldsymbol{\beta} + \eta_{\boldsymbol{\beta}}(t) + \sum_{k=1}^{p}(T_i - t)^k\eta_{\boldsymbol{\beta}}^{(k)}(t)/k!$; $g_1(\tau;t,\boldsymbol{\beta}) = \mathrm{E}\{\mathrm{p}_1(Y_i;\eta_i(t;\boldsymbol{\beta}))w_1(\boldsymbol{X}_i) \mid T_i = \tau\}$; $g_2(\tau;t,\boldsymbol{\beta}) = \mathrm{E}\{\mathrm{p}_2(Y_i;\eta_i(t;\boldsymbol{\beta}))w_1(\boldsymbol{X}_i) \mid T_i = \tau\}$.

Condition A:

- A1. $\eta_{\boldsymbol{\beta}}(t)$ is the unique minimizer of $S(a;t,\boldsymbol{\beta})$ with respect to $a \in \mathbb{R}^1$.
- A2. $\boldsymbol{\beta}_{\mathrm{o}} \in \mathbb{R}^d$ is the unique minimizer of $J(\boldsymbol{\beta}, \eta_{\boldsymbol{\beta}})$ with respect to $\boldsymbol{\beta}$, where $d \geq 1$.
- A3. $\eta^{\mathrm{o}}(\cdot) = \eta_{\boldsymbol{\beta}_{\mathrm{o}}}(\cdot)$.

Condition B:

- B1. The function $\rho_q(y,\mu)$ is continuous and bounded. The functions $\mathrm{p}_1(y;\theta)$, $\mathrm{p}_2(y;\theta)$, $\mathrm{p}_3(y;\theta)$, $w_1(\cdot)$ and $w_2(\cdot)$ are bounded; $\mathrm{p}_2(y;\theta)$ is continuous in θ.
- B2. The kernel function K is Lipschitz continuous, a symmetric probability density function with bounded support. The matrix $\mathcal{S}$ is positive definite.
- B3. The marginal density $f_T(t)$ of T is a continuous function, uniformly bounded away from zero and ∞ for $t \in \mathcal{T}_0$.
- B4. The function $S(a;t,\boldsymbol{\beta})$ is continuous and $\eta_{\boldsymbol{\beta}}(t)$ is a continuous function of $(t,\boldsymbol{\beta})$.
- B5. Assume $g_2(\tau;t,\boldsymbol{\beta})$ is continuous in τ; $g_2(t;t,\boldsymbol{\beta})$ is continuous in $t \in \mathcal{T}_0$.
- B6. Functions $\eta_{\boldsymbol{\beta}}(t)$ and $\eta^{\mathrm{o}}(t)$ are $(p+1)$-times continuously differentiable at t.
- B7. The link function $F(\cdot)$ is monotone increasing and a bijection, $F^{(3)}(\cdot)$ is continuous, and $F^{(1)}(\cdot) > 0$. The matrix $\mathrm{var}(\boldsymbol{X} \mid T = t)$ is positive definite for a.e. t.
- B8. The matrix $\mathbf{H}_0$ in (25) is invertible; Ω_0^* in (26) is positive-definite.
- B9. $\widehat{\eta}_{\boldsymbol{\beta}}(t)$ and $\eta_{\boldsymbol{\beta}}(t)$ are continuously differentiable with respect to $(t,\boldsymbol{\beta})$, and twice continuously differentiable with respect to $\boldsymbol{\beta}$ such that for any $1 \leq j,k \leq d$, $\frac{\partial^2}{\partial\beta_j\partial\beta_k}\eta_{\boldsymbol{\beta}}(t)|_{\boldsymbol{\beta}=\boldsymbol{\beta}_{\mathrm{o}}}$ is bounded. Furthermore, for any $1 \leq j,k \leq d$, $\frac{\partial^2}{\partial\beta_j\partial\beta_k}\eta_{\boldsymbol{\beta}}(t)$ satisfies the equicontinuity condition:

$$\forall \varepsilon > 0, \exists\, \delta_\varepsilon > 0 : \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{\mathrm{o}}\| < \delta_\varepsilon \Longrightarrow \Big\| \frac{\partial^2}{\partial\beta_j\partial\beta_k}\eta_{\boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_1} - \frac{\partial^2}{\partial\beta_j\partial\beta_k}\eta_{\boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_{\mathrm{o}}} \Big\|_\infty < \varepsilon.$$

Note that Conditions A, B2–B5 and B8–B9 were similarly used in [9]. Conditions B1 and B7 follow [10]. Condition B6 is due to the local p-th-degree polynomial regression estimation.

Proof of Lemma 1: From Condition A1, we obtain $\mathrm{E}\{\mathrm{p}_1(Y; \boldsymbol{X}^T\boldsymbol{\beta} + \eta_\beta(t))w_1(\boldsymbol{X}) \mid T = t\} = 0$ and $\mathrm{E}\{\mathrm{p}_2(Y; \boldsymbol{X}^T\boldsymbol{\beta} + \eta_\beta(t))w_1(\boldsymbol{X}) \mid T = t\} > 0$, i.e.,

$$g_1(t; t, \boldsymbol{\beta}) = \mathrm{E}\{\mathrm{p}_1(Y; \boldsymbol{X}^T\boldsymbol{\beta} + \eta_\beta(t))w_1(\boldsymbol{X}) \mid T = t\} = 0, \quad \text{(A1)}$$

$$g_2(t; t, \boldsymbol{\beta}) = \mathrm{E}\{\mathrm{p}_2(Y; \boldsymbol{X}^T\boldsymbol{\beta} + \eta_\beta(t))w_1(\boldsymbol{X}) \mid T = t\} > 0. \quad \text{(A2)}$$

Define by $\boldsymbol{\eta}_\beta^{(0,\ldots,p)}(t) = (\eta_\beta(t), \eta_\beta^{(1)}(t), \ldots, \eta_\beta^{(p)}(t)/p!)^T$ the vector of $\eta_\beta(t)$ along with re-scaled derivatives with respect to t up to the order p. Note that:

$$\begin{aligned}
\eta_i(t; \boldsymbol{\beta}) &= \boldsymbol{X}_i^T\boldsymbol{\beta} + \sum_{k=0}^{p}(T_i - t)^k \frac{\eta_\beta^{(k)}(t)}{k!} \\
&= \boldsymbol{X}_i^T\boldsymbol{\beta} + \mathbf{t}_i(t)^T\boldsymbol{\eta}_\beta^{(0,\ldots,p)}(t) \\
&= \boldsymbol{X}_i^T\boldsymbol{\beta} + \{H^{-1}\mathbf{t}_i(t)\}^T H\boldsymbol{\eta}_\beta^{(0,\ldots,p)}(t) \\
&= \boldsymbol{X}_i^T\boldsymbol{\beta} + \mathbf{t}_i^*(t)^T H\boldsymbol{\eta}_\beta^{(0,\ldots,p)}(t),
\end{aligned}$$

where $H = \mathrm{diag}\{(1, h, \ldots, h^p)\}$ and $\mathbf{t}_i^*(t) = H^{-1}\mathbf{t}_i(t) = (1, (T_i - t)/h, \ldots, (T_i - t)^p/h^p)^T$ denotes the re-scaled $\mathbf{t}_i(t)$. Then:

$$\begin{aligned}
& \boldsymbol{X}_i^T\boldsymbol{\beta} + \mathbf{t}_i(t)^T\boldsymbol{a} \\
= \; & \boldsymbol{X}_i^T\boldsymbol{\beta} + \mathbf{t}_i^*(t)^T H\boldsymbol{a} \\
= \; & \boldsymbol{X}_i^T\boldsymbol{\beta} + \mathbf{t}_i^*(t)^T H\boldsymbol{\eta}_\beta^{(0,\ldots,p)}(t) + \mathbf{t}_i^*(t)^T H\{\boldsymbol{a} - \boldsymbol{\eta}_\beta^{(0,\ldots,p)}(t)\} \\
= \; & \eta_i(t; \boldsymbol{\beta}) + \mathbf{t}_i^*(t)^T H\{\boldsymbol{a} - \boldsymbol{\eta}_\beta^{(0,\ldots,p)}(t)\}.
\end{aligned}$$

Hence, we rewrite (16) as:

$$S_n(\boldsymbol{a}; t, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\rho_q(Y_i, F^{-1}(\eta_i(t; \boldsymbol{\beta}) + \mathbf{t}_i^*(t)^T H\{\boldsymbol{a} - \boldsymbol{\eta}_\beta^{(0,\ldots,p)}(t)\}))\, w_1(\boldsymbol{X}_i)\, K_h(T_i - t).$$

Therefore, $\widehat{\boldsymbol{a}}(t, \boldsymbol{\beta})$ minimizing $S_n(\boldsymbol{a}; t, \boldsymbol{\beta})$ is equivalent to the one minimizing:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\Big\{&\rho_q(Y_i, F^{-1}(\eta_i(t; \boldsymbol{\beta}) + \mathbf{t}_i^*(t)^T H\{\boldsymbol{a} - \boldsymbol{\eta}_\beta^{(0,\ldots,p)}(t)\})) \\
&-\rho_q(Y_i, F^{-1}(\eta_i(t; \boldsymbol{\beta})))\Big\} w_1(\boldsymbol{X}_i)\, K_h(T_i - t)
\end{aligned}$$

with respect to $\boldsymbol{a}$. It follows that $\widehat{\boldsymbol{a}}^*(t, \boldsymbol{\beta})$, defined by $\widehat{\boldsymbol{a}}^*(t, \boldsymbol{\beta}) = \sqrt{nh}\, H\{\widehat{\boldsymbol{a}}(t, \boldsymbol{\beta}) - \boldsymbol{\eta}_\beta^{(0,\ldots,p)}(t)\}$, minimizes:

$$\begin{aligned}
G_n(\boldsymbol{a}^*; t, \boldsymbol{\beta}) = nh\Bigg[&\frac{1}{n}\sum_{i=1}^{n}\Big\{\rho_q(Y_i, F^{-1}(\eta_i(t; \boldsymbol{\beta}) + \{a_n\mathbf{t}_i^*(t)^T\boldsymbol{a}^*\})) - \rho_q(Y_i, F^{-1}(\eta_i(t; \boldsymbol{\beta})))\Big\} \\
&w_1(\boldsymbol{X}_i)\, K_h(T_i - t)\Bigg]
\end{aligned}$$

with respect to $\boldsymbol{a}^* \in \mathbb{R}^{p+1}$, where $a_n = 1/\sqrt{nh}$. Note that for any fixed $\boldsymbol{a}^*$, $|\mathbf{t}_i^*(t)^T\boldsymbol{a}^*| \leq C$. By Taylor expansion,

$$\begin{aligned}
G_n(\boldsymbol{a}^*; t, \boldsymbol{\beta}) = \; & nh\Bigg(a_n\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{p}_1(Y_i; \eta_i(t; \boldsymbol{\beta}))\{\mathbf{t}_i^*(t)^T\boldsymbol{a}^*\}w_1(\boldsymbol{X}_i)\, K_h(T_i - t)\right] \\
&+a_n^2\frac{1}{2}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{p}_2(Y_i; \eta_i(t; \boldsymbol{\beta}))\{\mathbf{t}_i^*(t)^T\boldsymbol{a}^*\}^2 w_1(\boldsymbol{X}_i)\, K_h(T_i - t)\right]
\end{aligned}$$

$$
\begin{aligned}
& + a_n^3 \frac{1}{6}\left[\frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_3(Y_i;\eta_i^*(t;\boldsymbol{\beta}))\{\mathbf{t}_i^*(t)^T\boldsymbol{a}^*\}^3 w_1(\boldsymbol{X}_i)\, K_h(T_i-t)\right]\Big) \\
= \;& I_{n,1} + I_{n,2} + I_{n,3},
\end{aligned}
$$

where $\eta_i^*(t;\boldsymbol{\beta})$ is located between $\eta_i(t;\boldsymbol{\beta})$ and $\eta_i(t;\boldsymbol{\beta}) + \{a_n \mathbf{t}_i^*(t)^T \boldsymbol{a}^*\}$. We notice that:

$$
I_{n,1} \equiv \sqrt{nh}\, \boldsymbol{W}_n(t,\boldsymbol{\beta})^T \boldsymbol{a}^*,
$$

where:

$$
\boldsymbol{W}_n(t,\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_1(Y_i;\eta_i(t;\boldsymbol{\beta}))\mathbf{t}_i^*(t) w_1(\boldsymbol{X}_i)\, K_h(T_i-t);
$$

also, Lemma A1 implies:

$$
\begin{aligned}
I_{n,2} &= nha_n^2 \frac{1}{2}\boldsymbol{a}^{*T}\left[\frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_2(Y_i;\eta_i(t;\boldsymbol{\beta}))\{\mathbf{t}_i^*(t)\mathbf{t}_i^*(t)^T\} w_1(\boldsymbol{X}_i)\, K_h(T_i-t)\right]\boldsymbol{a}^* \\
&= \frac{1}{2}\boldsymbol{a}^{*T}\mathbf{S}_2(t,\boldsymbol{\beta})\boldsymbol{a}^* + o_{\mathrm{P}}(1),
\end{aligned}
$$

where:

$$
\mathbf{S}_2(t,\boldsymbol{\beta}) = g_2(t;t,\boldsymbol{\beta}) f_T(t)\mathcal{S} \succ \mathbf{0}
$$

by (A2), Condition B2 and B5; and (by using $X_n = O_{\mathrm{P}}(\mathrm{E}(|X_n|))$):

$$
I_{n,3} \leq C\, O_{\mathrm{P}}(nha_n^3) = O_{\mathrm{P}}(1/\sqrt{nh}) = o_{\mathrm{P}}(1).
$$

Then:

$$
G_n(\boldsymbol{a}^*;t,\boldsymbol{\beta}) = \sqrt{nh}\, \boldsymbol{W}_n(t,\boldsymbol{\beta})^T\boldsymbol{a}^* + \frac{1}{2}\boldsymbol{a}^{*T}\mathbf{S}_2(t,\boldsymbol{\beta})\boldsymbol{a}^* + o_{\mathrm{P}}(1),
$$

where $\boldsymbol{a}^{*T}\mathbf{S}_2(t,\boldsymbol{\beta})\boldsymbol{a}^* = (\boldsymbol{a}^{*T}\mathcal{S}\boldsymbol{a}^*) g_2(t;t,\boldsymbol{\beta}) f_T(t)$ is continuous in $t \in \mathcal{T}_0$ by B3 and B5.

We now examine $\boldsymbol{W}_n(t,\boldsymbol{\beta})$. Note that:

$$
\begin{aligned}
\mathrm{var}\{\boldsymbol{W}_n(t,\boldsymbol{\beta})\} &= \frac{1}{n}\mathrm{var}\{\mathrm{p}_1(Y_i;\eta_i(t;\boldsymbol{\beta}))\mathbf{t}_i^*(t) w_1(\boldsymbol{X}_i)\, K_h(T_i-t)\} \\
&\leq \frac{1}{n}\mathrm{E}\Big[\mathrm{p}_1^2(Y_i;\eta_i(t;\boldsymbol{\beta}))\{\mathbf{t}_i^*(t)\mathbf{t}_i^*(t)^T\} w_1^2(\boldsymbol{X}_i)\,\{K_h(T_i-t)\}^2\Big] \\
&\leq \frac{C}{n}\mathrm{E}\Big[\frac{1}{h^2}\Big\{K\Big(\frac{T_i-t}{h}\Big)\Big\}^2\Big] \\
&= \frac{C}{nh}.
\end{aligned}
$$

To evaluate $\mathrm{E}\{\boldsymbol{W}_n(t,\boldsymbol{\beta})\}$, it is easy to see that for each $j \in \{0,1,\ldots,p\}$,

$$
\begin{aligned}
\boldsymbol{e}_{j+1,p+1}^T \mathrm{E}\{\boldsymbol{W}_n(t,\boldsymbol{\beta})\} &= \mathrm{E}\{\mathrm{p}_1(Y_i;\eta_i(t;\boldsymbol{\beta}))\boldsymbol{e}_{j+1,p+1}^T\mathbf{t}_i^*(t) w_1(\boldsymbol{X}_i)\, K_h(T_i-t)\} \\
&= \mathrm{E}\Big\{\mathrm{p}_1(Y_i;\eta_i(t;\boldsymbol{\beta}))\Big(\frac{T_i-t}{h}\Big)^j w_1(\boldsymbol{X}_i)\, K_h(T_i-t)\Big\} \\
&= \mathrm{E}\Big[\mathrm{E}\{\mathrm{p}_1(Y_i;\eta_i(t;\boldsymbol{\beta}))w_1(\boldsymbol{X}_i) \mid T_i\}\Big(\frac{T_i-t}{h}\Big)^j K_h(T_i-t)\Big] \\
&= \mathrm{E}\Big\{g_1(T_i;t,\boldsymbol{\beta})\Big(\frac{T_i-t}{h}\Big)^j K_h(T_i-t)\Big\} \\
&= \int g_1(y;t,\boldsymbol{\beta})\Big(\frac{y-t}{h}\Big)^j \frac{1}{h}K\Big(\frac{y-t}{h}\Big) f_T(y)dy \\
&= \int g_1(t+hx;t,\boldsymbol{\beta}) x^j K(x) f_T(t+hx)dx.
\end{aligned}
$$

Note that by Taylor expansion,

$$\eta_\beta(t+hx) = \sum_{k=0}^{p}(hx)^k\frac{\eta_\beta^{(k)}(t)}{k!} + (hx)^{p+1}\frac{\eta_\beta^{(p+1)}(t)}{(p+1)!} + o(h^{p+1}).$$

This combined with the facts (A1) and (A2) give that:

$$\begin{aligned}
& g_1(t+hx;t,\boldsymbol{\beta}) \\
= \;& \mathrm{E}\left\{\mathrm{p}_1\left(Y;\boldsymbol{X}^T\boldsymbol{\beta} + \sum_{k=0}^{p}(hx)^k\frac{\eta_\beta^{(k)}(t)}{k!}\right)w_1(\boldsymbol{X}) \;\middle|\; T = t+hx\right\} \\
= \;& \mathrm{E}\Bigg[\mathrm{p}_1(Y;\boldsymbol{X}^T\boldsymbol{\beta}+\eta_\beta(t+hx))w_1(\boldsymbol{X}) \\
& +\mathrm{p}_2(Y;\boldsymbol{X}^T\boldsymbol{\beta}+\eta_\beta(t+hx))\left\{\sum_{k=0}^{p}(hx)^k\frac{\eta_\beta^{(k)}(t)}{k!} - \eta_\beta(t+hx)\right\}w_1(\boldsymbol{X}) \;\middle|\; T = t+hx\Bigg] \\
& +o(h^{p+1}) \\
= \;& g_1(t+hx;t+hx,\boldsymbol{\beta}) - (hx)^{p+1}\frac{\eta_\beta^{(p+1)}(t)}{(p+1)!}g_2(t+hx;t+hx,\boldsymbol{\beta}) + o(h^{p+1}) \\
= \;& -(hx)^{p+1}\frac{\eta_\beta^{(p+1)}(t)}{(p+1)!}g_2(t+hx;t+hx,\boldsymbol{\beta}) + o(h^{p+1}).
\end{aligned}$$

Thus, using the continuity of $g_2(t;t,\boldsymbol{\beta})$ and $f_T(t)$ in t, we obtain:

$$\mathrm{E}\{\boldsymbol{W}_n(t,\boldsymbol{\beta})\} = -\boldsymbol{c}_p\frac{\eta_\beta^{(p+1)}(t)}{(p+1)!}g_2(t;t,\boldsymbol{\beta})f_T(t)h^{p+1} + o(h^{p+1})$$

uniformly in $(t,\boldsymbol{\beta})$. Thus, we conclude that $\sqrt{nh}\,\boldsymbol{W}_n(t,\boldsymbol{\beta}) = O_{\mathrm{P}}(1)$ when $nh^{2p+3} = O(1)$.

By Lemma A2,

$$\sup_{\boldsymbol{a}^*\in\Theta,\,t\in\mathcal{T}_0,\,\boldsymbol{\beta}\in\mathcal{K}}\left|G_n(\boldsymbol{a}^*;t,\boldsymbol{\beta}) - \sqrt{nh}\,\boldsymbol{W}_n(t,\boldsymbol{\beta})^T\boldsymbol{a}^* - \frac{1}{2}\boldsymbol{a}^{*T}\mathbf{S}_2(t,\boldsymbol{\beta})\boldsymbol{a}^*\right| = o_{\mathrm{P}}(1).$$

This along with Lemma A.1 of [18] yields:

$$\sup_{t\in\mathcal{T}_0,\,\boldsymbol{\beta}\in\mathcal{K}}\|\widehat{\boldsymbol{a}}^*(t,\boldsymbol{\beta}) + \{\mathbf{S}_2(t,\boldsymbol{\beta})\}^{-1}\sqrt{nh}\,\boldsymbol{W}_n(t,\boldsymbol{\beta})\| = o_{\mathrm{P}}(1),$$

the first entry of which satisfies:

$$\sup_{t\in\mathcal{T}_0,\,\boldsymbol{\beta}\in\mathcal{K}}|\sqrt{nh}\{\widehat{\eta}_\beta(t) - \eta_\beta(t)\} + \boldsymbol{e}_{1,p+1}^T\{\mathbf{S}_2(t,\boldsymbol{\beta})\}^{-1}\sqrt{nh}\,\boldsymbol{W}_n(t,\boldsymbol{\beta})| = o_{\mathrm{P}}(1),$$

namely, $\sup_{t\in\mathcal{T}_0,\boldsymbol{\beta}\in\mathcal{K}}|\widehat{\eta}_\beta(t) - \eta_\beta(t) + \boldsymbol{e}_{1,p+1}^T\{\mathbf{S}_2(t,\boldsymbol{\beta})\}^{-1}\boldsymbol{W}_n(t,\boldsymbol{\beta})| = o_{\mathrm{P}}(1/\sqrt{nh})$. By [31], $\sup_{t\in\mathcal{T}_0,\boldsymbol{\beta}\in\mathcal{K}}\|\boldsymbol{W}_n(t,\boldsymbol{\beta}) - \mathrm{E}\{\boldsymbol{W}_n(t,\boldsymbol{\beta})\}\| = O_{\mathrm{P}}(\{\frac{\log(1/h)}{nh}\}^{1/2})$. Furthermore,

$$\{\mathbf{S}_2(t,\boldsymbol{\beta})\}^{-1}\mathrm{E}\{\boldsymbol{W}_n(t,\boldsymbol{\beta})\} = -\mathcal{S}^{-1}\boldsymbol{c}_p\frac{\eta_\beta^{(p+1)}(t)}{(p+1)!}h^{p+1} + o(h^{p+1})$$

uniformly in $(t,\boldsymbol{\beta})$. Therefore,

$$\sup_{t\in\mathcal{T}_0,\,\boldsymbol{\beta}\in\mathcal{K}}\left|\widehat{\eta}_\beta(t) - \eta_\beta(t) - \boldsymbol{e}_{1,p+1}^T\mathcal{S}^{-1}\boldsymbol{c}_p\frac{\eta_\beta^{(p+1)}(t)}{(p+1)!}h^{p+1}\right| = o_{\mathrm{P}}(1).$$

This yields:

$$\begin{aligned} & \sup_{\boldsymbol{\beta}\in\mathcal{K}}\sup_{t\in\mathcal{T}_0}\Big|\widehat{\eta}_{\boldsymbol{\beta}}(t)-\eta_{\boldsymbol{\beta}}(t)-\boldsymbol{e}_{1,p+1}^{T}\mathcal{S}^{-1}\boldsymbol{c}_p\frac{\eta_{\boldsymbol{\beta}}^{(p+1)}(t)}{(p+1)!}h^{p+1}\Big| \\ \leq & \sup_{t\in\mathcal{T}_0,\boldsymbol{\beta}\in\mathcal{K}}\Big|\widehat{\eta}_{\boldsymbol{\beta}}(t)-\eta_{\boldsymbol{\beta}}(t)-\boldsymbol{e}_{1,p+1}^{T}\mathcal{S}^{-1}\boldsymbol{c}_p\frac{\eta_{\boldsymbol{\beta}}^{(p+1)}(t)}{(p+1)!}h^{p+1}\Big| = o_{\mathrm{P}}(1). \end{aligned}$$

Note that for $p=1$, $\boldsymbol{e}_{1,p+1}^{T}\mathcal{S}^{-1}\boldsymbol{c}_p=\mu_2(K)$. This completes the proof. □

Lemma A1. *Assume Condition* B *in the Appendix. If $n\to\infty$, $h\to 0$ and $nh\to\infty$, then for given $t\in\mathcal{T}_0$ an $\boldsymbol{\beta}\in\mathcal{K}$,*

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{p}_2(Y_i;\eta_i(t;\boldsymbol{\beta}))\{\mathbf{t}_i^*(t)\mathbf{t}_i^*(t)^T\}w_1(\boldsymbol{X}_i)\,K_h(T_i-t)=\mathbf{S}_2(t,\boldsymbol{\beta})+o_{\mathrm{P}}(1),$$

where $\mathbf{S}_2(t,\boldsymbol{\beta})=g_2(t;t,\boldsymbol{\beta})f_T(t)\mathcal{S}$, with $\mathcal{S}=(\mu_{j+k-2}(K))_{1\leq j,k\leq p+1}$ and $\mu_j(K)=\int u^jK(u)du$, $j=0,1,\ldots,2p$.

Proof. Recall the $(p+1)\times(p+1)$ matrix $\mathbf{t}_i^*(t)\mathbf{t}_i^*(t)^T=((\frac{T_i-t}{h})^{j+k-2})_{1\leq j,k\leq p+1}$. Set $X_j=\frac{1}{n}\sum_{i=1}^{n}\mathrm{p}_2(Y_i;\eta_i(t;\boldsymbol{\beta}))(\frac{T_i-t}{h})^jw_1(\boldsymbol{X}_i)\,K_h(T_i-t)$ for $j=0,1,\ldots,2p$. We observe that:

$$\begin{aligned} \mathrm{E}(X_j) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\Big[\mathrm{E}\{\mathrm{p}_2(Y_i;\eta_i(t;\boldsymbol{\beta}))w_1(\boldsymbol{X}_i)\mid T_i\}\Big(\frac{T_i-t}{h}\Big)^jK_h(T_i-t)\Big] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\Big\{g_2(T_i;t,\boldsymbol{\beta})\Big(\frac{T_i-t}{h}\Big)^jK_h(T_i-t)\Big\} \\ &= \mathrm{E}\Big\{g_2(T;t,\boldsymbol{\beta})\Big(\frac{T-t}{h}\Big)^jK_h(T-t)\Big\} \\ &= \int g_2(y;t,\boldsymbol{\beta})\Big(\frac{y-t}{h}\Big)^j\frac{1}{h}K\Big(\frac{y-t}{h}\Big)f_T(y)dy \\ &= \int g_2(t+hx;t,\boldsymbol{\beta})x^jK(x)f_T(t+hx)dx \\ &= g_2(t;t,\boldsymbol{\beta})f_T(t)\mu_j(K)+o(1), \end{aligned}$$

using the continuity of $g_2(\tau;t,\boldsymbol{\beta})$ in τ and $f_T(t)$ in t. Similarly,

$$\begin{aligned} \mathrm{var}(X_j) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}\Big\{\mathrm{p}_2(Y_i;\eta_i(t;\boldsymbol{\beta}))\Big(\frac{T_i-t}{h}\Big)^jw_1(\boldsymbol{X}_i)\,K_h(T_i-t)\Big\} \\ &\leq \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}\Big[\mathrm{p}_2^2(Y_i;\eta_i(t;\boldsymbol{\beta}))\Big(\frac{T_i-t}{h}\Big)^{2j}w_1^2(\boldsymbol{X}_i)\,\{K_h(T_i-t)\}^2\Big] \\ &\leq \frac{C}{nh}. \end{aligned}$$

This completes the proof. □

Lemma A2. *Assume Condition* B. *If $n\to\infty$, $h\to 0$, $nh\to\infty$, $\log(1/h)/(nh)\to 0$, then $\sup_{\boldsymbol{a}^*\in\Theta,t\in\mathcal{T}_0,\boldsymbol{\beta}\in\mathcal{K}}|G_n(\boldsymbol{a}^*;t,\boldsymbol{\beta})-\sqrt{nh}\,\mathbf{W}_n(t,\boldsymbol{\beta})^T\boldsymbol{a}^*-2^{-1}\boldsymbol{a}^{*T}\mathbf{S}_2(t,\boldsymbol{\beta})\boldsymbol{a}^*|=o_{\mathrm{P}}(1)$, with a compact set $\Theta\subseteq\mathbb{R}^{p+1}$.*

Proof. Let $D_n(\boldsymbol{a}^*;t,\boldsymbol{\beta})=G_n(\boldsymbol{a}^*;t,\boldsymbol{\beta})-\sqrt{nh}\,\mathbf{W}_n(t,\boldsymbol{\beta})^T\boldsymbol{a}^*$. Note that:

$$\begin{aligned} & D_n(\boldsymbol{a}^*;t,\boldsymbol{\beta}) \\ = & nh\Big[\frac{1}{n}\sum_{i=1}^{n}\rho_q(Y_i,F^{-1}(\eta_i(t;\boldsymbol{\beta})+\{a_n\mathbf{t}_i^*(t)^T\boldsymbol{a}^*\}))w_1(\boldsymbol{X}_i)\,K_h(T_i-t) \\ & -\frac{1}{n}\sum_{i=1}^{n}\rho_q(Y_i,F^{-1}(\eta_i(t;\boldsymbol{\beta})))w_1(\boldsymbol{X}_i)\,K_h(T_i-t) \end{aligned}$$

$$
\begin{aligned}
& -\frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_1(Y_i;\eta_i(t;\boldsymbol{\beta}))\{a_n \mathbf{t}_i^*(t)^T \boldsymbol{a}^*\} w_1(\boldsymbol{X}_i)\, K_h(T_i - t)\Big] \\
= \; & \frac{1}{2}\boldsymbol{a}^{*T}\left[\frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_2(Y_i;\widetilde{\eta}_i(t;\boldsymbol{\beta}))\{\mathbf{t}_i^*(t)\mathbf{t}_i^*(t)^T\} w_1(\boldsymbol{X}_i)\, K_h(T_i - t)\right]\boldsymbol{a}^*,
\end{aligned}
$$

where $a_n = 1/\sqrt{nh}$ and $\widetilde{\eta}_i(t;\boldsymbol{\beta})$ is between $\eta_i(t;\boldsymbol{\beta})$ and $\eta_i(t;\boldsymbol{\beta}) + \{a_n \mathbf{t}_i^*(t)^T \boldsymbol{a}^*\}$. Then:

$$
\begin{aligned}
& |D_n(\boldsymbol{a}^*;t,\boldsymbol{\beta}) - 2^{-1}\boldsymbol{a}^{*T}\mathbf{S}_2(t,\boldsymbol{\beta})\boldsymbol{a}^*| \\
= \; & \frac{1}{2}\left|\boldsymbol{a}^{*T}\left[\frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_2(Y_i;\widetilde{\eta}_i(t;\boldsymbol{\beta}))\{\mathbf{t}_i^*(t)\mathbf{t}_i^*(t)^T\} w_1(\boldsymbol{X}_i)\, K_h(T_i - t) - \mathbf{S}_2(t,\boldsymbol{\beta})\right]\boldsymbol{a}^*\right| \\
\le \; & \|\boldsymbol{a}^*\|^2\left|\frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_2(Y_i;\widetilde{\eta}_i(t;\boldsymbol{\beta}))\{\mathbf{t}_i^*(t)\mathbf{t}_i^*(t)^T\} w_1(\boldsymbol{X}_i)\, K_h(T_i - t) - \mathbf{S}_2(t,\boldsymbol{\beta})\right|.
\end{aligned}
$$

The proof completes by applying [31]. □

Proof of Theorem 1. Before showing Theorem 1, we need Proposition A1 (whose proof is omitted), where the following notation will be used. Denote by $\mathcal{C}^1(\mathcal{T})$ the set of continuously differentiable functions in $\mathcal{T}$. Let $\mathcal{V}(\boldsymbol{\beta})$ denote the neighborhood of $\boldsymbol{\beta} \in \mathcal{K}$. Let $\mathcal{H}_\delta(\boldsymbol{\beta})$ denote the neighborhood of $\eta_{\boldsymbol{\beta}}$ such that $\mathcal{V}(\boldsymbol{\beta}) \subseteq \mathcal{K}$ and $\mathcal{H}_\delta(\boldsymbol{\beta}) = \{u \in \mathcal{C}^1(\mathcal{T}) : \|u - \eta_{\boldsymbol{\beta}}\|_\infty \le \delta,\ \|\frac{\partial}{\partial t}u - \frac{\partial}{\partial t}\eta_{\boldsymbol{\beta}}\|_\infty \le \delta\}$.

Proposition A1. *Let $\{(Y_i, \boldsymbol{X}_i, T_i)\}_{i=1}^n$ be independent observations of $(Y, \boldsymbol{X}, T)$ modeled by (2) and (5). Assume that a random variable T is distributed on $\mathcal{T}$. Let $\mathcal{K}$ and $\mathcal{H}_1(\boldsymbol{\beta})$ be compact sets, $g(\cdot;\cdot) : \mathbb{R}^2 \to \mathbb{R}$ be a continuous and bounded function, $W(\boldsymbol{x}, t) : \mathbb{R}^{d+1} \to \mathbb{R}$ be such that $\mathrm{E}\{|W(\boldsymbol{X}, T)|\} < \infty$ and $\eta_{\boldsymbol{\beta}}(t) = \eta(t, \boldsymbol{\beta}) : \mathbb{R}^{d+1} \to \mathbb{R}$ be a continuous function of $(t, \boldsymbol{\beta})$. Then:*

(i) $\mathrm{E}\{g(Y; \boldsymbol{X}^T\boldsymbol{\theta} + v(T))\, W(\boldsymbol{X}, T)\} \to \mathrm{E}\{g(Y; \boldsymbol{X}^T\boldsymbol{\beta} + \eta_{\boldsymbol{\beta}}(T))\, W(\boldsymbol{X}, T)\}$ *as* $\|\boldsymbol{\theta} - \boldsymbol{\beta}\| + \|v - \eta_{\boldsymbol{\beta}}\|_\infty \to 0$;

(ii) $\sup_{\boldsymbol{\theta}\in\mathcal{K}} |n^{-1}\sum_{i=1}^n g(Y_i; \boldsymbol{X}_i^T\boldsymbol{\theta} + \eta_\theta(T_i))\, W(\boldsymbol{X}, T) - \mathrm{E}\{g(Y; \boldsymbol{X}^T\boldsymbol{\theta} + \eta_\theta(T))\, W(\boldsymbol{X}, T)\} \xrightarrow{\mathrm{P}} 0$ *as* $n \to \infty$;

(iii) *if, in addition, $\mathcal{T}$ is compact and $\eta_{\boldsymbol{\beta}} \in \mathcal{C}^1(\mathcal{T})$, then* $\sup_{\boldsymbol{\theta}\in\mathcal{K},\, v\in\mathcal{H}_1(\boldsymbol{\beta})}$
$|n^{-1}\sum_{i=1}^n g(Y_i; \boldsymbol{X}_i^T\boldsymbol{\theta} + v(T_i))\, W(\boldsymbol{X}_i, T_i) - \mathrm{E}\{g(Y; \boldsymbol{X}^T\boldsymbol{\theta} + v(T))\, W(\boldsymbol{X}, T)\}| \xrightarrow{\mathrm{P}} 0$ *as* $n \to \infty$.

For part (i), we first show that for any compact set $\mathcal{K}$ in $\mathbb{R}^d$,

$$
\sup_{\boldsymbol{\beta}\in\mathcal{K}} |J_n(\boldsymbol{\beta}, \widehat{\eta}_{\boldsymbol{\beta}}) - J(\boldsymbol{\beta}, \eta_{\boldsymbol{\beta}})| \xrightarrow{\mathrm{P}} 0. \tag{A3}
$$

It suffices to show $\sup_{\boldsymbol{\beta}\in\mathcal{K}} |J_n(\boldsymbol{\beta}, \eta_{\boldsymbol{\beta}}) - J(\boldsymbol{\beta}, \eta_{\boldsymbol{\beta}})| \xrightarrow{\mathrm{P}} 0$, which follows from Proposition A1 (ii), and:

$$
\sup_{\boldsymbol{\beta}\in\mathcal{K}} |J_n(\boldsymbol{\beta}, \widehat{\eta}_{\boldsymbol{\beta}}) - J_n(\boldsymbol{\beta}, \eta_{\boldsymbol{\beta}})| \quad \xrightarrow{\mathrm{P}} \quad 0. \tag{A4}
$$

To show (A4), we note that for any $\varepsilon > 0$, let $\mathcal{T}_0$ be a compact set such that $\mathrm{P}(T_i \notin \mathcal{T}_0) < \varepsilon$. Then:

$$
\begin{aligned}
& J_n(\boldsymbol{\beta}, \widehat{\eta}_{\boldsymbol{\beta}}) - J_n(\boldsymbol{\beta}, \eta_{\boldsymbol{\beta}}) \\
= \; & \frac{1}{n}\sum_{i=1}^{n}\{\rho_q(Y_i, F^{-1}(\boldsymbol{X}_i^T\boldsymbol{\beta} + \widehat{\eta}_{\boldsymbol{\beta}}(T_i))) - \rho_q(Y_i, F^{-1}(\boldsymbol{X}_i^T\boldsymbol{\beta} + \eta_{\boldsymbol{\beta}}(T_i)))\}\, w_2(\boldsymbol{X}_i)\, \mathrm{I}(T_i \in \mathcal{T}_0) \\
& + \frac{1}{n}\sum_{i=1}^{n}\{\rho_q(Y_i, F^{-1}(\boldsymbol{X}_i^T\boldsymbol{\beta} + \widehat{\eta}_{\boldsymbol{\beta}}(T_i))) - \rho_q(Y_i, F^{-1}(\boldsymbol{X}_i^T\boldsymbol{\beta} + \eta_{\boldsymbol{\beta}}(T_i)))\}\, w_2(\boldsymbol{X}_i)\, \mathrm{I}(T_i \notin \mathcal{T}_0).
\end{aligned}
$$

For $T_i \in \mathcal{T}_0$, by the mean-value theorem,

$$
\begin{aligned}
& |\rho_q(Y_i, F^{-1}(\boldsymbol{X}_i^T\boldsymbol{\beta} + \widehat{\eta}_{\boldsymbol{\beta}}(T_i))) - \rho_q(Y_i, F^{-1}(\boldsymbol{X}_i^T\boldsymbol{\beta} + \eta_{\boldsymbol{\beta}}(T_i)))| \\
= \; & |\mathrm{p}_1(Y_i; \boldsymbol{X}_i^T\boldsymbol{\beta} + \eta_{i,\boldsymbol{\beta}}^*)\{\widehat{\eta}_{\boldsymbol{\beta}}(T_i) - \eta_{\boldsymbol{\beta}}(T_i)\}|
\end{aligned}
$$

$$\leq \|\mathrm{p}_1(\cdot;\cdot)\|_\infty \sup_{\boldsymbol{\beta}\in\mathcal{K}} \|\widehat{\eta}_{\boldsymbol{\beta}} - \eta_{\boldsymbol{\beta}}\|_{\mathcal{T}_0;\infty},$$

where $\eta^*_{i,\boldsymbol{\beta}}$ is located between $\widehat{\eta}_{\boldsymbol{\beta}}(T_i)$ and $\eta_{\boldsymbol{\beta}}(T_i)$. For $T_i \notin \mathcal{T}_0$, it follows that:

$$\begin{aligned} & |\rho_q(Y_i, F^{-1}(\boldsymbol{X}_i^T\boldsymbol{\beta} + \widehat{\eta}_{\boldsymbol{\beta}}(T_i))) - \rho_q(Y_i, F^{-1}(\boldsymbol{X}_i^T\boldsymbol{\beta} + \eta_{\boldsymbol{\beta}}(T_i)))| \\ \leq\ & 2\|\rho_q(\cdot,\cdot)\|_\infty. \end{aligned}$$

Hence,

$$\begin{aligned} |J_n(\boldsymbol{\beta}, \widehat{\eta}_{\boldsymbol{\beta}}) - J_n(\boldsymbol{\beta}, \eta_{\boldsymbol{\beta}})| &\leq \left\{\|\mathrm{p}_1(\cdot;\cdot)\|_\infty \sup_{\boldsymbol{\beta}\in\mathcal{K}} \|\widehat{\eta}_{\boldsymbol{\beta}} - \eta_{\boldsymbol{\beta}}\|_{\mathcal{T}_0;\infty} + 2\|\rho_q(\cdot,\cdot)\|_\infty T_n^*\right\}\| w_2\|_\infty \\ &\leq 2\varepsilon, \end{aligned}$$

where the last inequality is entailed by Lemma 1 and the law of large numbers for $T_n^* = n^{-1}\sum_{i=1}^n \mathrm{I}(T_i \notin \mathcal{T}_0)$. This completes the proof of (A3). The proof of $\widehat{\boldsymbol{\beta}} \xrightarrow{\mathrm{P}} \boldsymbol{\beta}_\mathrm{o}$ follows from combining Lemma A-1 of [1] with (A3) and Condition A2.

Part (ii) follows from Lemma 1, Part (i) and Condition B5 for $\eta_{\boldsymbol{\beta}}(t)$. □

Proof of Theorem 2. Similar to the proof of Lemma 1, it can be shown that $|\widehat{\eta}_{\boldsymbol{\beta}}(t) - \eta_{\boldsymbol{\beta}}(t) + \boldsymbol{e}_{1,p+1}^T\{\mathbf{S}_2(t,\boldsymbol{\beta})\}^{-1}\frac{1}{n}\sum_{i=1}^n \mathrm{p}_1(Y_i;\eta_i(t;\boldsymbol{\beta}))\mathbf{t}_i^*(t)w_1(\boldsymbol{X}_i)\,K_h(T_i - t)| = O_\mathrm{P}(h^2a_n + a_n^2\sqrt{\log(1/h)})$. Note that for $p = 1$,

$$\begin{aligned} \boldsymbol{e}_{1,p+1}^T\{\mathbf{S}_2(t,\boldsymbol{\beta})\}^{-1}\mathbf{t}_i^*(t) &= \frac{1}{g_2(t;t,\boldsymbol{\beta})f_T(t)}(1,0)\begin{pmatrix}1 & 0\\ 0 & 1/\mu_2(K)\end{pmatrix}\begin{pmatrix}1\\ (T_i - t)/h\end{pmatrix} \\ &= \frac{1}{g_2(t;t,\boldsymbol{\beta})f_T(t)}. \end{aligned}$$

Thus:

$$\left|\widehat{\eta}_{\boldsymbol{\beta}}(t) - \eta_{\boldsymbol{\beta}}(t) + \frac{1}{nf_T(t)g_2(t;t,\boldsymbol{\beta})}\sum_{i=1}^n \mathrm{p}_1(Y_i;\eta_i(t;\boldsymbol{\beta}))w_1(\boldsymbol{X}_i)\,K_h(T_i - t)\right| = O_\mathrm{P}(h^2a_n + a_n^2\sqrt{\log(1/h)}).$$

Consider $\widehat{\boldsymbol{\beta}}$ defined in (23). Note that:

$$\begin{aligned} \boldsymbol{X}_i^T\boldsymbol{\beta} + \widehat{\eta}_{\boldsymbol{\beta}}(T_i) &= \boldsymbol{X}_i^T\boldsymbol{\beta}_\mathrm{o} + \boldsymbol{X}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}_\mathrm{o}) + \widehat{\eta}_{(\boldsymbol{\beta}-\boldsymbol{\beta}_\mathrm{o})+\boldsymbol{\beta}_\mathrm{o}}(T_i) \\ &= \boldsymbol{X}_i^T\boldsymbol{\beta}_\mathrm{o} + c_n\boldsymbol{X}_i^T\{\sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}_\mathrm{o})\} + \widehat{\eta}_{c_n\{\sqrt{n}(\boldsymbol{\beta}-\boldsymbol{\beta}_\mathrm{o})\}+\boldsymbol{\beta}_\mathrm{o}}(T_i), \end{aligned}$$

where $c_n = 1/\sqrt{n}$. Then, $\widehat{\boldsymbol{\theta}} = \sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\mathrm{o})$ minimizes:

$$\begin{aligned} J_n(\boldsymbol{\theta}) = n\Bigg[\frac{1}{n}\sum_{i=1}^n \Big\{&\rho_q(Y_i, F^{-1}(\boldsymbol{X}_i^T\boldsymbol{\beta}_\mathrm{o} + c_n\boldsymbol{X}_i^T\boldsymbol{\theta} + \widehat{\eta}_{c_n\boldsymbol{\theta}+\boldsymbol{\beta}_\mathrm{o}}(T_i)))w_2(\boldsymbol{X}_i) \\ &-\rho_q(Y_i, F^{-1}(\boldsymbol{X}_i^T\boldsymbol{\beta}_\mathrm{o} + \widehat{\eta}_{\boldsymbol{\beta}_\mathrm{o}}(T_i)))w_2(\boldsymbol{X}_i)\Big\}\Bigg] \end{aligned}$$

with respect to $\boldsymbol{\theta}$. By Taylor expansion,

$$\begin{aligned} & J_n(\boldsymbol{\theta}) \\ =\ & n\Bigg(\frac{1}{n}\sum_{i=1}^n \mathrm{p}_1(Y_i;\boldsymbol{X}_i^T\boldsymbol{\beta}_\mathrm{o} + \widehat{\eta}_{\boldsymbol{\beta}_\mathrm{o}}(T_i))[c_n\boldsymbol{X}_i^T\boldsymbol{\theta} + \{\widehat{\eta}_{c_n\boldsymbol{\theta}+\boldsymbol{\beta}_\mathrm{o}}(T_i) - \widehat{\eta}_{\boldsymbol{\beta}_\mathrm{o}}(T_i)\}]w_2(\boldsymbol{X}_i) \end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2n}\sum_{i=1}^{n}\mathrm{p}_2(Y_i;\boldsymbol{X}_i^T\boldsymbol{\beta}_{\mathrm{o}}+\widehat{\eta}_{\beta_{\mathrm{o}}}(T_i))[c_n\boldsymbol{X}_i^T\boldsymbol{\theta}+\{\widehat{\eta}_{c_n\boldsymbol{\theta}+\beta_{\mathrm{o}}}(T_i)-\widehat{\eta}_{\beta_{\mathrm{o}}}(T_i)\}]^2w_2(\boldsymbol{X}_i)\\
&+\frac{1}{6n}\sum_{i=1}^{n}\mathrm{p}_3(Y_i;\eta_i^*)[c_n\boldsymbol{X}_i^T\boldsymbol{\theta}+\{\widehat{\eta}_{c_n\boldsymbol{\theta}+\beta_{\mathrm{o}}}(T_i)-\widehat{\eta}_{\beta_{\mathrm{o}}}(T_i)\}]^3w_2(\boldsymbol{X}_i)\Big)\\
=\;& I_{n,1}+I_{n,2}+I_{n,3},
\end{aligned}
$$

where η_i^* is located between $\boldsymbol{X}_i^T\boldsymbol{\beta}_{\mathrm{o}}+\widehat{\eta}_{\beta_{\mathrm{o}}}(T_i)$ and $\boldsymbol{X}_i^T\boldsymbol{\beta}_{\mathrm{o}}+c_n\boldsymbol{X}_i^T\boldsymbol{\theta}+\widehat{\eta}_{c_n\boldsymbol{\theta}+\beta_{\mathrm{o}}}(T_i)$,

$$
\begin{aligned}
I_{n,1} &= \sum_{i=1}^{n}\mathrm{p}_1(Y_i;\boldsymbol{X}_i^T\boldsymbol{\beta}_{\mathrm{o}}+\widehat{\eta}_{\beta_{\mathrm{o}}}(T_i))\Big\{c_n\boldsymbol{X}_i^T\boldsymbol{\theta}+\frac{\partial\widehat{\eta}_{\beta}(T_i)}{\partial\boldsymbol{\beta}}^T\Big|_{\beta=\beta_n}c_n\boldsymbol{\theta}\Big\}w_2(\boldsymbol{X}_i)\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{p}_1(Y_i;\boldsymbol{X}_i^T\boldsymbol{\beta}_{\mathrm{o}}+\widehat{\eta}_{\beta_{\mathrm{o}}}(T_i))\Big\{\boldsymbol{X}_i+\frac{\partial\widehat{\eta}_{\beta}(T_i)}{\partial\boldsymbol{\beta}}\Big|_{\beta=\beta_n}\Big\}^T\boldsymbol{\theta}w_2(\boldsymbol{X}_i)\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{p}_1(Y_i;\boldsymbol{X}_i^T\boldsymbol{\beta}_{\mathrm{o}}+\widehat{\eta}_{\beta_{\mathrm{o}}}(T_i))\Big\{\boldsymbol{X}_i+\frac{\partial\eta_{\beta}(T_i)}{\partial\boldsymbol{\beta}}\Big|_{\beta=\beta_{\mathrm{o}}}\Big\}^T\boldsymbol{\theta}w_2(\boldsymbol{X}_i)+o_{\mathrm{P}}(1),\\
I_{n,2} &= \frac{1}{2}\boldsymbol{\theta}^T\Big[\frac{1}{n}\sum_{i=1}^{n}\mathrm{p}_2(Y_i;\boldsymbol{X}_i^T\boldsymbol{\beta}_{\mathrm{o}}+\widehat{\eta}_{\beta_{\mathrm{o}}}(T_i))\\
&\quad\Big\{\boldsymbol{X}_i+\frac{\partial\widehat{\eta}_{\beta}(T_i)}{\partial\boldsymbol{\beta}}\Big|_{\beta=\beta_n}\Big\}\Big\{\boldsymbol{X}_i+\frac{\partial\widehat{\eta}_{\beta}(T_i)}{\partial\boldsymbol{\beta}}\Big|_{\beta=\beta_n}\Big\}^Tw_2(\boldsymbol{X}_i)\Big]\boldsymbol{\theta}\\
&= \frac{1}{2}\boldsymbol{\theta}^T\mathbf{B}_2\boldsymbol{\theta}+o_{\mathrm{P}}(1),\\
I_{n,3} &= o_{\mathrm{P}}(1),
\end{aligned}
$$

with $\boldsymbol{\beta}_n$ located between $\boldsymbol{\beta}_{\mathrm{o}}$ and $c_n\boldsymbol{\theta}+\boldsymbol{\beta}_{\mathrm{o}}$, and $\mathbf{B}_2=\mathbf{H}_0$ following Lemma 1, Condition A3 and Proposition A1. Thus:

$$
J_n(\boldsymbol{\theta})={I_{n,1}^*}^T\boldsymbol{\theta}+\frac{1}{2}\boldsymbol{\theta}^T\mathbf{B}_2\boldsymbol{\theta}+o_{\mathrm{P}}(1), \tag{A5}
$$

where $I_{n,1}^*=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{p}_1(Y_i;\boldsymbol{X}_i^T\boldsymbol{\beta}_{\mathrm{o}}+\widehat{\eta}_{\beta_{\mathrm{o}}}(T_i))\{\boldsymbol{X}_i+\frac{\partial\eta_{\beta}(T_i)}{\partial\boldsymbol{\beta}}|_{\beta=\beta_{\mathrm{o}}}\}w_2(\boldsymbol{X}_i)$. Note that:

$$
\begin{aligned}
I_{n,1}^* &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\mathrm{p}_1(Y_i;\boldsymbol{X}_i^T\boldsymbol{\beta}_{\mathrm{o}}+\eta_{\beta_{\mathrm{o}}}(T_i))\Big\{\boldsymbol{X}_i+\frac{\partial\eta_{\beta}(T_i)}{\partial\boldsymbol{\beta}}\Big|_{\beta=\beta_{\mathrm{o}}}\Big\}w_2(\boldsymbol{X}_i)\\
&\quad+\mathrm{p}_2(Y_i;\boldsymbol{X}_i^T\boldsymbol{\beta}_{\mathrm{o}}+\eta_{\beta_{\mathrm{o}}}(T_i))\Big\{\boldsymbol{X}_i+\frac{\partial\eta_{\beta}(T_i)}{\partial\boldsymbol{\beta}}\Big|_{\beta=\beta_{\mathrm{o}}}\Big\}w_2(\boldsymbol{X}_i)\{\widehat{\eta}_{\beta_{\mathrm{o}}}(T_i)-\eta_{\beta_{\mathrm{o}}}(T_i)\}\\
&\quad+\frac{1}{2}\mathrm{p}_3(Y_i;\eta_i^{**})\Big\{\boldsymbol{X}_i+\frac{\partial\eta_{\beta}(T_i)}{\partial\boldsymbol{\beta}}\Big|_{\beta=\beta_{\mathrm{o}}}\Big\}w_2(\boldsymbol{X}_i)\{\widehat{\eta}_{\beta_{\mathrm{o}}}(T_i)-\eta_{\beta_{\mathrm{o}}}(T_i)\}^2\Big]\\
&= T_{n,1}+T_{n,2}+T_{n,3},
\end{aligned}
$$

where η_i^{**} is between $\boldsymbol{X}_i^T\boldsymbol{\beta}_{\mathrm{o}}+\widehat{\eta}_{\beta_{\mathrm{o}}}(T_i)$ and $\boldsymbol{X}_i^T\boldsymbol{\beta}_{\mathrm{o}}+\eta_{\beta_{\mathrm{o}}}(T_i)$,

$$
\begin{aligned}
T_{n,3} &= o_{\mathrm{P}}(1),\\
T_{n,2} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{p}_2(Y_i;\boldsymbol{X}_i^T\boldsymbol{\beta}_{\mathrm{o}}+\eta_{\beta_{\mathrm{o}}}(T_i))\Big\{\boldsymbol{X}_i+\frac{\partial\eta_{\beta}(T_i)}{\partial\boldsymbol{\beta}}\Big|_{\beta=\beta_{\mathrm{o}}}\Big\}w_2(\boldsymbol{X}_i)\\
&\quad\times\frac{(-1)}{nf_T(T_i)g_2(T_i;T_i,\boldsymbol{\beta}_{\mathrm{o}})}\sum_{j=1}^{n}\mathrm{p}_1(Y_j;\eta_j(T_i;\boldsymbol{\beta}_{\mathrm{o}}))w_1(\boldsymbol{X}_j)\,K_h(T_j-T_i)\\
&= -\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{\mathrm{p}_1(Y_j;\eta_j)w_1(\boldsymbol{X}_j)}{g_2(T_j;T_j,\boldsymbol{\beta}_{\mathrm{o}})}\mathrm{E}\Big[\mathrm{p}_2(Y_j;\eta_j)\Big\{\boldsymbol{X}_j+\frac{\partial\eta_{\beta}(T_j)}{\partial\boldsymbol{\beta}}\Big|_{\beta=\beta_{\mathrm{o}}}\Big\}w_2(\boldsymbol{X}_j)\Big|T_j\Big]\\
&\equiv -\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathrm{p}_1(Y_j;\eta_j)\frac{\gamma(T_j)}{g_2(T_j;T_j,\boldsymbol{\beta}_{\mathrm{o}})}w_1(\boldsymbol{X}_j),
\end{aligned}
$$

with:

$$\gamma(t) = \mathrm{E}\Big[\mathrm{p}_2(Y;\eta(\boldsymbol{X},T))\Big\{\boldsymbol{X} + \frac{\partial \eta_\beta(T)}{\partial \boldsymbol{\beta}}\Big|_{\beta=\beta_\mathrm{o}}\Big\}w_2(\boldsymbol{X})\Big|T=t\Big].$$

Therefore,

$$I^*_{n,1} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{p}_1(Y_i;\eta_i)\Big[\Big\{\boldsymbol{X}_i + \frac{\partial \eta_\beta(T_i)}{\partial \boldsymbol{\beta}}\Big|_{\beta=\beta_\mathrm{o}}\Big\}w_2(\boldsymbol{X}_i) - \frac{\gamma(T_i)}{g_2(T_i;T_i,\boldsymbol{\beta}_\mathrm{o})}w_1(\boldsymbol{X}_i)\Big] + o_\mathrm{P}(1).$$

By the central limit theorem,

$$I^*_{n,1} \xrightarrow{\mathcal{D}} N(\mathbf{0},\Omega^*_0), \tag{A6}$$

where:

$$\begin{aligned}\Omega^*_0 \;=\; & \mathrm{E}\Big(\mathrm{p}_1^2(Y;\eta(\boldsymbol{X},T))\Big[\Big\{\boldsymbol{X} + \frac{\partial \eta_\beta(T)}{\partial \boldsymbol{\beta}}\Big|_{\beta=\beta_\mathrm{o}}\Big\}w_2(\boldsymbol{X}) - \frac{\gamma(T)}{g_2(T;T,\boldsymbol{\beta}_\mathrm{o})}w_1(\boldsymbol{X})\Big]\\ & \Big[\Big\{\boldsymbol{X} + \frac{\partial \eta_\beta(T)}{\partial \boldsymbol{\beta}}\Big|_{\beta=\beta_\mathrm{o}}\Big\}w_2(\boldsymbol{X}) - \frac{\gamma(T)}{g_2(T;T,\boldsymbol{\beta}_\mathrm{o})}w_1(\boldsymbol{X})\Big]^T\Big).\end{aligned}$$

From (A5) and (A6), $\widehat{\boldsymbol{\theta}} = -\mathbf{B}_2^{-1}I^*_{n,1} + o_\mathrm{P}(1)$. This implies that $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\mathrm{o}) \xrightarrow{\mathcal{D}} N(\mathbf{0},\mathbf{H}_0^{-1}\Omega^*_0\mathbf{H}_0^{-1})$. □

Proof of Theorem 3. Denote $\mathbf{V}_0 = \mathbf{H}_0^{-1}\Omega^*_0\mathbf{H}_0^{-1}$ and $\widehat{\mathbf{V}}_n = \widehat{\mathbf{H}}_0^{-1}\widehat{\Omega}^*_0\widehat{\mathbf{H}}_0^{-1}$. Note that $\mathrm{A}\widehat{\boldsymbol{\beta}} - \boldsymbol{g}_0 = \mathrm{A}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\mathrm{o}) + (\mathrm{A}\boldsymbol{\beta}_\mathrm{o} - \boldsymbol{g}_0)$. Thus:

$$\begin{aligned} & (\mathrm{A}\widehat{\mathbf{V}}_n\mathrm{A}^T)^{-1/2}\sqrt{n}(\mathrm{A}\widehat{\boldsymbol{\beta}} - \boldsymbol{g}_0)\\ = \;& (\mathrm{A}\widehat{\mathbf{V}}_n\mathrm{A}^T)^{-1/2}\{\mathrm{A}\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\mathrm{o})\} + (\mathrm{A}\widehat{\mathbf{V}}_n\mathrm{A}^T)^{-1/2}\{\sqrt{n}(\mathrm{A}\boldsymbol{\beta}_\mathrm{o} - \boldsymbol{g}_0)\}\\ \equiv \;& I_1 + I_2,\end{aligned}$$

which implies that $W_n = \|I_1 + I_2\|^2$. Arguments for Theorem 2 give $I_1 \xrightarrow{\mathcal{D}} N(\mathbf{0},\mathbf{I}_\mathsf{k})$. Under H_0 in (4), $I_2 \equiv \mathbf{0}$ and thus $(I_1 + I_2) \xrightarrow{\mathcal{D}} N(\mathbf{0},\mathbf{I}_\mathsf{k})$, which completes the proof. □

Proof of Theorem 4. Follow the notation and proof in Theorem 3. Under H_{1n} in (29), $I_2 \xrightarrow{\mathrm{P}} (\mathrm{A}\mathbf{V}_0\mathrm{A}^T)^{-1/2}\boldsymbol{c}$ and thus $(I_1 + I_2) \xrightarrow{\mathcal{D}} N((\mathrm{A}\mathbf{V}_0\mathrm{A}^T)^{-1/2}\boldsymbol{c},\mathbf{I}_\mathsf{k})$. This completes the proof. □

Proof of Theorem 5. Following the notation and proof in Theorem 3, $W_n = \|I_1\|^2 + 2I_1^TI_2 + \|I_2\|^2$. We see that $\|I_1\|^2 \xrightarrow{\mathcal{D}} \chi^2_\mathsf{k}$. Under H_1 in (4), $I_2 = (\mathrm{A}\mathbf{V}_0\mathrm{A}^T)^{-1/2}\sqrt{n}(\mathrm{A}\boldsymbol{\beta}_\mathrm{o} - \boldsymbol{g}_0)\{1 + o_\mathrm{P}(1)\}$, which means $\|I_2\|^2 = n(\mathrm{A}\boldsymbol{\beta}_\mathrm{o} - \boldsymbol{g}_0)^T(\mathrm{A}\mathbf{V}_0\mathrm{A}^T)^{-1}(\mathrm{A}\boldsymbol{\beta}_\mathrm{o} - \boldsymbol{g}_0)\{1 + o_\mathrm{P}(1)\}$ and thus $I_1^TI_2 = O_\mathrm{P}(\sqrt{n})$. Hence, $n^{-1}W_n \geq \lambda_{\min}\{(\mathrm{A}\mathbf{V}_0\mathrm{A}^T)^{-1}\}\|\mathrm{A}\boldsymbol{\beta}_\mathrm{o} - \boldsymbol{g}_0\|^2 + o_\mathrm{P}(1)$. This completes the proof. □

Proof of Theorem 6. Denote $\mathrm{J}_n(\boldsymbol{\beta}) = J_n(\boldsymbol{\beta},\widehat{\eta}_\beta)$. For the matrix A in (4), there exists a $(d-\mathsf{k})\times d$ matrix B satisfying $BB^T = \mathbf{I}_{d-\mathsf{k}}$ and $AB^T = \mathbf{0}$. Therefore, $\mathrm{A}\boldsymbol{\beta} = \boldsymbol{g}_0$ is equivalent to $\boldsymbol{\beta} = B^T\boldsymbol{\gamma} + \boldsymbol{b}_0$ for some vector $\boldsymbol{\gamma} \in \mathbb{R}^{d-\mathsf{k}}$ and $\boldsymbol{b}_0 = \mathrm{A}^T(\mathrm{A}\mathrm{A}^T)^{-1}\boldsymbol{g}_0$. Then, minimizing $\mathrm{J}_n(\boldsymbol{\beta})$ subject to $\mathrm{A}\boldsymbol{\beta} = \boldsymbol{g}_0$ is equivalent to minimizing $\mathrm{J}_n(B^T\boldsymbol{\gamma} + \boldsymbol{b}_0)$ with respect to $\boldsymbol{\gamma}$, and we denote by $\widehat{\boldsymbol{\gamma}}$ the minimizer. Furthermore, under H_0 in (4), we have $\boldsymbol{\beta}_\mathrm{o} = B^T\boldsymbol{\gamma}_0 + \boldsymbol{b}_0$ for $\boldsymbol{\gamma}_0 = B\boldsymbol{\beta}_\mathrm{o}$, and $\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0 \xrightarrow{\mathrm{P}} \mathbf{0}$.

For Part (i), using the Taylor expansion around $\widehat{\boldsymbol{\beta}}$, we get:

$$\mathrm{J}_n(B^T\widehat{\boldsymbol{\gamma}} + \boldsymbol{b}_0) - \mathrm{J}_n(\widehat{\boldsymbol{\beta}}) = \frac{1}{2n}\{\sqrt{n}(B^T\widehat{\boldsymbol{\gamma}} + \boldsymbol{b}_0 - \widehat{\boldsymbol{\beta}})\}^T\mathrm{J}''_n(\widetilde{\boldsymbol{\beta}})\{\sqrt{n}(B^T\widehat{\boldsymbol{\gamma}} + \boldsymbol{b}_0 - \widehat{\boldsymbol{\beta}})\}, \tag{A7}$$

where $\widetilde{\boldsymbol{\beta}}$ is between $B^T\widehat{\boldsymbol{\gamma}}+\boldsymbol{b}_0$ and $\widehat{\boldsymbol{\beta}}$. We now discuss $B^T\widehat{\boldsymbol{\gamma}}+\boldsymbol{b}_0-\widehat{\boldsymbol{\beta}}$. From the proof in Theorem 2, $(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{\mathrm{o}})=-\mathbf{H}_0^{-1}\mathrm{J}_n'(\boldsymbol{\beta}_{\mathrm{o}})\{1+o_{\mathrm{P}}(1)\}$, where $\mathrm{J}_n'(\boldsymbol{\beta}_{\mathrm{o}})=\{I_{n,1}^*+o_{\mathrm{P}}(1)\}/\sqrt{n}$. Similar arguments deduce $\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0=-(B\mathbf{H}_0B^T)^{-1}B\mathrm{J}_n'(\boldsymbol{\beta}_{\mathrm{o}})\{1+o_{\mathrm{P}}(1)\}$. Thus, under H_0 in (4),

$$B^T\widehat{\boldsymbol{\gamma}}+\boldsymbol{b}_0-\widehat{\boldsymbol{\beta}}=B^T(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)-(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{\mathrm{o}})=\mathbf{H}_0^{-1/2}P_{\mathbf{H}_0^{-1/2}\mathrm{A}^T}\mathbf{H}_0^{-1/2}\mathrm{J}_n'(\boldsymbol{\beta}_{\mathrm{o}})\{1+o_{\mathrm{P}}(1)\},$$

and thus by (A6),

$$\sqrt{n}(B^T\widehat{\boldsymbol{\gamma}}+\boldsymbol{b}_0-\widehat{\boldsymbol{\beta}})\xrightarrow{\mathcal{D}}\mathbf{H}_0^{-1/2}P_{\mathbf{H}_0^{-1/2}\mathrm{A}^T}\mathbf{H}_0^{-1/2}\Omega_0^{*1/2}\boldsymbol{Z}, \tag{A8}$$

where $\boldsymbol{Z}=(Z_1,\ldots,Z_d)^T\sim N(\mathbf{0},\mathbf{I}_d)$. Combining the fact $\mathrm{J}_n''(\widetilde{\boldsymbol{\beta}})\xrightarrow{\mathrm{P}}\mathbf{H}_0$, (A7) and (A8) gives:

$$\begin{aligned}\Lambda_n &= \{\sqrt{n}(B^T\widehat{\boldsymbol{\gamma}}+\boldsymbol{b}_0-\widehat{\boldsymbol{\beta}})\}^T\mathbf{H}_0\{\sqrt{n}(B^T\widehat{\boldsymbol{\gamma}}+\boldsymbol{b}_0-\widehat{\boldsymbol{\beta}})\}\{1+o_{\mathrm{P}}(1)\}\\ &\xrightarrow{\mathcal{D}} \boldsymbol{Z}^T\Omega_0^{*1/2}\mathbf{H}_0^{-1/2}P_{\mathbf{H}_0^{-1/2}\mathrm{A}^T}\mathbf{H}_0^{-1/2}\Omega_0^{*1/2}\boldsymbol{Z}\\ &= \sum_{j=1}^{d}\lambda_j(\Omega_0^{*1/2}\mathbf{H}_0^{-1/2}P_{\mathbf{H}_0^{-1/2}\mathrm{A}^T}\mathbf{H}_0^{-1/2}\Omega_0^{*1/2})Z_j^2\\ &= \sum_{j=1}^{\mathrm{k}}\lambda_j\{(\mathrm{A}\mathbf{H}_0^{-1}\mathrm{A}^T)^{-1}(\mathrm{A}\mathbf{V}_0\mathrm{A}^T)\}Z_j^2. \end{aligned}\tag{A9}$$

This proves Part (i).

For Part (ii), using $\psi(r)=r$, $w_1(\boldsymbol{x})=w_2(\boldsymbol{x})\equiv 1$ and (31), we obtain $\Omega_0^*=\Omega_0=C\mathbf{H}_0$, and thus, $\mathrm{A}\mathbf{V}_0\mathrm{A}^T=C(\mathrm{A}\mathbf{H}_0^{-1}\mathrm{A}^T)$. Thus, (A9) $=C\sum_{j=1}^{\mathrm{k}}Z_j^2\sim C\chi_{\mathrm{k}}^2$, which completes the proof. □

Proof of Theorem 7. The proofs are similar to those used in Theorem 4 and Theorems 5 and 6. The lengthy details are omitted. □

References

1. Andrews, D. Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica* **1994**, *62*, 43–72.
2. Robinson, P.M. Root-n consistent semiparametric regression. *Econometrica* **1988**, *56*, 931–954.
3. Speckman, P. Kernel smoothing in partial linear models. *J. R. Statist. Soc. B* **1988**, *50*, 413–436.
4. Yatchew, A. An elementary estimator of the partial linear model. *Econ. Lett.* **1997**, *57*, 135–143.
5. Fan, J.; Li, R. New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *J. Am. Stat. Assoc.* **2004**, *99*, 710–723.
6. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*, 2nd ed.; Chapman & Hall: London, UK, 1989.
7. Zhang, C.M.; Yu, T. Semiparametric detection of significant activation for brain fMRI. *Ann. Stat.* **2008**, *36*, 1693–1725.
8. Fan, J.; Huang, T. Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **2005**, *11*, 1031–1057.
9. Boente, G.; He, X.; Zhou, J. Robust estimates in generalized partially linear models. *Ann. Stat.* **2006**, *34*, 2856–2878.
10. Zhang, C.M.; Guo, X.; Cheng, C.; Zhang, Z.J. Robust-BD estimation and inference for varying-dimensional general linear models. *Stat. Sin.* **2014**, *24*, 653–673.
11. Fan, J.; Gijbels, I. *Local Polynomial Modeling and Its Applications*; Chapman and Hall: London, UK, 1996.
12. Brègman, L.M. A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Math. Phys.* **1967**, *7*, 620–631.
13. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer: New York, NY, USA, 2001.
14. Zhang, C.M.; Jiang, Y.; Shang, Z. New aspects of Bregman divergence in regression and classification with parametric and nonparametric estimation. *Can. J. Stat.* **2009**, *37*, 119–139.
15. Huber, P. Robust estimation of a location parameter. *Ann. Math. Statist.* **1964**, *35*, 73–101.
16. Van der Vaart, A.W. *Asymptotic Statistics*; Cambridge University Press: Cambridge, UK, 1998.

17. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33*, 1–22.
18. Carroll, R.; Fan, J.; Gijbels, I.; Wand, M. Generalized partially linear single-index models. *J. Am. Stat. Assoc.* **1997**, *92*, 477–489.
19. Mukarjee, H.; Stern, S. Feasible nonparametric estimation of multiargument monotone functions. *J. Am. Stat. Assoc.* **1994**, *89*, 77–80.
20. Albright, S.C.; Winston, W.L.; Zappe, C.J. *Data Analysis and Decision Making with Microsoft Excel*; Duxbury Press: Pacific Grove, CA, USA, 1999.
21. Fan, J.; Peng, H. Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Stat.* **2004**, *32*, 928–961.
22. Dette, H. A consistent test for the functional form of a regression based on a difference of variance estimators. *Ann. Stat.* **1999**, *27*, 1012–1050.
23. Dette, H.; von Lieres und Wilkau, C. Testing additivity by kernel-based methods. *Bernoulli* **2001**, *7*, 669–697.
24. Fan, J.; Zhang, C.M.; Zhang, J. Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Stat.* **2001**, *29*, 153–193.
25. Hong, Y.M.; Lee, Y.J. A loss function approach to model specification testing and its relative efficiency. *Ann. Stat.* **2013**, *41*, 1166–1203.
26. Zheng, J.X. A consistent test of functional form via nonparametric estimation techniques. *J. Econ.* **1996**, *75*, 263–289.
27. Opsomer, J.D.; Ruppert D. A root-n consistent backfitting estimator for semiparametric additive modeling. *J. Comput. Graph. Stat.* **1999**, *8*, 715–732.
28. Belloni, A.; Chernozhukov, V.; Hansen, C. Inference on treatment effects after selection amongst high-dimensional controls. *Rev. Econ. Stud.* **2014**, *81*, 608–650.
29. Cattaneo, M.D.; Jansson, M.; Newey, W.K. Alternative asymptotics and the partially linear model with many regressors. *Econ. Theory* **2016**, 1–25.
30. Cattaneo, M.D.; Jansson, M.; Newey, W.K. Treatment effects with many covariates and heteroskedasticity. *arXiv* **2015**, arXiv:1507.02493.
31. Mack, Y.P.; Silverman, B.W. Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrsch. Verw. Gebiete* **1982**, *61*, 405–415.