Research Collection School Of Economics

School of Economics

Non-separable models with high-dimensional data

Liangjun SU
Singapore Management University, ljsu@smu.edu.sg

T URA

YC ZHANG

Citation

Non-separable models with high-dimensional data☆

Liangjun Su [a], Takuya Ura [b], Yichong Zhang [a],*

[a] *School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903, Singapore*
[b] *Department of Economics, University of California, Davis. One Shields Avenue, Davis, CA 95616, United States*

ARTICLE INFO

ABSTRACT

This paper studies non-separable models with a continuous treatment when the dimension of the control variables is high and potentially larger than the effective sample size. We propose a three-step estimation procedure to estimate the average, quantile, and marginal treatment effects. In the first stage we estimate the conditional mean, distribution, and density objects by penalized local least squares, penalized local maximum likelihood estimation, and numerical differentiation, respectively, where control variables are selected via a localized method of L_1-penalization at each value of the continuous treatment. In the second stage we estimate the average and marginal distribution of the potential outcome via the plug-in principle. In the third stage, we estimate the quantile and marginal treatment effects by inverting the estimated distribution function and using the local linear regression, respectively. We study the asymptotic properties of these estimators and propose a weighted-bootstrap method for inference. Using simulated and real datasets, we demonstrate that the proposed estimators perform well in finite samples.

1. Introduction

Non-separable models without additivity appear frequently in econometric analyses, because economic theory motivates a nonlinear role of the unobserved individual heterogeneity (Altonji and Matzkin, 2005) and its multidimensionality (Browning and Carro, 2007; Carneiro et al., 2003; Cunha et al., 2010). A large fraction of the previous literature on non-separable models has used control variables to achieve the unconfoundedness condition (Rosenbaum and Rubin, 1983), that is, the conditional independence between a regressor of interest (or a treatment) and the unobserved individual heterogeneity given the control variables. Although including high-dimensional control variables makes unconfoundedness more plausible, the estimation and inference become more challenging, as well. It remains unanswered how to select control variables among potentially very many variables and conduct proper statistical inference for parameters of interest in non-separable models with a continuous treatment.

This paper proposes estimation and inference for unconditional parameters,[1] including unconditional means of the potential outcomes, the unconditional cumulative distribution function, the unconditional quantile function, and the unconditional quantile partial derivative with the presence of both continuous treatment and high-dimensional covariates.[2] The proposed method estimates the parameters of interest in three stages. The first stage selects controls by the method of least absolute shrinkage and selection operator (Lasso) and predicts reduced-form parameters such as the conditional expectation and distribution of the outcome given the variables and treatment level and the conditional density of the treatment given the control variables. We allow for different control variables to be selected at different values of the continuous treatment. The second stage recovers the average and the marginal distribution of the potential outcome by plugging the reduced-form parameters into doubly robust moment conditions. The last stage recovers the quantile of the potential outcome and its derivative with respect to the treatment by inverting the estimated distribution function and using the local linear regression, respectively. The inference is implemented via a weighted-bootstrap without recalculating the first stage variable selections, which saves considerable computation time.

To motivate our parameters of interest, we relate our estimands (the population objects that our procedure aims to recover) with the structural outcome function. Notably, we extend Hoderlein and Mammen (2007) and Sasaki (2015) to demonstrate that the unconditional derivative of the quantile of the potential outcome with respect to the treatment is equal to the weighted average of the marginal effects over individuals with same outcomes and treatments.

This paper contributes to two important strands of the econometric literature. The first is the literature on non-separable models with a continuous treatment, in which previous analyses have focused on a fixed and small number of control variables; see, e.g., Chesher (2003), Chernozhukov et al. (2007), Hoderlein and Mammen (2007), Imbens and Newey (2009), Matzkin (1994) and Matzkin (2003). The second is a growing literature on recovering the causal effect from the high-dimensional data; see, e.g., Belloni et al. (2012), Belloni et al. (2014a), Chernozhukov et al. (2015a), Chernozhukov et al. (2015b), Farrell (2015), Athey and Imbens (2016), Chernozhukov et al. (2017), Belloni et al. (2014b), Wager and Athey (2018), Belloni et al. (2017), and Belloni et al. (2016b). Our paper complements the previous works by studying both the variable selection and post-selection inference of causal parameters in a *non-separable* model with a *continuous* treatment. Recently, Cattaneo et al. (2016), Cattaneo et al. (2019), and Cattaneo et al. (2018) have considered the semiparametric estimation of the causal effect in a setting with many included covariates and proposed novel bias-correction methods to conduct valid inference. Comparing with them, we deal with the fully nonparametric model with an ultra-high dimension of potential covariates, and rely on the approximate sparsity to reduce dimensionality.

The treatment variable being continuous imposes difficulties in both variable selection and post-selection inference. To address the former, we use penalized local Maximum Likelihood and Least Square estimations (hereafter, MLE and LS, respectively) to select control variables for each value of the continuous treatment. The penalized local LS was previously studied by Kong et al. (2015) and Lee and Mammen (2016).[3] The local MLE complements the LS method by estimating a nonlinear and high-dimensional model with varying coefficients indexed by not only the continuous treatment variable but also a location variable. Our approach directly extend the distribution regression proposed in Chernozhukov et al. (2013) to the high-dimensional varying coefficient setting. By relying on the kernel smoothing method, we require a different penalty loading than the traditional Lasso method. Chu et al. (2011) and Ning and Liu (2017) develop general theories of estimation, inference, and hypothesis testing of penalized (Pseudo) MLE. We complement their results by considering the local likelihood with an L_1 penalty term. Belloni et al. (2018) construct uniformly valid confidence bands for the Z-estimators of unconditional moment equalities. Our results are not covered by theirs, either, as our parameters are defined based on conditional moment equalities. To prove the statistical properties of the penalized local MLE, we establish a local version of the compatibility condition (Bühlmann and van de Geer, 2011), which itself is new to the best of our knowledge.

For the post-selection inference, we establish doubly robust moment conditions for the continuous treatment effect model. Our parameters of interest are irregularly identified by the definition in Khan and Tamer (2010), as they are identified by a thin-set. Therefore, by averaging observations only when their treatment levels are close to the one of interest, the convergence rates of our estimators are nonparametric, which is in contrast with the $\sqrt{n}$-rate obtained in Belloni et al. (2017) and Farrell (2015). Albeit motivated by distinct models, Belloni et al. (2016a) also estimate the irregular identified parameters in the high-dimensional setting. However, the irregularity faced by Belloni et al. (2016a) is not due to the continuity of the variable of interest. Consequently, Belloni et al. (2016a) do not study the regularized estimator with localization as we do in this paper.

Estimation based on doubly robust moments is also related to the literature of semiparametric efficiency. The idea of doubly robust estimation can be traced back to the nonparametric efficiency theory for functional estimation developed by Begun et al. (1983), Pfanzagl (1990), Bickel et al. (1993), and Newey (1994). Robins and Rotnitzky (2001) and van der

---

[1] To be more specific, the parameters of interest are unconditional on covariates but conditional on the treatment level.

[2] We focus on unconditional parameters, in which (potentially high-dimensional) covariates are employed to achieve the unconfoundedness but the parameters of interest are unconditional on the covariates. Unconditional parameters are simple to display and the simplicity is crucial especially when the covariates are high dimensional. As emphasized in Frölich and Melly (2013) and Powell (2010), unconditional parameters have two additional attractive features. First, by definition, they capture all the individuals in the sample at the same time instead of investigating the underlying structure separately for each subgroup defined by the covariates X. The treatment effect for the whole population is more policy-relevant. Second, an estimator for unconditional parameters can have better finite/large sample properties.

[3] We thank a referee for the reference.

Laan and Robins (2003) study the semiparametric doubly robust estimators by modeling both the treatment and outcome processes. van der Laan and Dudoit (2003) allow for nonparametric modeling in causal inference problems. When both processes are nonparametrically estimated, the doubly robust methods can achieve faster rates of convergence than their nuisance estimator, making the estimator less sensitive to the curse of dimensionality and model selection bias. Their use in causal inference is also considered by Robins and Rotnitzky (1995), Hahn (1998), van der Laan and Robins (2003), Hirano et al. (2003), van der Laan and Rubin (2006), Firpo (2007), Tsiatis (2007), van der Laan and Rose (2011), Kennedy et al. (2017), and Robins et al. (2017), among others.

Among the works above, our paper is most closely related to Kennedy et al. (2017), who consider the doubly robust estimation for the average treatment effect when the treatment variable is continuous. Our paper complements theirs in four aspects. First, the estimation procedures are different. Kennedy et al. (2017) first estimate the efficient influence function for the weighted average of the mean effect over all treatment levels, and then, use kernel smoothing to estimate the mean effect at each treatment level. On the contrary, we consider the doubly-robust moment for the parameters of interest. Second, Kennedy et al. (2017) mainly focus on the mean effect, while we also consider quantile and marginal treatment effects. We obtain linear expansions for our estimators uniformly over both the quantile index and the treatment variable. Third, Kennedy et al. (2017) do not construct detailed estimators of their nuisance parameters, but instead, impose high-level assumptions. To verify such high-level assumptions in the high-dimensional setting is nontrivial. In contrast, we provide valid estimators for our nuisance parameters via both regularization and localization, and derive their statistical properties. Fourth, we take into account the fact that the dimension of covariates may increase with the sample size so that the complexity of our nuisance parameter estimator measured by the uniform entropy will diverge to infinity. Such a situation is ruled out by Kennedy et al. (2017).

To obtain uniformly valid results over values of the continuous treatment, we derive linear expansions of the rearrangement operator for a local process which is not tight, extending the existing results in Chernozhukov et al. (2010).

We study the finite sample performance of our estimation procedure via Monte Carlo simulations and an empirical application. The simulations suggest that the proposed estimators perform reasonably well in finite samples. In the empirical exercise, we estimate the distributional effect of parental income on son's income and intergenerational elasticity using the 1979 National Longitudinal Survey of Youth (NLSY79). We control for a large dimension of demographic variables. The quantiles of son's potential income are in general upward slopping with respect to parental income, but for the subsample of blacks, the intergenerational elasticities are not statistically significant.

The rest of this paper is organized as follows. Section 2 presents the model and the parameters of interest. Section 3 proposes an estimation method in the presence of high-dimensional covariates. Section 4 demonstrates the validity of a bootstrap inference procedure. Section 5 presents Monte Carlo simulations. Section 6 illustrates the proposed estimator using NLSY79. Section 7 concludes. Proofs of the main theorems and Lemma 3.1 are reported in the Appendix. Proofs of the rest of the lemmas are collected in an online supplement.

Throughout this paper, we adopt the convention that the capital letters, such as A, Y, X, denote random elements while their corresponding lower cases denote realizations. C denotes an arbitrary positive constant that may not be the same in different contexts. For a sequence of random variables $\{U_n\}_{n=1}^{\infty}$ and a random variable U, $U_n \rightsquigarrow U$ indicates weak convergence in the sense of van der Vaart and Wellner (1996). When U_n and U are k-dimensional elements, the space of the sample path is $\Re^k$ equipped with Euclidean norm. When U_n and U are stochastic processes, the space of sample path is $L^{\infty}(\{v \in \Re^k : |v| < B\})$ for some positive B equipped with sup norm. The letters $\mathbb{P}_n$, $\mathbb{P}$, and $\mathcal{U}_n$ denote the empirical process, expectation, and U-process, respectively. In particular, $\mathbb{P}_n$ assigns probability $\frac{1}{n}$ to each observation and $\mathcal{U}_n$ assigns probability $\frac{1}{n(n-1)}$ to each pair of observations. $\mathbb{E}$ also denotes expectation. We use $\mathbb{P}$ and $\mathbb{E}$ exchangeably. For any positive (random) sequence (u_n, v_n), if there exists a positive constant C independent of n such that $u_n \leq C v_n$, then we write $u_n \lesssim v_n$. $\| \cdot \|_{Q,q}$ denotes L^q norm under measure Q, where $q = 1, 2, \infty$. If measure Q is omitted, the underlying measure is assumed to be the counting measure. For any vector θ, $\|\theta\|_0$ denotes the number of its nonzero coordinates. Supp(θ), the support of a p-dimensional vector θ, is defined as $\{j : \theta_j \neq 0\}$. For $T \subset \{1, 2, \ldots, p\}$, let $|T|$ be the cardinality of T, T^c be the complement of T, and θ_T be the vector in $\Re^p$ that has the same coordinates as θ on T and zero coordinates on T^c. Last, let $a \vee b = \max(a, b)$.

2. Model and parameters of interest

Econometricians observe an outcome Y, a continuous treatment T, and a set of covariates X, which may be high-dimensional. They are connected by a measurable function $\Gamma(\cdot)$, i.e.,

$$Y = \Gamma(T, X, A),$$

where A is an unobservable random vector and may not be weakly separable from observables (T, X), and Γ may not be monotone in either T or A.

Let $Y(t) = \Gamma(t, X, A)$. We are interested in the average $\mathbb{E}Y(t)$, the marginal distribution $\mathbb{P}(Y(t) \leq u)$ for some $u \in \Re$, and the quantile $q_\tau(t)$, where we denote $q_\tau(t)$ as the τ-th quantile of $Y(t)$ for some $\tau \in (0, 1)$. We are also interested in the causal effect of moving T from t to t', i.e., $\mathbb{E}(Y(t) - Y(t'))$ and $q_\tau(t) - q_\tau(t')$. Last, we are interested in the average marginal effect $\mathbb{E}[\partial_t \Gamma(t, X, A)]$ and quantile partial derivative $\partial_t q_\tau(t)$. Next, we specify conditions under which the above parameters are identified.

Assumption 1. The random variables A and T are conditionally independent given X.

Assumption 1 is known as the unconfoundedness condition, which is commonly assumed in the treatment effect literature. See Cattaneo (2010), Cattaneo and Farrell (2011), Hirano et al. (2003) and Firpo (2007) for the case of discrete treatment and Graham et al. (2014), Galvao and Wang (2015), and Hirano and Imbens (2004) for the case of continuous treatment. It is also called the conditional independence assumption in Hoderlein and Mammen (2007), which is weaker than the full joint independence between A and (T, X). Note that X can be arbitrarily correlated with the unobservables A. This assumption is more plausible when we control for sufficiently many and potentially high-dimensional covariates.

Theorem 2.1. *Suppose* Assumption 1 *holds and* $\Gamma(\cdot)$ *is differentiable in its first argument. Then the marginal distribution of* $Y(t)$ *and the average marginal effect* $\partial_t \mathbb{E}Y(t)$ *are identified. In addition, if* Assumption 6 *in* Appendix A *holds and* X *is continuously distributed, then* $\partial_t q_\tau(t) = \mathbb{E}_{\mu_{\tau,t}}[\partial_t \Gamma(t, X, A)]$, *where, for* $f_{(X,A)}$ *denoting the joint density of* (X, A), $\mu_{\tau,t}$ *is the probability measure on* $\{(x, a) : \Gamma(t, x, a) = q_\tau(t)\}$ *with density* $\frac{f_{(X,A)}}{c_f \|\nabla_{(x,a)}\Gamma(t,\cdot,\cdot)\|}$, *where*

$$c_f = \int_{(x,a):\Gamma(t,x,a)=q_\tau(t)} \frac{f_{(X,A)}(x, a)}{\|\nabla_{(x,a)}\Gamma(t, \cdot, \cdot)\|} dx da.$$

Several comments are in order. First, because the marginal distribution of $Y(t)$ is identified, so be its average, quantile, average marginal effect, and quantile partial derivative. As pointed out by Imbens and Newey (2009), a non-separable outcome with a general disturbance is equivalent to treatment effect models. Therefore, we can view $Y(t)$ as the potential outcome. Under unconfoundedness, the identification of the marginal distribution of the potential outcome with a continuous treatment has already been established in Hirano and Imbens (2004) and Galvao and Wang (2015). The first part of Theorem 2.1 just re-states their results. Second, the second result indicates that the partial quantile derivative identifies the weighted average marginal effect for the subpopulation with the same potential outcome, i.e., $\{Y(t) = q_\tau(t)\}$. The result is closely related to, but different from Sasaki (2015). We consider the unconditional quantile of $Y(t)$, whereas he considered the conditional quantile of $Y(t)$ given X. Third, we require X to be continuous just for the simplicity of derivation. If some elements of X are discrete, a similar result can be established in a conceptually straightforward manner by focusing on the continuous covariates within samples homogeneous in the discrete covariates, at the expense of additional notation. Finally, we do not require X to be continuous when establishing the estimation and inference results below.

3. Estimation

Let $f_t(x) = f_{T|X}(t|x)$ denote the conditional density of T evaluated at t given $X = x$ and $d_t(\cdot)$ denote the Dirac function such that for any function $g(\cdot)$,

$$\int g(s)d_t(s)ds = g(t).$$

In addition, let $Y_u(t) = 1\{Y(t) \le u\}$ and $Y_u = 1\{Y \le u\}$ for some $u \in \Re$. Then $\mathbb{E}(Y(t))$ and $\mathbb{E}(Y_u(t))$ can be identified by the method of generalized propensity score as proposed in Hirano and Imbens (2004), i.e.,

$$\mathbb{E}(Y(t)) = \mathbb{E}\left(\frac{Yd_t(T)}{f_t(X)}\right) \quad \text{and} \quad \mathbb{E}(Y_u(t)) = \mathbb{E}\left(\frac{Y_ud_t(T)}{f_t(X)}\right). \tag{3.1}$$

There is a direct analogy between (3.1) for the continuous treatment and $\mathbb{E}(Y_u(t)) = \mathbb{E}(\frac{Y_u 1\{T=t\}}{\mathbb{P}(T=t|X)})$ when the treatment T is discrete: the indicator function shrinks to a Dirac function and the propensity score is replaced by the conditional density. Following this analogy, Hirano and Imbens (2004) called $f_t(X)$ the generalized propensity.

Belloni et al. (2017) and Farrell (2015) considered the model with a discrete treatment and high-dimensional control variables, and proposed to use the doubly robust moment for inference. Following their lead, we propose the corresponding doubly robust moment when the treatment status is continuous. Let $v_t(x) = \mathbb{E}(Y|X = x, T = t)$ and $\phi_{t,u}(x) = \mathbb{E}(Y_u|X = x, T = t)$, then

$$\mathbb{E}(Y(t)) = \mathbb{E}\left[\left(\frac{(Y - v_t(X))d_t(T)}{f_t(X)}\right) + v_t(X)\right] \tag{3.2}$$

and

$$\mathbb{E}(Y_u(t)) = \mathbb{E}\left[\left(\frac{(Y_u - \phi_{t,u}(X))d_t(T)}{f_t(X)}\right) + \phi_{t,u}(X)\right]. \tag{3.3}$$

We propose the following three-stage procedure to estimate $\mu(t) := \mathbb{E}Y(t)$, $\alpha(t, u) := \mathbb{P}(Y(t) \le u)$, $q_\tau(t)$, and $\partial_t q_\tau(t)$:

1. Estimate $v_t(x)$, $\phi_{t,u}(x)$, and $f_t(x)$ by $\widehat{v}_t(x)$, $\widehat{\phi}_{t,u}(x)$ and $\hat{f}_t(x)$, respectively, using the first-stage bandwidth h_1.

2. Estimate $\mu(t)$ and $\alpha(t, u)$ by

$$\hat{\mu}(t) = \frac{1}{n} \sum_{i=1}^{n} \left[\left(\frac{(Y - \widehat{v}_t(X_i))}{\hat{f}_t(X_i) h_2} K(\frac{T_i - t}{h_2}) \right) + \widehat{v}_t(X_i) \right]$$

and

$$\hat{\alpha}(t, u) = \frac{1}{n} \sum_{i=1}^{n} \left[\left(\frac{(Y_u - \widehat{\phi}_{t,u}(X_i))}{\hat{f}_t(X_i) h_2} K(\frac{T_i - t}{h_2}) \right) + \widehat{\phi}_{t,u}(X_i) \right], \quad \text{respectively,}$$

where $K(\cdot)$ and h_2 are a kernel function and the second-stage bandwidth, respectively. Then rearrange $\hat{\alpha}(t, u)$ to obtain $\hat{\alpha}^r(t, u)$, which is monotone in u.

3. Estimate $q_\tau(t)$ by inverting $\hat{a}^r(t, u)$ with respect to (w.r.t.) u, i.e., $\hat{q}_\tau(t) = \inf\{u : \hat{a}^r(t, u) \geq \tau\}$; estimate $\partial_t \mu(t) = \mathbb{E}\partial_t \Gamma(t, X, A)$ by $\breve{\beta}^1(t)$, which is the estimator of the slope coefficient in the local linear regression of $\hat{\mu}(T_i)$ on T_i; estimate $\partial_t q_\tau(t)$ by $\breve{\beta}_\tau^1(t)$, which is the estimator of the slope coefficient in the local linear regression of $\hat{q}_\tau(T_i)$ on T_i.

3.1. The first stage estimation

In this section, we define the first stage estimators and derive their asymptotic properties. Since $v_t(x)$, $\phi_{t,u}(x)$, and $f_t(x)$ are local parameters w.r.t. $T = t$, in addition to using L_1 penalty to select relevant covariates, we rely on a kernel function to implement the localization. In particular, we propose to estimate $v_t(x)$, $\phi_{t,u}(x)$, and $f_t(x)$ by a penalized local LS, a penalized local MLE, and numerical differentiation, respectively.

3.1.1. Penalized local LS and MLE

Recall $v_t(x) = \mathbb{E}(Y|X = x, T = t)$ and $\phi_{t,u}(x) = \mathbb{E}(Y_u|X = x, T = t)$ where $Y_u = 1\{Y \leq u\}$. We approximate $v_t(x)$ and $\phi_{t,u}(x)$ by $b(x)'\gamma_t$ and $\Lambda(b(x)'\theta_{t,u})$, respectively, where $\Lambda(\cdot)$ is the logistic CDF and $b(X)$ is a $p \times 1$ vector of basis functions with potentially large p. In the case of high-dimensional covariates, $b(X)$ is just X, while in the case of nonparametric sieve estimation, $b(X)$ is a series of bases of X. The approximation errors for $v_t(x)$ and $\phi_{t,u}(x)$ are given by $r_t^v(x) = v_t(x) - b(x)'\gamma_t$ and $r_{t,u}^\phi(x) = \phi_{t,u}(x) - \Lambda(b(x)'\theta_{t,u})$, respectively.

Note that we only approximate $v_t(x)$ and $\phi_{t,u}(x)$ by a linear regression and a logistic regression, respectively, with the approximation errors satisfying Assumption 2. Assumption 2 puts a sparsity structure on $v_t(x)$ and $\phi_{t,u}(x)$ so that the number of effective covariates that can affect them is much smaller than p. If the effective covariates are a few discrete variables that have a few categories, then we can saturate the regressions by low-dimensional dummy variables so that there is no approximate error. If some of the effective covariates are continuous, then we can include sieve bases in the linear regression so that the approximation error can still satisfy Assumption 2. One possible scenario that the approximate sparsity condition may fail is when there are a substantial amount of discrete variables that are all on the same footing (e.g., job occupation dummies). In this case, it is hard to define a sparse approximation.[4] Last, the coefficients γ_t and $\theta_{t,u}$ are both functional parameters that can vary with their indexes. This provides additional flexibility of our setup against misspecification.

We estimate $v_t(x)$ and $\phi_{t,u}(x)$ by $\widehat{v}_t(x) = b(x)'\hat{\gamma}_t$ and $\widehat{\phi}_{t,u}(x) = \Lambda(b(x)'\hat{\theta}_{t,u})$, respectively, where

$$\hat{\gamma}_t = \arg\min_\gamma \frac{1}{2n} \sum_{i=1}^{n} (Y_i - b(X_i)'\gamma)^2 K(\frac{T_i - t}{h_1}) + \frac{\lambda}{n} \|\widehat{\Xi}_t \gamma\|_1, \tag{3.4}$$

$$\hat{\theta}_{t,u} = \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} M(1\{Y_i \leq u\}, X_i; \theta) K(\frac{T_i - t}{h_1}) + \frac{\lambda}{n} \|\widehat{\Psi}_{t,u} \theta\|_1, \tag{3.5}$$

$\|\cdot\|_1$ denotes the L_1 norm, h_1 is the first-stage bandwidth, $\lambda = \ell_n (\log(p \vee nh_1)nh_1)^{1/2}$ for some slowly diverging sequence ℓ_n, and $M(y, x; g) = -[y \log(\Lambda(b(x)'g)) + (1 - y) \log(1 - \Lambda(b(x)'g))]$. Our penalty term λ is different from the one used in Belloni et al. (2017) and Belloni et al. (2019), i.e., $\lambda^* = 1.1\Phi^{-1}(1 - \gamma/p)n^{1/2}$, where $\gamma = o(1)$ is some user-supplied constant, and $\Phi(\cdot)$ is the standard normal CDF. Belloni et al. (2017) suggest $\gamma = C/(n \log(n))$, which implies that

$$\Phi^{-1}(1 - \gamma/p) \sim [\log(1/C) + \log(p) + \log(n) + \log(\log(n))]^{1/2} \sim \sqrt{\log(p \vee n)}.$$

Therefore, our penalty term λ is of same order of magnitude of λ^* if nh_1 is replaced with n and ℓ_n is removed. We need to use nh_1 in our penalty due to the presence of the kernel function in our estimation procedure. In particular, the effective sample size is of the same order of nh_1.[5] We will specify the order of magnitude of h_1 in Assumption 2. The role played by ℓ_n in our penalty is similar to that of γ in λ^*, which is to control the selection error uniformly. We refer readers to Belloni

---

[4] We thank the Associate Editor for this point.

[5] Note that $\log(n)$ and $\log(nh_1)$ are of the same order of magnitude.

et al. (2017, Equation (6.4)) for a more detailed discussion on this point. Since we do not use the advanced technique of self-normalized process as in Belloni et al. (2017), we multiply the sequence ℓ_n with $\sqrt{\log(p \vee n)}$ while in λ^*, $\log(\gamma)$ is additive to $\log(pn)$ inside the square root. We propose a rule-of-thumb λ in Section 5 and study the sensitivity of our inference method against the choice of λ in Section S.3 of the online supplementary material.

In (3.4) and (3.5), $\widehat{\Xi}_t = \mathrm{diag}(\tilde{l}_{t,1}, \ldots, \tilde{l}_{t,p})$ and $\widehat{\Psi}_{t,u} = \mathrm{diag}(l_{t,u,1}, \ldots, l_{t,u,p})$ are generic penalty loading matrices. The infeasible loading matrices we would like to use are $\widehat{\Xi}_{t,0} = \mathrm{diag}(\tilde{l}_{t,0,1}, \ldots, \tilde{l}_{t,0,p})$ and $\widehat{\Psi}_{t,u,0} = \mathrm{diag}(l_{t,u,0,1}, \ldots, l_{t,u,0,p})$ in which

$$\tilde{l}_{t,0,j} = \left\| (Y - \nu_t(X))b_j(X)K(\frac{T-t}{h_1})h_1^{-1/2} \right\|_{\mathbb{P}_n,2}$$

and

$$l_{t,u,0,j} = \left\| (Y_u - \phi_{t,u}(X))b_j(X)K(\frac{T-t}{h_1})h_1^{-1/2} \right\|_{\mathbb{P}_n,2},$$

respectively. Since $\nu_t(\cdot)$ and $\phi_{t,u}(\cdot)$ are not known, we follow Belloni et al. (2017) and propose an iterative algorithm to obtain the feasible versions of the loading matrices. The statistical properties of the feasible loading matrices are summarized in Lemma A.3 in Appendix A.

Algorithm 3.1.

1. Let $\widehat{\Xi}_t^0 = \mathrm{diag}(\tilde{l}_{t,1}^0, \ldots, \tilde{l}_{t,p}^0)$ and $\widehat{\Psi}_{t,u}^0 = \mathrm{diag}(l_{t,u,1}^0, \ldots, l_{t,u,p}^0)$, where $\tilde{l}_{t,j}^0 = \|Yb_j(X)K(\frac{T-t}{h_1})h_1^{-1/2}\|_{\mathbb{P}_n,2}$ and $l_{t,u,j}^0 = \|Y_u b_j(X)K(\frac{T-t}{h_1})h_1^{-1/2}\|_{\mathbb{P}_n,2}$. Using $\widehat{\Xi}_t^0$ and $\widehat{\Psi}_{t,u}^0$, we can compute $\hat{\gamma}_t^0$ and $\hat{\theta}_{t,u}^0$ by (3.4) and (3.5). Let $\widehat{\nu}_t^0(x) = b(x)'\hat{\gamma}_t^0$ and $\widehat{\phi}_{t,u}^0(x) = \Lambda(b(x)'\hat{\theta}_{t,u}^0)$ for $x = X_1, \ldots, X_n$.

2. For $k = 1, \ldots, K$ for some fixed positive integer K, we compute $\widehat{\Xi}_t^k = \mathrm{diag}(\tilde{l}_{t,1}^k, \ldots, \tilde{l}_{t,p}^k)$ and $\widehat{\Psi}_{t,u}^k = \mathrm{diag}(l_{t,u,1}^k, \ldots, l_{t,u,p}^k)$, where

$$\tilde{l}_{t,j}^k = \left\| (Y - \widehat{\nu}_t^{k-1}(X))b_j(X)K(\frac{T-t}{h_1})h_1^{-1/2} \right\|_{\mathbb{P}_n,2}$$

and

$$l_{t,u,j}^k = \left\| (Y_u - \widehat{\phi}_{t,u}^{k-1}(X))b_j(X)K(\frac{T-t}{h_1})h_1^{-1/2} \right\|_{\mathbb{P}_n,2}.$$

Using $\widehat{\Xi}_t^k$ and $\widehat{\Psi}_{t,u}^k$, we can compute $\hat{\gamma}_t^k$ and $\hat{\theta}_{t,u}^k$ by (3.4) and (3.5). Let $\widehat{\nu}_t^k(x) = b(x)'\hat{\gamma}_t^k$ and $\widehat{\phi}_{t,u}^k(x) = \Lambda(b(x)'\hat{\theta}_{t,u}^k)$ for $x = X_1, \ldots, X_n$. The final penalty loading matrices $\widehat{\Xi}_t^K$ and $\widehat{\Psi}_{t,u}^K$ will be used for $\widehat{\Xi}_t$ and $\widehat{\Psi}_{t,u}$ in (3.4) and (3.5).

Let $\widetilde{\mathcal{S}}_t^\mu$ and $\widetilde{\mathcal{S}}_{t,u}$ contain the supports of $\hat{\gamma}_t$ and $\hat{\theta}_{t,u}$, respectively, such that $|\widetilde{\mathcal{S}}_t^\mu| \lesssim \sup_{t \in \mathcal{T}} \|\widehat{\gamma}_t\|_0$, and $|\widetilde{\mathcal{S}}_{t,u}| \lesssim \sup_{(t,u) \in \mathcal{TU}} \|\widehat{\theta}_{t,u}\|_0$. For each $(t, u) \in \mathcal{TU} := \mathcal{T} \times \mathcal{U}$ where $\mathcal{T}$ and $\mathcal{U}$ are compact subsets of the supports of T and Y, respectively, the post-Lasso estimator of γ_t and $\theta_{t,u}$ based on the set of covariates $\widetilde{\mathcal{S}}_t^\mu$ and $\widetilde{\mathcal{S}}_{t,u}$ are defined as

$$\tilde{\gamma}_t \in \arg\min_\gamma \sum_{i=1}^n (Y_i - b(X_i)'\gamma)^2 K(\frac{T_i - t}{h_1}), \quad s.t. \quad \mathrm{Supp}(\gamma) \in \widetilde{\mathcal{S}}_t^\mu,$$

and

$$\tilde{\theta}_{t,u} \in \arg\min_\theta \sum_{i=1}^n M(1\{Y_i \leq u\}, X_i; \theta) K(\frac{T_i - t}{h_1}), \quad s.t. \quad \mathrm{Supp}(\theta) \in \widetilde{\mathcal{S}}_{t,u}.$$

The post-Lasso estimators of $\nu_t(x)$ and $\phi_{t,u}(X)$ are given by $\widetilde{\nu}_t(X) = b(X)'\tilde{\gamma}_t$ and $\widetilde{\phi}_{t,u}(X) = \Lambda(b(X)'\tilde{\theta}_{t,u})$, respectively.

3.1.2. Conditional density estimation

Following Belloni et al. (2019), we propose to first estimate $F_t(X)$, the conditional CDF of T given X, by the (logistic) distributional Lasso regression studied in Belloni et al. (2017) and then take the numerical derivative. Following Belloni et al. (2017), we approximate $F_t(X)$ by a logistic CDF $\Lambda(b(X)'\beta_t)$ and the approximation error is denoted as $r_t^F(x) = F_t(x) - \Lambda(b(x)'\beta_t)$. We estimate β_t by $\hat{\beta}_t$, which is computed as

$$\hat{\beta}_t = \arg\min_\beta \frac{1}{n} \sum_{i=1}^n M(1\{T_i \leq t\}, X_i; \beta) + \frac{\tilde{\lambda}}{n} \|\widehat{\Psi}_t \beta\|_1 \quad \text{and} \quad \hat{F}_t(x) = \Lambda(b(x)'\hat{\beta}_t), \tag{3.6}$$

where $M(\cdot)$ is the logistic likelihood as defined previously, the penalty

$$\tilde{\lambda} = 1.1\Phi^{-1}(1 - \gamma/\{p \vee nh_1\})n^{1/2}$$

is slightly modified from but of the same order of magnitude as λ^* used in Belloni et al. (2017) and Belloni et al. (2019), for some $\gamma \to 0$ specified in Section 5, and the penalty loading $\hat{\Psi}_t$ is estimated in Algorithm 3.2, which is also due to Belloni et al. (2017):

Algorithm 3.2.

1. Let $\widehat{\Psi}_t^0 = \text{diag}(l_{t,1}^0, \ldots, l_{t,p}^0)$ where $l_{t,j}^0 = \|1\{T \leq t\}b_j(X)\|_{\mathbb{P}_n,2}$. Using $\widehat{\Psi}_t^0$, we can compute $\hat{\beta}_t^0$ and $\hat{F}_t(X)$ by the (logistic) distributional Lasso regression.
2. For $k = 1, \ldots, K$, we compute $\widehat{\Psi}_t^k = \text{diag}(l_{t,1}^k, \ldots, l_{t,p}^k)$ where

$$l_{t,j}^k = \left\| \left(1\{T \leq t\} - \hat{F}_t^{k-1}(X) \right) b_j(X) \right\|_{\mathbb{P}_n,2}.$$

Using $\widehat{\Psi}_t^k$, we can compute $\hat{\beta}_t^k$ and $\hat{F}_t^k(X)$ by the (logistic) distributional Lasso regression. The final penalty loading matrix $\widehat{\Psi}_t^K$ will be used as $\widehat{\Psi}_t$ in (3.6).

Then, $f_t(X)$, the conditional density of $T = t$ given X is computed as

$$\hat{f}_t(X) = \frac{\hat{F}_{t+h_1}(X) - \hat{F}_{t-h_1}(X)}{2h_1},$$

where h_1 is the first-stage bandwidth.

3.1.3. Asymptotic properties of the first stage estimators

To study the asymptotic properties of the first stage estimators, we need some assumptions.

Assumption 2. Let $\mathcal{TU}$ be a compact subset of the support of (T, Y) and $\mathcal{X}$ be the support of X.

1. The sample $\{Y_i, T_i, X_i\}_{i=1}^n$ is i.i.d.
2. $\| \max_{j \leq p} |b_j(X)| \|_{\mathbb{P},\infty} \leq \zeta_n$ and $\underline{C} \leq \mathbb{E}b_j(X)^2 \leq 1/\underline{C}$ $j = 1, \ldots, p$.
3. $\sup_{(t,u) \in \mathcal{TU}} \max(\|\gamma_t\|_0, \|\beta_t\|_0, \|\theta_{t,u}\|_0) \leq s$ for some s which possibly depends on the sample size n.
4. $\sup_{t \in \mathcal{T}} \|r_t^F(X)\|_{\mathbb{P}_n,2} = O_p((s\log(p \vee n)/(n))^{1/2})$ and

$$\sup_{(t,u) \in \mathcal{TU}} \left[\|r_{t,u}^v(X)K(\frac{T-t}{h_1})^{1/2}\|_{\mathbb{P}_n,2} + \|r_{t,u}^\phi(X)K(\frac{T-t}{h_1})^{1/2}\|_{\mathbb{P}_n,2} \right] = O_p((s\log(p \vee n)/n)^{1/2}).$$

5. $\sup_{t \in \mathcal{T}} \|r_t^F(X)\|_{\mathbb{P},\infty} = O((\log(p \vee n)s^2\zeta_n^2/(n))^{1/2})$ and

$$\sup_{(t,u) \in \mathcal{TU}} \left[\|r_{t,u}^v(X)\|_{\mathbb{P},\infty} + \|r_{t,u}^\phi(X)\|_{\mathbb{P},\infty} \right] = O((\log(p \vee n)s^2\zeta_n^2/(nh_1))^{1/2}).$$

6. $f_t(x)$ is second-order differentiable w.r.t. t with bounded derivatives uniformly over $(t, x) \in \mathcal{TX}$, where $\mathcal{T}$ is a compact subset of the support of T and $\mathcal{X}$ is the support of X.
7. $\zeta_n^2 s^2 \ell_n^2 \log(p \vee n)/(nh_1) \to 0$, $nh_1^5/(\log(p \vee n)) \to 0$.

Assumption 2.1 is common for cross-sectional observations. Assumption 2.2 is the same as Assumption 6.1(a) in Belloni et al. (2017). Assumption 2.3 requires that $v_t(x)$, $\phi_{t,u}(x)$, and $F_t(x)$ are approximately sparse, i.e., they can be well-approximated by using at most s elements of $b(x)$. This approximate sparsity condition is common in the literature on high-dimensional data (see, e.g., Belloni et al. (2017)). Assumptions 2.4 and 2.5 specify how well the approximations are in terms of $L_{\mathbb{P}_n,2}$ and $L_{\mathbb{P},\infty}$ norms. The exact rate for $r_t^F(X)$ follows Belloni et al. (2017). The rates for $r_{t,u}^v(X)$ and $r_{t,u}^\phi(X)$ are different from that for $r_t^F(X)$ because their approximations are local in $T = t$. If the models for $v_t(\cdot)$, $\phi_{t,u}(\cdot)$, and $F_t(\cdot)$ are correctly specified and exactly sparse, i.e., the coefficients for all but s regressors are zero, then there are no approximate errors. This implies $r_t^F(\cdot)$, $r_{t,u}^v(\cdot)$, and $r_{t,u}^\phi(\cdot)$ equal to zero so that Assumptions 2.4 and 2.5 hold automatically. In the sieve estimation, X is finite dimensional and $b(X)$ is just a sequence of sieve bases of X. Then $r_t^F(\cdot)$, $r_{t,u}^v(\cdot)$, and $r_{t,u}^\phi(\cdot)$ are the sieve approximation bias. Assumptions 2.3 and 2.4 can be verified under some smoothness conditions (see, e.g., Chen (2007)). Therefore, Assumptions 2.4 and 2.5 are in spirit close to the smoothness condition. Assumption 2.6 is the smoothness of the true density, which is needed for the theoretical analysis of the numerical derivative. Because $\mathcal{T}$ needs not be the whole support of T, this condition is plausible. In a simple case, if $T = \mu(X) + U$, $|\mu(x)|$ is bounded uniformly over $x \in \mathcal{X}$, and U is independent of X and logistically distributed, then this condition holds. Assumption 2.7 imposes conditions on the rates at which s, ζ_n, and p grow with sample size n. It ensures that the first stage nuisance parameters are estimated with sufficient accuracy. In particular, we require $s^2/(nh_1) \to 0$. Comparing with the condition that $s^2/n \to 0$ imposed in Belloni et al. (2017), our condition reflects the local nature of our estimation procedure in the sense that our effective sample size is of order of magnitude nh_1.

Assumption 3.

1. $K(\cdot)$ is a symmetric probability density function (PDF) with

$$\int uK(u)du = 0, \quad \text{and} \quad \kappa_2 := \int u^2K(u)du < \infty.$$

   There exists a positive constant $\overline{C}_K$ such that $\sup_u u^l K(u) \le \overline{C}_K$ for $l = 0, 1$.
2. There exists some positive constant $\underline{C} < 1$ such that $\underline{C} \le f_t(x) \le 1/\underline{C}$ uniformly over $(t, x) \in \mathcal{TX}$.
3. $\nu_t(x)$ and $\phi_{t,u}(x)$ are three times differentiable w.r.t. t, with all three derivatives being bounded uniformly over $(t, x, u) \in \mathcal{TXU}$.
4. For the same $\underline{C}$ as above, $\underline{C} \le \mathbb{E}(Y_u(t)|X = x) \le 1 - \underline{C}$ uniformly over $(t, x, u) \in \mathcal{TXU} := \mathcal{TX} \times \mathcal{U}$.

Assumption 3.1 holds for many kernel functions, e.g., uniform and Gaussian kernels. Since $f_T(X)$ was referred to as the generalized propensity by Hirano and Imbens (2004), Assumption 3.2 is analogous to the overlapping support condition commonly assumed in the treatment effect literature; see, e.g., Hirano et al. (2003) and Firpo (2007). Since the conditional density also has the sparsity structure as assumed in Assumption 2, at most s members of X's affect the conditional density, which makes Assumption 3.2 more plausible. It is possible to modify our proof to allow for the infimum of the density to decay to zero slowly, as sample size increases. On the other hand, we rule out the case that

$$\inf_{(t,x)\in\mathcal{TX}} f_t(x) = 0. \tag{3.7}$$

Recall that $f_t(X)$ is referred to as the generalized propensity score by Hirano and Imbens (2004). It is the counterpart of propensity score when the treatment status is binary. Khan and Tamer (2010) have already pointed out that, when the propensity score can approach zero, the semi-parametric efficiency bound for the average treatment effect is zero, which implies that there does not exist a regular estimator for the mean potential outcomes. Similarly, if (3.7) holds, the identification of $\mu(t)$ and $\alpha(t, u)$ is not only based on the "thin set" (a neighborhood of t) but also at infinity ($f_t(X)$ approaches zero). Sample selection and special regressor models are also identified at infinity. As shown in Andrews and Schafgans (1998) and Khan and Tamer (2010), their estimations are difficult and involve asymptotic trimming. In our case, asymptotic trimming means that we need to truncate the density estimator from below by $\varepsilon = \varepsilon_n \to 0$ as $n \to \infty$. Similar to Tan and Zhang (2018), ε_n controls the trade-off between bias and variance, which play key roles for inference. Such an extension is left as a future research topic. Assumption 3.3 imposes some smoothness conditions that are widely assumed in the nonparametric kernel literature. Assumption 3.4 holds if $\mathcal{XU}$ is compact.

Assumption 4. There exists a sequence $\ell_n \to \infty$ such that, with probability approaching one,

$$0 < \kappa' \le \inf_{\delta\neq 0, \|\delta\|_0\le s\ell_n} \frac{\|b(X)'\delta\|_{\mathbb{P}_n,2}}{\|\delta\|_2} \le \sup_{\delta\neq 0, \|\delta\|_0\le s\ell_n} \frac{\|b(X)'\delta\|_{\mathbb{P}_n,2}}{\|\delta\|_2} \le \kappa'' < \infty.$$

Assumption 4 is the restricted eigenvalue condition commonly assumed in the high-dimensional data literature. Based on Bickel et al. (2009),

$$\inf_{\delta\neq 0, \|\delta\|_0\le s\ell_n} \frac{\|b(X)'\delta\|_{\mathbb{P}_n,2}}{\|\delta\|_2} \quad \text{and} \quad \sup_{\delta\neq 0, \|\delta\|_0\le s\ell_n} \frac{\|b(X)'\delta\|_{\mathbb{P}_n,2}}{\|\delta\|_2}$$

are the minimal and maximal eigenvalues of Gram submatrices formed by any $s\ell_n$ components of $b(X)$. Because $p \gg n$, the matrix $b(X)'b(X)$ is not invertible. However, because $s\ell_n \ll n$, Assumption 4 implies that the Gram submatrices can still be invertible. We refer interested readers to Bickel et al. (2009) for more details and Bühlmann and van de Geer (2011) for a textbook treatment.

Since there is a kernel in the Lasso objective functions in (3.4) and (3.5), the asymptotic properties of $\hat{\gamma}_t$ and $\hat{\theta}_{t,u}$ cannot be established by directly applying the results in Belloni et al. (2017). The key missing piece is the following local version of the compatibility condition. Let $\mathcal{S}_{t,u}$ be an arbitrary subset of $\{1, \ldots, p\}$ such that $\sup_{(t,u)\in\mathcal{TU}} |\mathcal{S}_{t,u}| \le s$ and $\Delta_{c,t,u} = \{\delta : \|\delta_{\mathcal{S}_{t,u}^c}\|_1 \le c\|\delta_{\mathcal{S}_{t,u}}\|_1\}$ for some $c < \infty$ independent of (t, u).

Lemma 3.1. *If Assumptions 1–4 hold, then there exists $\underline{\kappa} = \kappa'\underline{C}^{1/2}/4 > 0$ such that, w.p.a.1,*

$$\inf_{(t,u)\in\mathcal{TU}} \inf_{\delta\in\Delta_{c,t,u}} \frac{\|b(X)'\delta K(\frac{T-t}{h_1})^{1/2}\|_{\mathbb{P}_n,2}}{\|\delta_{\mathcal{S}_{t,u}}\|_2\sqrt{h_1}} \ge \underline{\kappa}.$$

Note $\mathcal{S}_{t,u}$ in Lemma 3.1 is either the support of $\theta_{t,u}$ or the support of γ_t. For the latter case, the index u is not needed. We refer to Lemma 3.1 as the local compatibility condition because (1) there is a kernel function implementing the localization; and (2) by the Cauchy inequality, Lemma 3.1 implies

$$\inf_{(t,u)\in\mathcal{TU}} \inf_{\delta\in\Delta_{c,t,u}} \frac{\sqrt{s}\|b(X)'\delta K(\frac{T-t}{h_1})^{1/2}\|_{\mathbb{P}_n,2}}{\|\delta_{\mathcal{S}_{t,u}}\|_1\sqrt{h_1}} \ge \underline{\kappa}.$$

Bickel et al. (2009, Lemma 4.2) show that, under Assumption 4, we have the following compatibility condition:

$$\inf_{(t,u)\in\mathcal{TU}}\inf_{\delta\in\Delta_{c,t,u}}\frac{\sqrt{s}\|b(X)'\delta\|_{\mathbb{P}_n,2}}{\|\delta_{\mathcal{S}_{t,u}}\|_1}\geq\inf_{(t,u)\in\mathcal{TU}}\inf_{\delta\in\Delta_{c,t,u}}\frac{\|b(X)'\delta\|_{\mathbb{P}_n,2}}{\|\delta_{\mathcal{S}_{t,u}}\|_2}\geq\underline{\kappa},\tag{3.8}$$

which is the key convertibility condition used in high-dimensional analysis. We refer interested readers to Bühlmann and van de Geer (2011, Equation 6.4), the remarks after that, and Bühlmann and van de Geer (2011, Section 6.13) for more detailed discussions and further references. Under Assumption 4 and some regularity conditions assumed in the paper, Lemma 3.1 establishes a local version of (3.8). Based on Lemma 3.1, we can establish the following asymptotic probability bounds for the first stage estimators.

Theorem 3.1. *Suppose Assumptions 1–2, 3.1–3.3, and 4 hold. Then*

$$\sup_{t\in\mathcal{T}}\|(\widehat{v}_t(X)-v_t(X))\|_{\mathbb{P}_n,2}=O_p(\ell_n(\log(p\vee n)s)^{1/2}(nh_1)^{-1/2}),$$

$$\sup_{t\in\mathcal{T}}\|\widehat{v}_t(X)-v_t(X)\|_{\mathbb{P},\infty}=O_p(\ell_n(\log(p\vee n)s^2\zeta_n^2/(nh_1))^{1/2}),$$

$$\sup_{t\in\mathcal{T}}\|(\widetilde{v}_t(X)-v_t(X))\|_{\mathbb{P}_n,2}=O_p(\ell_n(\log(p\vee n)s)^{1/2}(nh_1)^{-1/2}),$$

$$\sup_{t\in\mathcal{T}}\|\widetilde{v}_t(X)-v_t(X)\|_{\mathbb{P},\infty}=O_p(\ell_n(\log(p\vee n)s^2\zeta_n^2/(nh_1))^{1/2}),$$

and $\sup_{t\in\mathcal{T}}\|\hat{\gamma}_t\|_0=O_p(s)$. *If in addition, Assumption 3.4 holds, then*

$$\sup_{(t,u)\in\mathcal{TU}}\|(\widehat{\phi}_{t,u}(X)-\phi_{t,u}(X))\|_{\mathbb{P}_n,2}=O_p(\ell_n(\log(p\vee n)s)^{1/2}(nh_1)^{-1/2}),$$

$$\sup_{(t,u)\in\mathcal{TU}}\|\widehat{\phi}_{t,u}(X)-\phi_{t,u}(X)\|_{\mathbb{P},\infty}=O_p(\ell_n(\log(p\vee n)s^2\zeta_n^2/(nh_1))^{1/2}),$$

$$\sup_{(t,u)\in\mathcal{TU}}\|(\widetilde{\phi}_{t,u}(X)-\phi_{t,u}(X))\|_{\mathbb{P}_n,2}=O_p(\ell_n(\log(p\vee n)s)^{1/2}(nh_1)^{-1/2}),$$

$$\sup_{(t,u)\in\mathcal{TU}}\|\widetilde{\phi}_{t,u}(X)-\phi_{t,u}(X)\|_{\mathbb{P},\infty}=O_p(\ell_n(\log(p\vee n)s^2\zeta_n^2/(nh_1))^{1/2}),$$

and $\sup_{(t,u)\in\mathcal{TU}}\|\hat{\theta}_{t,u}\|_0=O_p(s)$.

Several comments are in order. First, due to the nonlinearity of the logistic link function, Assumption 3.4 is needed for deriving the asymptotic properties of the penalized local MLE estimators $\widehat{\phi}_{t,u}(x)$ and $\widetilde{\phi}_{t,u}(x)$. Second, the $L_{\mathbb{P}_n,2}$ bounds in Theorem 3.1 are faster than $(nh_2)^{-1/4}$ by Assumption 5. This implies the estimators are sufficiently accurate so that in the second stage, their second and higher order impacts are asymptotically negligible. Last, the numbers of nonzero coordinates of $\hat{\gamma}_t$ and $\hat{\theta}_{t,u}$ determine the complexity of our first stage estimators, which are uniformly controlled with a high probability.

For the conditional density estimation, we have the following results.

Theorem 3.2. *Suppose Assumptions 1–2, 3.1–3.3, and 4 hold. Then*

$$\sup_{t\in\mathcal{T}}\|\hat{f}_t(X)-f_t(X)\|_{\mathbb{P}_n,2}=O_p((\log(p\vee n)s/n)^{1/2}h_1^{-1}),$$

$$\sup_{t\in\mathcal{T}}\|\hat{f}_t(X)-f_t(X)\|_{\mathbb{P},\infty}=O_p((\log(p\vee n)s^2\zeta_n^2/n)^{1/2}h_1^{-1}),$$

$$\sup_{t\in\mathcal{T}}\|\tilde{f}_t(X)-f_t(X)\|_{\mathbb{P}_n,2}=O_p((\log(p\vee n)s/n)^{1/2}h_1^{-1}),$$

$$\sup_{t\in\mathcal{T}}\|\tilde{f}_t(X)-f_t(X)\|_{\mathbb{P},\infty}=O_p((\log(p\vee n)s^2\zeta_n^2/n)^{1/2}h_1^{-1}),$$

and $\sup_{t\in\mathcal{T}}\|\hat{\beta}_t\|_0=O_p(s)$.

The rates of convergence in Theorem 3.2 are the same as those derived in Belloni et al. (2019, Section 8).

3.2. The second stage estimation

Let $W=\{Y,T,X\}$ and $W_u=\{Y_u,T,X\}$. For three generic functions $\breve{v}(\cdot)$, $\breve{\phi}(\cdot)$ and $\breve{f}(\cdot)$ of X, denote

$$\Pi_t'(W,\breve{v},\breve{f})=\frac{(Y-\breve{v}(X))}{\breve{f}(X)h_2}K(\frac{T-t}{h_2})+\breve{v}(X)$$

and

$$\Pi_{t,u}(W_u, \check{\phi}, \check{f}) = \frac{(Y_u - \check{\phi}(X))}{\check{f}(X)h_2}K(\frac{T-t}{h_2}) + \check{\phi}(X).$$

Then the estimators $\hat{\mu}(t)$ and $\hat{\alpha}(t, u)$ can be written as

$$\hat{\mu}(t) = \mathbb{P}_n \Pi'_t(W, \overline{\nu}_t, \overline{f}) \quad \text{and} \quad \hat{\alpha}(t, u) = \mathbb{P}_n \Pi_{t,u}(W_u, \overline{\phi}_{t,u}, \overline{f}),$$

where $\overline{\nu}_t(\cdot)$, $\overline{\phi}_{t,u}(\cdot)$, and $\overline{f}_t(\cdot)$ are either the Lasso estimators (i.e., $\widehat{\nu}_t(\cdot)$, $\widehat{\phi}_{t,u}(\cdot)$, and $\hat{f}_t(\cdot)$) or the post-Lasso estimators (i.e., $\widetilde{\nu}_t(\cdot)$, $\widetilde{\phi}_{t,u}(\cdot)$, and $\tilde{f}_t(\cdot)$) as defined in Section 3.1.

Assumption 5. Let $h_2 = C_2 n^{-H_2}$ for some positive constant C_2.

1. $H_2 \in [1/5, 1/3)$, $\log^2(n)s^2 \log^2(p \vee n)/(nh_2) \to 0$, and $\ell_n^2 s^2 \log^2(p \vee n)/(nh_1^2) \to 0$, and $\ell_n^2 s^2 \log^2(p \vee n)h_2/(nh_1^3) \to 0$.
2. $H_2 \in (1/4, 1/3)$, $\log^2(n)s^2 \log^2(p \vee n)/(nh_2^2) \to 0$, $\ell_n^2 s^2 \log^2(p \vee n)/(nh_1^2 h_2) \to 0$, and $\ell_n^2 s^2 \log^2(p \vee n)/(nh_1^3) \to 0$.

It is possible to allow for different sparsity levels for the infinite-dimensional nuisance functions (i.e., $\nu_t(\cdot)$, $\phi_{t,u}(\cdot)$, $f_t(\cdot)$) with more complicate notation. Then, following Farrell (2015),[6] we conjecture that Assumption 5 can be relaxed to

$$\sup_{t \in \mathcal{T}} \|\overline{\nu}_t(\cdot) - \nu_t(\cdot)\|_{P_n,2} \|\overline{f}_t(\cdot) - f_t(\cdot)\|_{P_n,2} = o_p((nh_2)^{-1/2})$$

and

$$\sup_{(t,u) \in \mathcal{TU}} \|\overline{\phi}_{t,u}(\cdot) - \nu_t(\cdot)\|_{P_n,2} \|\overline{f}_t(\cdot) - f_t(\cdot)\|_{P_n,2} = o_p((nh_2)^{-1/2}).$$

In addition, by implementing the cross-fitting technique proposed by Chernozhukov et al. (2018), it is possible to apply other machine-learning method other than Lasso for the estimation of nuisance functions and further relax the requirement on the complexity of the function spaces their estimators live in. Such an extension is left as a future research topic.

Theorem 3.3. *Suppose Assumptions 1–4 and 5.1 hold. Then*

$$\hat{\alpha}(t, u) - \alpha(t, u) = (\mathbb{P}_n - \mathbb{P})\Pi_{t,u}(W_u, \phi_{t,u}, f_t) + \mathcal{B}_\alpha(t, u)h_2^2 + R_n(t, u),$$

where

$$\mathcal{B}_\alpha(t, u) = \frac{\kappa_2}{2}\left[\mathbb{E}\left(\partial_t^2 \phi_{t,u}(X) + \frac{2\partial_t \phi_{t,u}(X)\partial_t f_t(X)}{f_t(X)}\right)\right],$$

$\kappa_2 = \int u^2 K(u)du$ *and* $\sup_{(t,u) \in \mathcal{TU}} |R_n(t, u)| = o_p((nh_2)^{-1/2})$. *If Assumption 5.1 is replaced by Assumption 5.2, then*

$$\sup_{(t,u) \in \mathcal{TU}} |\mathcal{B}_\alpha(t, u)h_2^2 + R_n(t, u)| = o_p(n^{-1/2}).$$

Theorem 3.3 presents the Bahadur representations of the nonparametric estimator $\hat{\alpha}(t, u)$ with a uniform control on the remainder term. For most purposes (e.g., to obtain the asymptotic distributions of the intermediate estimator or to obtain the results below), Assumption 5.1 is sufficient. Occasionally, one needs to impose Assumption 5.2 to have a better control on the remainder term, say, when one conducts an L_2-type specification test. See the remark after Theorem 3.4. Under an extra moment condition on Y (e.g., $\mathbb{E}Y^q < \infty$ for some $q \geq 4$), one can establish similar results for the average effect estimator, i.e.,

$$\hat{\mu}(t) - \mu(t) = (\mathbb{P}_n - \mathbb{P})\Pi'_t(W, \nu_t, f_t) + \mathcal{B}_\mu(t)h_2^2 + R'_n(t),$$

where $\sup_{t \in \mathcal{T}} |R'_n(t)| = o_p((nh_2)^{-1/2})$ and

$$\mathcal{B}_\mu(t) = \frac{\kappa_2}{2}\left[\mathbb{E}\left(\partial_t^2 \nu_t(X) + \frac{2\partial_t \nu_t(X)\partial_t f_t(X)}{f_t(X)}\right)\right].$$

If Assumption 5.1 is replaced by Assumption 5.2, then

$$\sup_{t \in \mathcal{T}} |\mathcal{B}_\mu(t)h_2^2 + R'_n(t)| = o_p(n^{-1/2}).$$

---

[6] We thank a referee for the reference and suggestion.

3.3. The third stage estimation

Recall that $q_\tau(t)$ denotes the τ-th quantile of $Y(t)$, which is the inverse of $\alpha(t, u)$ w.r.t. u. We propose to estimate $q_\tau(t)$ by $\hat{q}_\tau(t)$ where $\hat{q}_\tau(t) = \inf\{u : \hat{\alpha}^r(t, u) \geq \tau\}$ and $\hat{\alpha}^r(t, u)$ is the rearrangement of $\hat{\alpha}(t, u)$.

We rearrange $\hat{\alpha}(t, u)$ to make it monotonically increasing in $u \in \mathcal{U}$. Following Chernozhukov et al. (2010), for a generic function $Q(\cdot)$, we define $\overline{Q} = Q \circ \psi^\leftarrow$ where ψ can be any increasing bijective mapping: $\mathcal{U} \mapsto [0, 1]$ and $\psi^\leftarrow$ is the inverse of ψ. Then the rearrangement $\overline{Q}^r$ of $\overline{Q}$ is defined as

$$\overline{Q}^r(u) = F^\leftarrow(u) = \inf\{y : F(y) \geq u\},$$

where $F(y) = \int_0^1 1\{\overline{Q}(u) \leq y\}du$. The rearrangement Q^r for Q is $Q^r = \overline{Q}^r \circ \psi(u)$.

The rearrangement and inverse are two functionals operating on the process

$$\{\hat{\alpha}(t, u) : (t, u) \in \mathcal{TU}\}$$

and are shown to be Hadamard differentiable by Chernozhukov et al. (2010) and van der Vaart and Wellner (1996), respectively. However, by Theorem 3.3,

$$\sup_{(t,u)\in\mathcal{TU}} (nh_2)^{1/2}(\hat{\alpha}(t, u) - \alpha(t, u)) = O_p(\log^{1/2}(n)),$$

which is not asymptotically tight. Therefore, the standard functional delta method used in Chernozhukov et al. (2010) and van der Vaart and Wellner (1996) is not directly applicable. The next theorem overcomes this difficulty and establishes the linear expansion of the quantile estimator. Denote $\mathcal{TI}$, $\{q_\tau(t) : \tau \in \mathcal{I}\}^\varepsilon$, $\overline{\{q_\tau(t) : \tau \in \mathcal{I}\}^\varepsilon}$, and $\mathcal{U}_t$ as $\mathcal{T} \times \mathcal{I}$, the ε-enlarged set of $\{q_\tau(t) : \tau \in \mathcal{I}\}$, the closure of $\{q_\tau(t) : \tau \in \mathcal{I}\}^\varepsilon$, and the projection of $\mathcal{TU}$ on $T = t$, respectively.

Theorem 3.4. *Suppose that Assumptions 1–4 and 5.1 hold. If $\overline{\{q_\tau(t) : \tau \in \mathcal{I}\}^\varepsilon} \subset \mathcal{U}_t$ for any $t \in \mathcal{T}$, then*

$$\hat{q}_\tau(t) - q_\tau(t) = -(\mathbb{P}_n - \mathbb{P})\frac{\Pi_{t,u}(W_{q_\tau(t)}, \phi_{t,q_\tau(t)}, f_t)}{f_{Y(t)}(q_\tau(t))} - \beta_q(t, \tau)h_2^2 + R_n^q(t, \tau),$$

where $f_{Y(t)}$ is the density of $Y(t)$, $\beta_q(t, \tau) = \frac{\beta_\alpha(t, q_\tau(t))}{f_{Y(t)}(q_\tau(t))}$, and $\sup_{(t,\tau)\in\mathcal{TI}} R_n^q(t, \tau) = o_p((nh_2)^{-1/2})$. If Assumption 5.1 is replaced by Assumption 5.2, then

$$\sup_{(t,\tau)\in\mathcal{TI}} \left(|R_n^q(t, \tau)| + \left|\beta_q(t, \tau)\right|\right) h_2^2 = o_p(n^{-1/2}).$$

Under Assumption 5.2, the remainder term $R_n^q(t, \tau)$ is $o_p(n^{-1/2})$ uniformly in $(t, \tau) \in \mathcal{TI}$. This result is needed if one wants to establish an L_2-type specification test for $q_\tau(t)$. For example, one may be interested in testing the null hypotheses of the quantile partial derivative being homogeneous across treatment. In this case, the null hypothesis can be written as

$$H_0 : q_\tau(t) = \beta_0(\tau) + \beta_1(\tau)t \text{ for all } (t, \tau) \in \mathcal{TI},$$

and the alternative hypothesis is the negation of H_0. One way to conduct a consistent test for the above hypothesis is to employ the residuals of the linear regression of $\hat{q}_\tau(T_i)$ on T_i to construct the test statistic $\Upsilon_n(\tau)$, i.e.,

$$\Upsilon_n(\tau) = \frac{1}{n}\sum_{i=1}^n (\hat{q}_\tau(T_i) - \hat{\beta}_0 - \hat{\beta}_1 T_i)^2 1\{T_i \in \mathcal{T}\},$$

where $(\hat{\beta}_0, \hat{\beta}_1)$ are the linear coefficient estimators. This type of specification test has been previously studied by Su and Chen (2013), Lewbel et al. (2015), Su et al. (2015), Hoderlein et al. (2016), and Su and Hoshino (2016) in various contexts. One can follow them and apply the results in Theorem 3.4 to study the asymptotic distribution of $\Upsilon_n(\tau)$ for each τ. In addition, one can also consider either an integrated or a sup-version of $\Upsilon_n(\tau)$ and then study its asymptotic properties. For brevity we do not study such a specification test in this paper.

Given the estimators $\hat{\mu}(t)$ and $\hat{q}_\tau(t)$, we can run local linear regressions of $\hat{\mu}(T_i)$ and $\hat{q}_\tau(T_i)$ on $(1, T_i - t)$ and obtain estimators $\check{\beta}^1(t)$ and $\hat{\beta}_\tau^1(t)$ of $\partial\mu(t)$ and $\partial_t q_\tau(t)$, respectively, as estimators of the linear coefficients in the local linear regression.[7] Specifically, we define

$$(\check{\beta}^0(t), \check{\beta}^1(t)) = \arg\max_{\beta^0,\beta^1} \sum_{i=1}^n (\hat{\mu}(T_i) - \beta^0 - \beta^1(T_i - t))^2 K(\frac{T_i - t}{h_2})$$

and

$$(\hat{\beta}_\tau^0(t), \hat{\beta}_\tau^1(t)) = \arg\max_{\beta^0,\beta^1} \sum_{i=1}^n (\hat{q}_\tau(T_i) - \beta^0 - \beta^1(T_i - t))^2 K(\frac{T_i - t}{h_2}),$$

---

[7] Alternatively, one can consider the local quadratic or cubic regression.

where h_2 is the second-stage bandwidth. It is possible to use a third bandwidth h_3 in this step. Results similar to Theorem 3.5 still holds if $h_3/h_2 = O(1)$. Note that the usual optimal bandwidth for the kernel estimator of the derivative is $O(n^{-1/7})$. However, because $h_2 = O(n^{-1/5})$, the requirement that $h_3/h_2 = O(1)$ implies the optimal bandwidth is not achievable. The key reason is that, unlike the usual local linear regression, we need to plug in the estimates of $\mu(\cdot)$ and $q_\tau(\cdot)$. For simplicity, we just take $h_3 = h_2$.

The following theorem shows the asymptotic properties of $\breve{\beta}^1(t)$ and $\hat{\beta}^1_\tau(t)$.

Theorem 3.5. *Suppose Assumptions 1–4, and 5.1. If $\overline{\{q_\tau(t) : \tau \in \mathcal{I}\}^\varepsilon} \subset \mathcal{U}_t$ for any $t \in \mathcal{T}$, then*

$$\hat{\beta}^1_\tau(t) - \partial_t q_\tau(t) = -(\mathbb{P}_n - \mathbb{P})(\kappa_2 f_{Y(t)}(q_\tau(t)) f_t(X_j) h_2^2)^{-1} \left[Y_{q_\tau(t),j} - \phi_{t,q_\tau(t)}(X_j) \right] \overline{K}(\frac{T_j - t}{h_2}) + R_n^1(t, \tau),$$

where $\sup_{(t,\tau) \in \mathcal{T}\mathcal{I}} |R_n^1(t, \tau)| = o_p((nh_2^3)^{-1/2})$ *and* $\overline{K}(v) = \int wK(v - w)K(w)dw$.

Theorem 3.5 presents the Bahadur representation for $\hat{\beta}^1_\tau(t)$. Since it is the estimator for the first order derivative $\partial_t q_\tau(t)$, we can show that it converges to the true values at the $(nh_2^3)^{1/2}$-rate. Such a rate is common for kernel estimations of the first-order derivative of the conditional expectation, i.e., Li and Racine (2007, Theorem 2.10). If the addition moment condition for Y holds, the similar results hold for $\breve{\beta}^1(t)$, i.e.,

$$\breve{\beta}^1(t) - \partial_t \mu(t) = (\mathbb{P}_n - \mathbb{P})(\kappa_2 f_t(X_j) h_2^2)^{-1} \left[Y_j - \nu_t(X_j) \right] \overline{K}(\frac{T_j - t}{h_2}) + \breve{R}_n^1(t),$$

where $\sup_{t \in \mathcal{T}} |\breve{R}_n^1(t)| = o_p((nh_2^3)^{-1/2})$.

4. Inference

In this section, we study the inference for $\mu(t)$, $q_\tau(t)$, and $\partial_t q_\tau(t)$. We follow the lead of Belloni et al. (2017) and consider the weighted-bootstrap inference. Let $\{\eta_i\}_{i=1}^n$ be a sequence of i.i.d. random variables generated from the distribution of η such that it has sub-exponential tails and unit mean and variance.[8] For example, η can be a standard exponential random variable or a normal random variable with unit mean and standard deviation. We conduct the bootstrap inference based on the following procedure.

1. Obtain $\widehat{v}_t(x)$, $\widehat{\phi}_{t,u}(x)$, $\hat{f}_t(x)$, $\widetilde{v}_t(x)$, $\widetilde{\phi}_{t,u}(x)$ and $\tilde{f}_t(x)$ from the first stage.
2. For the bth bootstrap sample:

   - Generate $\{\eta_i\}_{i=1}^n$ from the distribution of η.
   - Compute

   $$\hat{\mu}^b(t) := \frac{1}{\sum_{i=1}^n \eta_i} \sum_{i=1}^n \eta_i \Pi'_t(W_i, \overline{v}_t, \overline{f}_t)$$

   and

   $$\hat{\alpha}^b(t, u) := \frac{1}{\sum_{i=1}^n \eta_i} \sum_{i=1}^n \eta_i \Pi_{t,u}(W_{ui}, \overline{\phi}_{t,u}, \overline{f}_t),$$

   where $(\overline{\phi}_{t,u}(\cdot), \overline{f}_t(\cdot))$ are either $(\widehat{\phi}_{t,u}(\cdot), \hat{f}_t(\cdot))$ or $(\widetilde{\phi}_{t,u}(\cdot), \tilde{f}_t(\cdot))$.
   - Rearrange $\hat{\alpha}^b(t, u)$ and obtain $\hat{\alpha}^{br}(t, u)$.
   - Invert $\hat{a}^{br}(t, u)$ w.r.t. u and obtain $\hat{q}^b_\tau(t) = \inf\{u : \hat{a}^{br}(t, u) \geq \tau\}$.
   - Compute $\breve{\beta}^{b1}(t)$ and $\hat{\beta}^{b1}_\tau(t)$ as the slope coefficients of local linear regressions of $\eta_i \hat{\mu}^b(T_i)$ on $(\eta_i, \eta_i(T_i - t))$ and $\eta_i \hat{q}^b_\tau(T_i)$ on $(\eta_i, \eta_i(T_i - t))$, respectively.

3. We repeat the above step for $b = 1, \ldots, B$ and obtain a bootstrap sample of

   $$\{\hat{\mu}^b(t), \hat{q}^b_\tau(t), \breve{\beta}^{b1}(t), \hat{\beta}^{b1}_\tau(t)\}_{b=1}^B.$$

4. Obtain $\widehat{Q}^\mu(\alpha)$, $\widehat{Q}^0(\alpha)$, $\widehat{Q}^{\mu 1}(\alpha)$, and $\widehat{Q}^1(\alpha)$ as the α-th quantile of the sequences $\{\hat{\mu}^b(t) - \hat{\mu}(t)\}_{b=1}^B$, $\{\hat{q}^b_\tau(t) - \hat{q}_\tau(t)\}_{b=1}^B$, $\{\breve{\beta}^{b1}(t) - \breve{\beta}^1(t)\}_{b=1}^B$, and $\{\hat{\beta}^{b1}_\tau(t) - \hat{\beta}^1_\tau(t)\}_{b=1}^B$, respectively.

The standard $100(1 - \alpha)\%$ percentile bootstrap confidence interval for $q_\tau(t)$ is

$$(\hat{Q}^0(\alpha/2) + \hat{q}_\tau(t), \hat{Q}^0(1 - \alpha/2) + \hat{q}_\tau(t)).$$

---

[8] A random variable η has sub-exponential tails if $P(|\eta| > x) \leq K \exp(-Cx)$ for every x and some constants K and C.

However, in our simulation study, we find that it slightly undercovers. Instead, we use the fact that normal CDF is symmetric and propose to use the modified percentile bootstrap confidence interval as follows:

$$(-\hat{Q}^{*0}(\alpha/2) + \hat{q}_\tau(t), \hat{Q}^{*0}(\alpha/2) + \hat{q}_\tau(t)),$$

where $\hat{Q}^{*0}(\alpha/2) = (-\hat{Q}^0(\alpha/2)) \vee \hat{Q}^0(1 - \alpha/2)$. We define $\widehat{Q}^{*\mu}(\alpha/2)$, $\widehat{Q}^{*\mu 1}(\alpha/2)$, and $\widehat{Q}^{*1}(\alpha/2)$ in the same manner. The following theorem summarizes the main results in this section.

Theorem 4.1. *Suppose that* Assumptions 1–4 *and* 5.1 *hold and* $nh_2^5 \to 0$. *Then*

$$\mathbb{P}(-\widehat{Q}^{*0}(\alpha/2) + \hat{q}_\tau(t) \leq q_\tau(t) \leq \widehat{Q}^{*0}(\alpha/2) + \hat{q}_\tau(t)) \to 1 - \alpha$$

and

$$\mathbb{P}(-\widehat{Q}^{*1}(\alpha/2) + \breve{\beta}_\tau^1(t) \leq \partial_t q_\tau(t) \leq \widehat{Q}^{*1}(\alpha/2) + \breve{\beta}_\tau^1(t)) \to 1 - \alpha.$$

Theorem 4.1 implies that, via under-smoothing, the $100(1 - \alpha)\%$ bootstrap confidence intervals for $q_\tau(t)$ and $\partial_t q_\tau(t)$ have the correct asymptotic coverage probability $1 - \alpha$. We need to under-smooth because, regardless of under-smoothing, the bootstrap estimator always center around the original estimator without the asymptotic bias. With more complicated notations and the arguments of strong approximation in Chernozhukov et al. (2014a) and Chernozhukov et al. (2014b), one can show that the validity of bootstrap inference holds uniformly over (t, τ). One of the key ingredients to verify Chernozhukov et al. (2014a, Condition H1) is the linear expansions of the estimators with a uniform control of the reminder terms, which has already been established in Theorems 3.4 and 3.5. Last, if the extra moment condition for Y holds, then the same results hold for $\hat{\mu}(t)$ and $\breve{\beta}^1(t)$, i.e.,

$$\mathbb{P}(-\widehat{Q}^{*\mu}(\alpha/2) + \hat{\mu}(t) \leq \mu(t) \leq \widehat{Q}^{*\mu}(\alpha/2) + \hat{\mu}(t)) \to 1 - \alpha,$$

and

$$\mathbb{P}(-\widehat{Q}^{*\mu 1}(\alpha/2) + \breve{\beta}^1(t) \leq \partial_t \mu(t) \leq \widehat{Q}^{*\mu 1}(\alpha/2) + \breve{\beta}^1(t)) \to 1 - \alpha.$$

5. Monte Carlo simulations

This section presents the results of Monte Carlo simulations, which demonstrate the finite sample performance of the estimation and inference procedure. Let Y be generated as

$$Y = \Lambda\left(\left(U + b(X)'\beta - \Phi^{-1}(0.5T + 0.25)\right) \exp\left((T - 0.5)^2\right)\right) \tag{5.1}$$

while T be generated as

$$T = \Lambda(V - b(X)'\beta), \tag{5.2}$$

where U and V are two standard logistic random variables such that $U \perp V$ and $(U, V) \perp X$, $\Lambda(\cdot)$ and $\Phi(\cdot)$ are the logistic and normal CDFs, respectively, $p = 100$, X is a p-dimensional random variables whose distribution is the Gaussian copula with covariance parameter $[0.5^{|j-k|}]_{jk}$, and $b(X)$ is a vector of basis functions constructed from X. Note that T ranges from 0 to 1. The parameters of interest are $q_\tau(t)$ and $\partial_t q_\tau(t)$, where $t = 0.25, 0.5, 0.75$ and $\tau \in (0.2, 0.8)$. We consider the following three designs:

1. (Exact sparse) $\beta_j = \frac{\pi^2}{24}$ for $j = 1, \ldots, 4$ $\beta_j = 0, j \geq 5$, and $b(X_j) = X_j, j = 1, \ldots, 100$;
2. (Approximate sparsity) $\beta_j = \frac{1}{j^2}$ for $j = 1, \ldots, 100$ and $b(X_j) = X_j, j = 1, \ldots, 100$;
3. (Sieve basis) $\beta_1 = \beta_2 = \frac{\pi^2}{12}$ and $\beta_j = 0, j \geq 3$. We construct $b(X)$ as the cubic spline basis functions of (X_1, X_2):

$$b(X) = \left[1, X_1, X_1^2, X_1^3, \max(X_1 - q^{(1)}(0.1), 0)^3, \ldots, \max(X_1 - q^{(1)}(0.9), 0)^3\right]$$
$$\times \left[1, X_2, X_2^2, X_2^3, \max(X_2 - q^{(2)}(0.1), 0)^3, \ldots, \max(X_2 - q^{(2)}(0.9), 0)^3\right],$$

where $q^{(j)}(\tau)$ denotes the τ-th empirical quantile of $X_j, j = 1, 2$. This results in 169 basis functions. We further remove the basis functions with variance less than 10^{-4}. We end up with about 128 basis functions on average.[9]

Note that the sum of the coefficients are (approximately) $\pi^2/6$ for all three designs. We normalize the basis functions $b(X)$ by their sample means and standard errors.

We use Gaussian kernel function in all three stages. We have four tuning parameters: λ, $\tilde{\lambda}$, h_1, and h_2. As we discussed in Section 3.1, we use

$$\lambda = \ell_n(\log(p \vee nh_1)nh_1)^{1/2} \quad \text{and} \quad \tilde{\lambda} = 1.1\Phi^{-1}(1 - \gamma/\{p \vee nh_1\})n^{1/2},$$

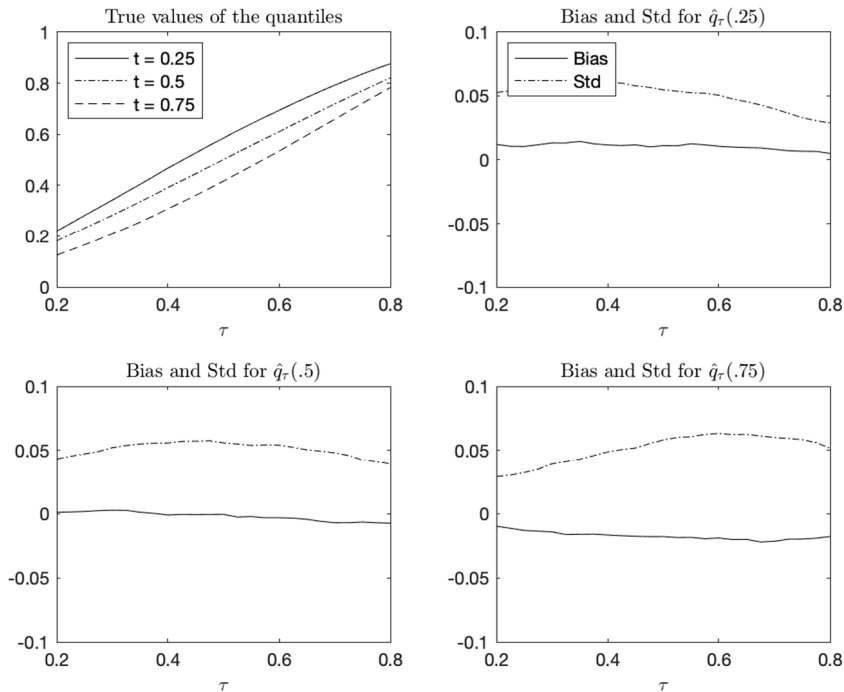[9] The number of basis functions slightly varies across simulations.

Fig. 1. DGP1, finite sample performance of $\hat{q}_\tau(t)$.

where $\ell_n = \sqrt{\log(\log(nh_1))}$ and $\gamma = 1/\log(n)$. We use the rule-of-thumb bandwidth for h_1, i.e., $h_1 = h^* = 1.06 \times sd(T) \times n^{-1/5}$. Last, we build h_2 based on the rule-of-thumb bandwidth for the local quantile regression suggested by Yu and Jones (1998). In particular, Yu and Jones (1998) propose the bandwidths $h_{RoT}(\tau) = C(\tau) \times h_{mean}$, where $C(\tau)$ is a constant dependent only on τ, and $C(0.5) = 1.095$ and $C(0.25) = C(0.75) = 1.13$ and h_{mean} is the bandwidth for the kernel estimation of $\mathbb{E}(Y|T)$.[10] We use the leave-one-out cross-validation to search for the optimal bandwidth of h_{mean} over a grid in $(0.8h^*, 1.2h^*)$. The resulting bandwidth is denoted as h^*_{mean}. In order to achieve under-smoothing, we define $h_2 = n^{-1/10} \times C(\tau) \times h^*_{mean}$, where our choice of the factor $n^{-1/10}$ follows Cai and Xiao (2012, p.418).

The upper-left subplots in Figs. 1, 4, 7 and 2, 5, 8 report the true functions of $q_\tau(t)$ and $\partial_t q_\tau(t)$ for $t = 0.25, 0.5, 0.75$, $\tau \in (0.2, 0.8)$ and DGP 1, 2, and 3, respectively. Both $q_\tau(t)$ and $\partial_t q_\tau(t)$ are heterogeneous across τ and t, which imposes difficulties for estimation and inference. The rest of the subplots in the above Figures show the estimation biases and standard errors. We observe that all the biases of our estimators are of smaller order of magnitude than the standard error (std) and the root mean squared error (rMSE), which indicates the doubly robust moments effectively remove the selection bias induced by the Lasso method. The estimators of the quantile functions are very accurate. The estimators of the quantile partial derivatives are less so because they have slower convergence rates. Figs. 3, 6 and 9 show that the 90% point-wise modified percentile bootstrap confidence intervals have reasonable performance for both the quantile functions and their derivatives, across all τ and t values considered, with slight over-coverage for the quantile derivative functions. The results of variable selections depend on the values of t and (t, u) for conditional density estimation and penalized local MLE, respectively, which are tedious to report. Thus, they are omitted for brevity. Overall, 2 to 4 covariates are selected.

In Section S.3 of the online supplementary material, we report the performance of oracle estimators for the three designs, in which oracle estimators are computed using the true conditional CDF and density functions. We also report the finite-sample performance of our mean potential outcome (i.e., $\mathbb{E}(Y(t))$) estimators, which is similar to that of the quantile effect estimates reported here. Last, we consider an extra design in which the approximate sparsity condition may be violated and show that our method breaks down. We use this design to illustrate the limitation of our method.

---

[10] We refer interested readers to Yu and Jones (1998, Table 1) for more details on $C(\tau)$. In our simulation studies, as $C(\tau)$ is nearly constant over $\tau \in [0.25, 0.75]$, we just choose $C(0.5) = 1.095$ for all the quantile index τ.
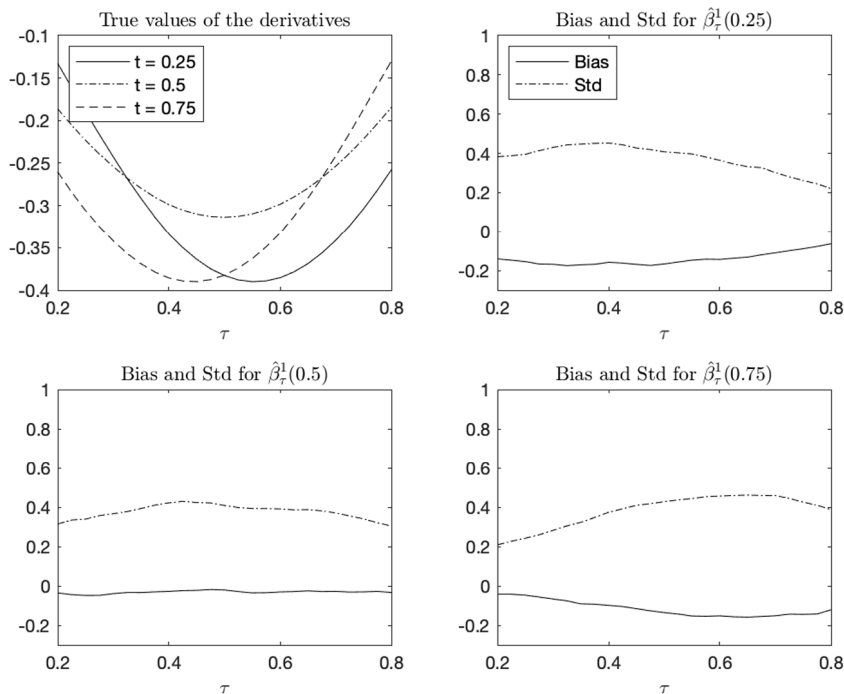
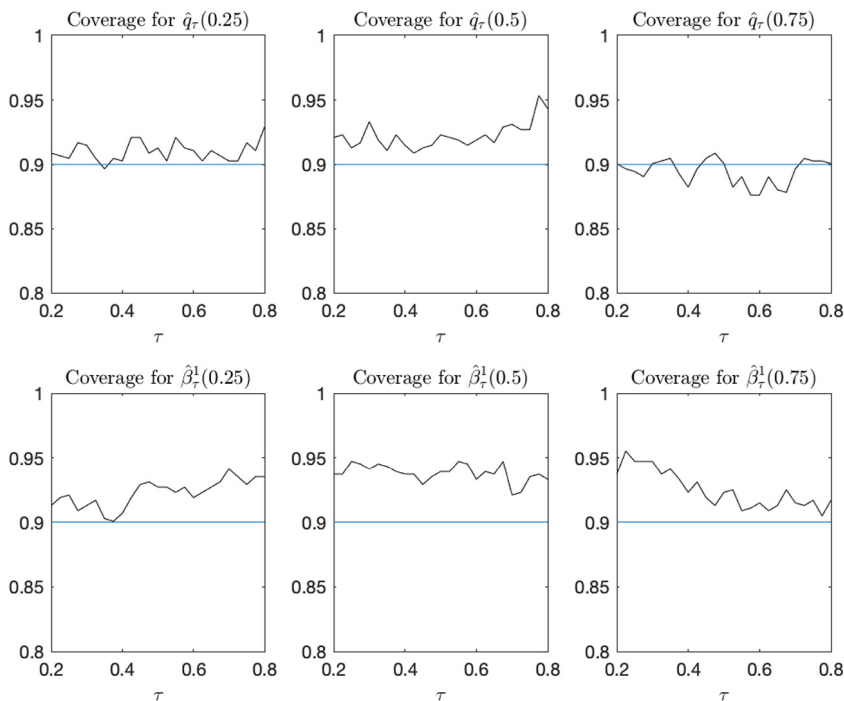**Fig. 2.** DGP1, finite sample performance of $\hat{\beta}_\tau^1(t)$.

Fig. 3. DGP1, coverage probability.

6. Empirical illustration

To investigate our proposed estimation and inference procedures, we use the 1979 National Longitudinal Survey of Youth (NLSY79) and consider the effect of father's income on son's income in the presence of many control variables.
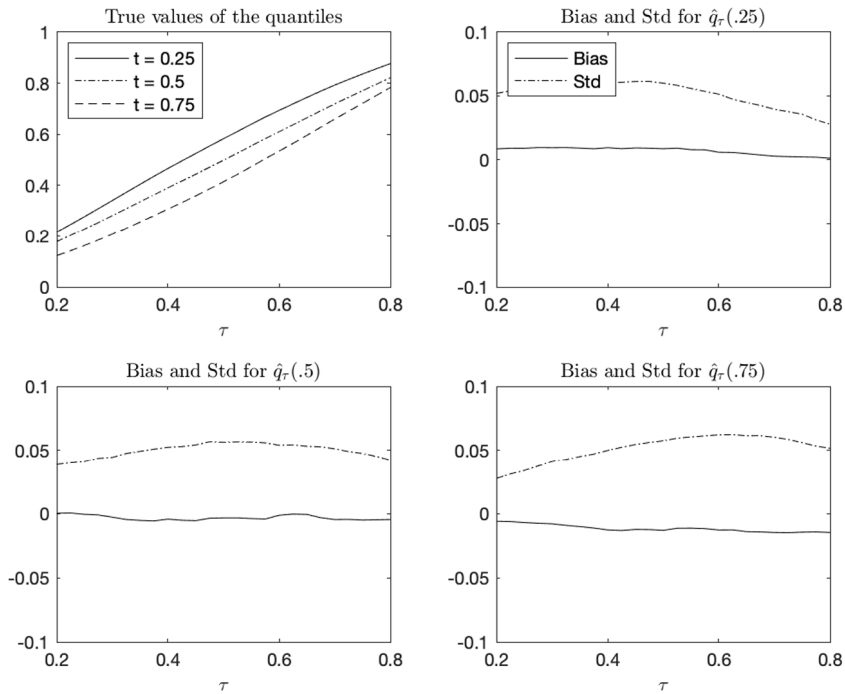
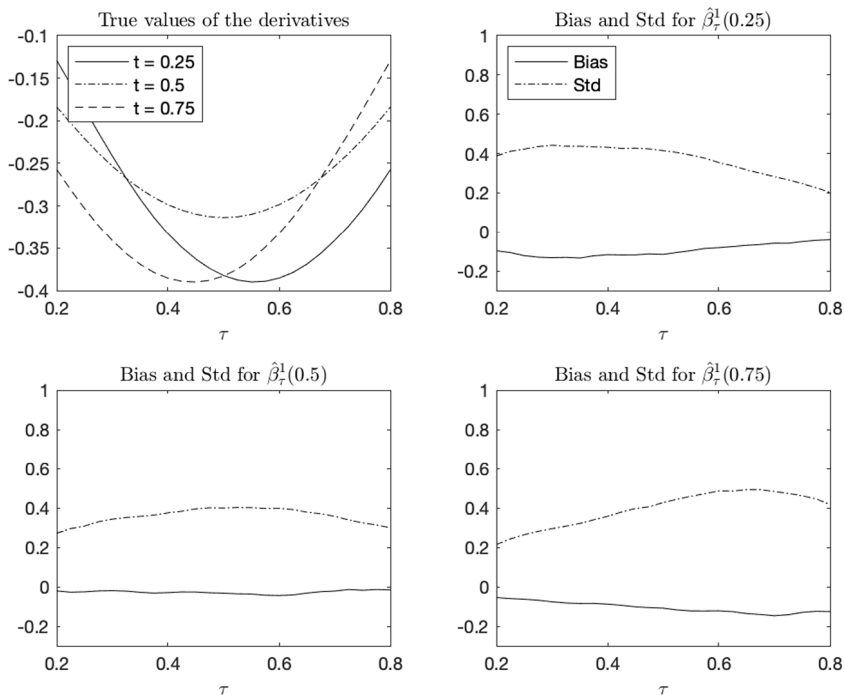**Fig. 4.** DGP2, finite sample performance of $\hat{q}_\tau(t)$.

Fig. 5. DGP2, finite sample performance of $\hat{\beta}_\tau^1(t)$.

Our analysis is based on Bhattacharya and Mazumder (2011). The data consist of a nationally representative sample of individuals with age 14–22 years old as of 1979. We use only white and black males and discard the individuals with missing values in the covariates we use. The resulting sample size is 1,795, out of which 1,302 individuals are white and 493 individuals are black.
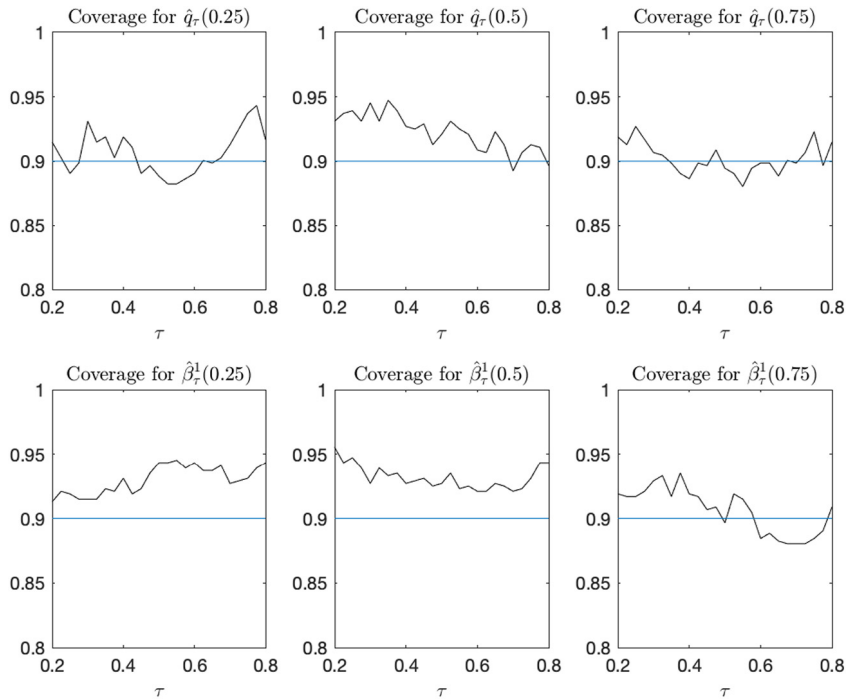
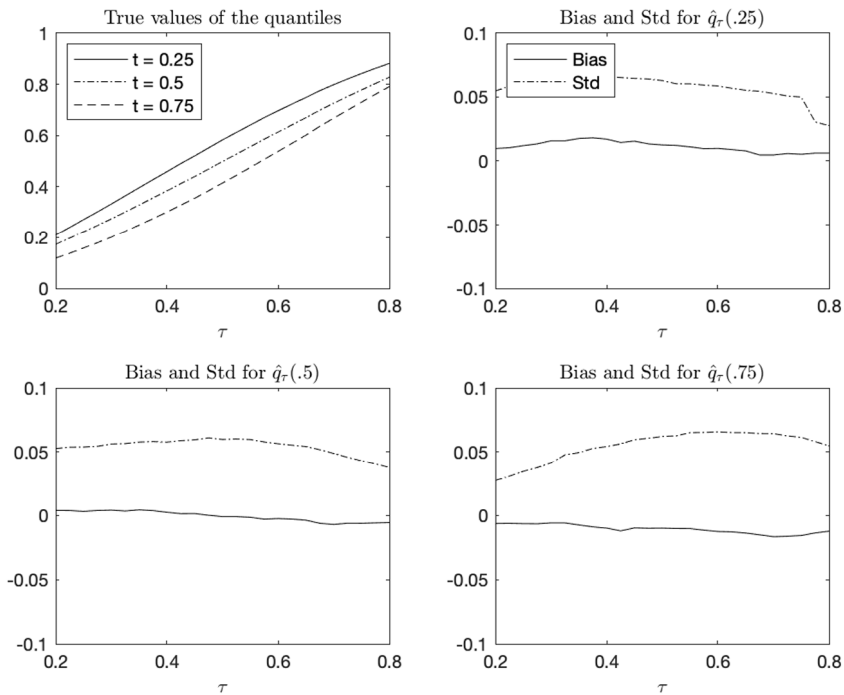**Fig. 6.** DGP2, coverage probability.

Fig. 7. DGP3, finite sample performance of $\hat{q}_\tau(t)$.

The treatment variable of interest is the logarithm of father's income, in which father's income is computed as the average family income for 1978, 1979, and 1980. The outcome variable is the logarithm of son income, in which son income is computed as the average family income for 1997, 1999, 2001 and 2003. We create control variables by
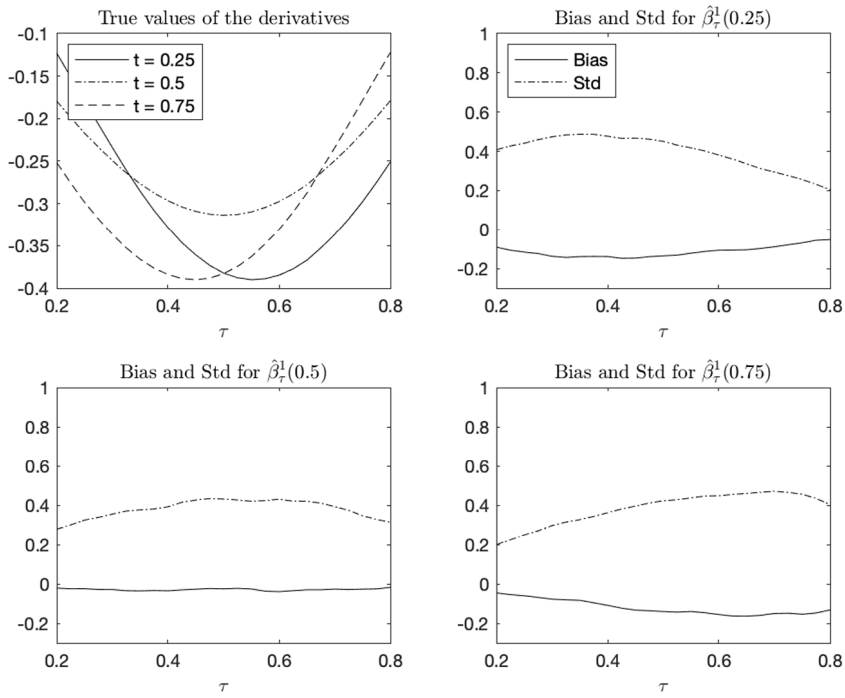
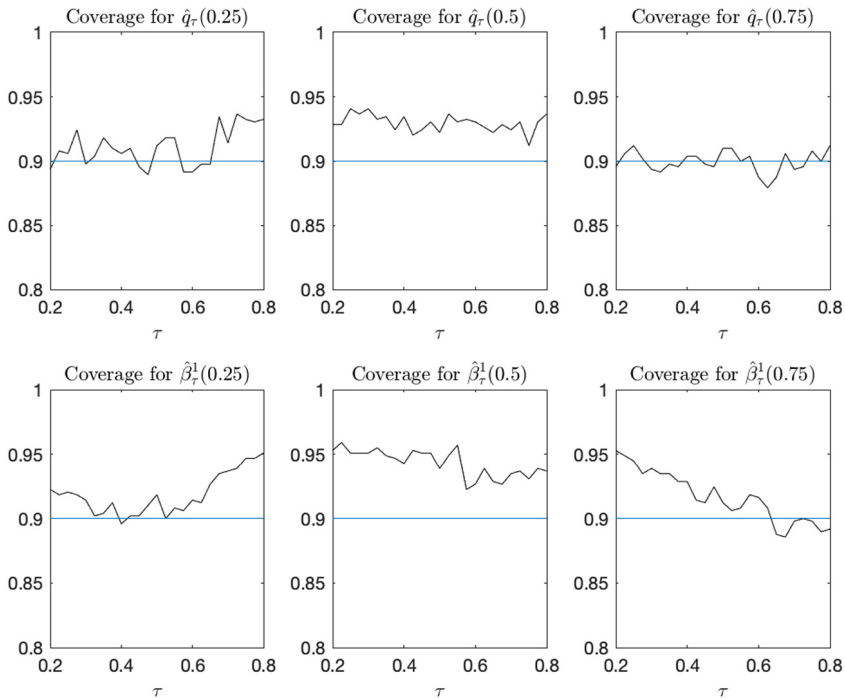**Fig. 8.** DGP3, finite sample performance of $\hat{\beta}_\tau^1(t)$.
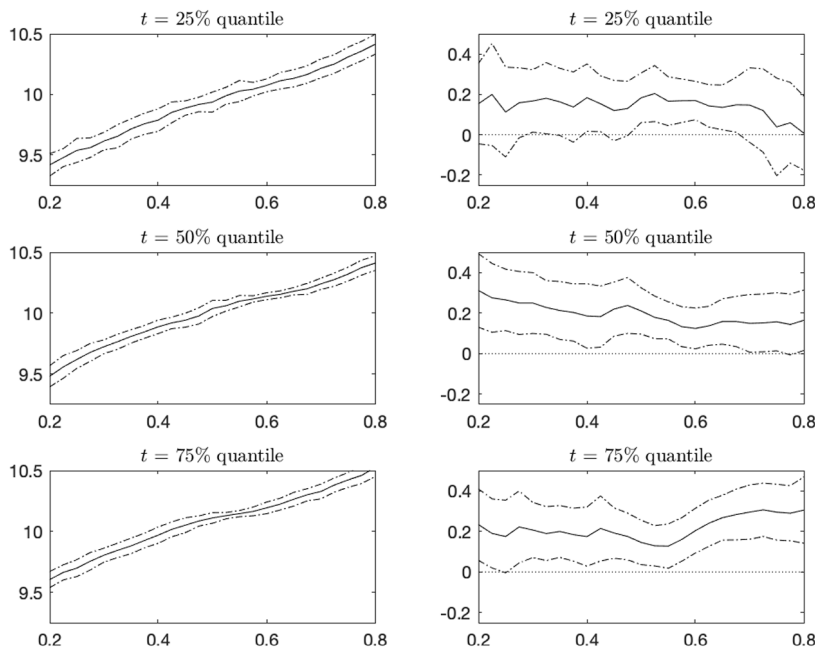

Fig. 9. DGP3, coverage probability.

Fig. 10. Whites. First column: the quantile index τ (X-axis), the son's log income (Y-axis), the estimated unconditional quantile function at τ (solid line), and its (point-wise) 90% confidence bands (dot–dash line). Second column: the quantile index τ (X-axis), the intergenerational elasticity (Y-axis), the estimated derivative of the unconditional quantile function at τ (solid line), and its (point-wise) 90% confidence bands (dot–dash line).

interacting a list of demographic variables with the cubic splines of the AFQT score and the years of education.[11] The list includes the age, the mother's education level, the father's education level, the indicators of (i) living in urban areas at age 14, (ii) living in the south, (iii) speaking a foreign language at childhood, and (iv) being born outside the U.S. We drop the variables whose variance is less than 10^{-4}. The resulting numbers of control variables are 120 for whites and 145 for blacks.

We apply the proposed estimation and inference procedures for black and white individuals separately. We use the same tuning parameter choices as in the previous section.[12] As a result, our effective sample sizes are of orders of magnitude $nh_1 \approx 462$ and 175 for whites and blacks, respectively. Figs. 10 and 11 show the estimated unconditional quantile functions and the estimated derivative, as well as the point-wise 90% confidence bands for $\tau \in [0.2, 0.8]$ and t taking values at the 25%, 50%, and 75% quantiles of the empirical distribution of T_i. Under the context of intergenerational income mobility, the unconditional quantile and its derivative represent the quantile of son's potential log income indexed by father's log income and the intergenerational elasticity, respectively. The unconditional quantile functions have a slight upward trend and the estimated derivative is positive in most parts of father's log income. The confidence bands for the unconditional quantile functions are quite narrow for both black and white individuals. For white individuals with the values of father's log income at the 50% or 75% quantile, we can reject the (locally) zero intergenerational elasticity for most of the values of $\tau \in [0.2, 0.8]$. For the other cases, we cannot reject the (locally) zero intergenerational elasticity for almost all τ's. This is considered as the cost of our fully nonparametric specification.

It is worthwhile to mention the variable selection in this application. The years of education, the AFQT score, the age, the father's education level, and the mother's education level are the leading control variables selected.[13]

7. Conclusion

This paper studies non-separable models with a continuous treatment and high-dimensional control variables. It extends the existing results on the causal inference in non-separable models to the case with both continuous treatment

---

[11] The cubic splines for the AFQT score are constructed based on the normalized value by scaling the raw AFQT score into [0,1], where the knots are taken at the quantiles of the normalized AFQT score at 10%, 20%, ..., 90%. The cubic splines for the years of education are constructed in the same way. In this exercise, we do not interact the cubic splines for the AFQT score and the years of education.

[12] In Section S.4 of the online supplementary material, we investigate the sensitivity of our estimation method with respect to the tuning parameters.

[13] More precisely, for whites, *dad_educ * afqt* and *mom_educ* are the two most selected control variables for the density estimations. *age * educ* and *age * afqt* are the two most selected control variables for the penalized local MLE. For blacks, *mom_educ* and *dad_educ * educ* are the two most selected control variables for the density estimations. *educ* and *age * afqt* are the two most selected control variables for the penalized local MLE.
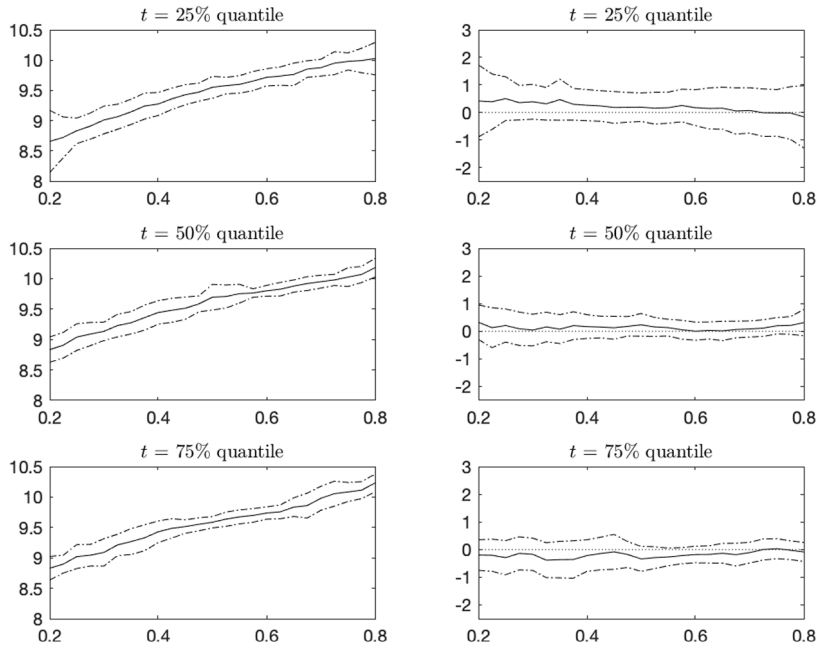
Fig. 11. Blacks. First column: the quantile index τ (X-axis), the son's log income (Y-axis), the estimated unconditional quantile function at τ (solid line), and its (point-wise) 90% confidence bands (dot–dash line). Second column: the quantile index τ (X-axis), the intergenerational elasticity (Y-axis), the estimated derivative of the unconditional quantile function at τ (solid line), and its (point-wise) 90% confidence bands (dot–dash line).

and high-dimensional covariates. It develops a method based on localized L_1-penalization to select covariates at each value of the continuous treatment. It then proposes a multi-stage estimation and inference procedure for average, quantile, and marginal treatment effects. The simulation and empirical exercises support the theoretical findings in finite samples.

Appendix A. Proof of the main results in the paper

Before proving the theorem, we first introduce some additional notation and Assumption 6, which is a restatement of Sasaki (2015, Assumptions 1 and 2) in our framework. Denote by $\dim_X$ (resp. $\dim_A$) the dimensionality of X (resp. A). We define $\partial V(y, t) = \{(x, a) : \Gamma(t, x, a) = y\}$ and $\partial V(y, t)$ can be parametrized as a mapping from a $(\dim_X + \dim_A - 1)$-dimensional rectangle, denoted by Σ, to $\partial V(y, t)$. $H^{\dim_X + \dim_A - 1}$ is the $(\dim_X + \dim_A - 1)$-dimensional Hausdorff measure restricted from $\mathbb{R}^{\dim_X + \dim_A}$ to $(\partial V(y, t), \mathcal{B}(y, t))$, where $\mathcal{B}(y, t)$ is the set of the interactions between $\partial V(y, t)$ and a Borel set in $\mathbb{R}^{\dim_X + \dim_A}$. $\partial v(y, \cdot; u)/\partial y$ (resp. $\partial v(\cdot, t; u)/\partial t$) is the velocity of $\partial V(y, t)$ at u with respect to y (resp. t).

Assumption 6.

1. Γ is continuously differentiable.
2. $\|\nabla_{(x,a)} \Gamma(t, \cdot, \cdot)\| \neq 0$ on $\partial V(y, t)$.
3. The conditional distribution of (X, A) given T is absolutely continuous with respect to the Lebesgue measure, and $f_{(X,A)|T}$ is a continuously differentiable function of $\mathcal{T}$ to $L^1(\mathbb{R}^{\dim_X + \dim_A})$.
4. $\int_{\partial V(y,t)} f_{(X,A)|T}(x, a \mid t) dH^{\dim_X + \dim_A - 1}(x, a) > 0$.
5. $t \mapsto \partial V(y, t)$ is a continuously differentiable function of $\Sigma \times \mathcal{T}$ to $\mathbb{R}^{\dim_X + \dim_A}$ for every y and $y \mapsto \partial V(y, t)$ is a continuously differentiable function of $\Sigma \times \mathcal{Y}$ to $\mathbb{R}^{\dim_X + \dim_A}$ for every t.
6. The mapping $\partial v(y, \cdot; \cdot)/\partial t$ is a continuously differentiable function of $\mathcal{T}$ to $\mathbb{R}^{\dim_X + \dim_A}$ and $\partial v(\cdot, t; \cdot)/\partial y$ is a continuously differentiable function of $\mathcal{Y}$ to $\mathbb{R}^{\dim_X + \dim_A}$.
7. There is $p, q \geq 1$ with $\frac{1}{p} + \frac{1}{q} = 1$ such that the mapping $(x, a) \mapsto \|\nabla_{(x,a)} \Gamma(t, x, a)\|^{-1}$ is bounded in $L^p(\partial V(y, t), H^{\dim_X + \dim_A - 1})$ and that the mapping $(x, a) \mapsto f_{(X,A)}(x, a)$ is bounded in $L^q(\partial V(y, t), H^{\dim_X + \dim_A - 1})$.

Assumption 6 is a combination of Assumptions 1 and 2 in Sasaki (2015). We refer the readers to the paper for detailed explanation.

Proof of Theorem 2.1. For the marginal distribution of $Y(t)$, we note that, by Assumption 1, $\mathbb{P}(Y(t) \leq u) = \mathbb{E}[\mathbb{E}(1\{Y(t) \leq u\}|X)] = \mathbb{E}[\mathbb{E}(1\{Y(t) \leq u\}|X, T = t)] = \mathbb{E}[\mathbb{E}(1\{Y \leq u\}|X, T = t)]$. The first result follows as $\mathbb{E}(1\{Y \leq u\}|X, T = t)$ is identified.

For the second result, consider a random variable T^* which has the same marginal distribution as T and is independent of (X, A). Define

$$Y^* = \Gamma(T^*, X, A). \tag{A.1}$$

Note that the (i) (X, A) and T^* are independent, and (ii) the τ-th quantile of Y^* given $T^* = t$ is $q_\tau(t)$ for all t, because $\mathbb{P}(Y^* \leq q_\tau(t) \mid T^* = t) = \mathbb{P}(\Gamma(t, X, A) \leq q_\tau(t)) = \tau$. Assumption 6 implies Assumptions 1 and 2 in Sasaki (2015) for (Y^*, T^*, U^*) with $U^* = (X, A)$, and then his Theorem 1 implies that the derivative of the τ-th quantile of Y^* given $T^* = t$ is equal to $\mathbb{E}_{\mu_{\tau,t}}[\partial_t \Gamma(t, X, A)]$. Therefore, $\partial_t q_\tau(t) = \mathbb{E}_{\mu_{\tau,t}}[\partial_t \Gamma(t, X, A)]$. Note that Theorem 1 in Sasaki (2015) does not apply directly to (Y, T, U^*), because our assumptions do not imply that T and U^* are independent. ∎

Lemma 3.1 is the local version of the compatibility condition, which is one of the key building blocks for Lemma A.1. Then, Lemma A.1 is used to prove Theorem 3.1.

Proof of Lemma 3.1. By Assumption 4, we can work on the set

$$\left\{ \{X_i\}_{i=1}^n : \sup_{|\delta|_0 \leq s\ell_n} \frac{\|b(X)'\delta\|_{\mathbb{P}_n,2}}{\|\delta\|_2} \leq \kappa'' < \infty \right\}.$$

We use the same partition as in Bickel et al. (2009). Let $\mathcal{S}_0 = \mathcal{S}_{t,u}$ and $m \geq s$ be an integer which will be specified later. Partition $\mathcal{S}_{t,u}^c$, the complement of $\mathcal{S}_{t,u}$, as $\sum_{l=1}^L \mathcal{S}_l$ such that $|\mathcal{S}_l| = m$ for $1 \leq l < L$, $|\mathcal{S}_L| \leq m$, where $\mathcal{S}_l$, for $l < L$, contains the indexes corresponding to m largest coordinates (in absolute value) of δ outside $\cup_{j=0}^{l-1} \mathcal{S}_j$, and $\mathcal{S}_L$ collects the remaining indexes. Further denote $\delta_j = \delta_{\mathcal{S}_j}$ and $\delta_{01} = \delta_{\mathcal{S}_0 \cup \mathcal{S}_1}$. Then

$$\|b(X)'\delta K(\frac{T-t}{h_1})^{1/2}\|_{\mathbb{P}_n,2} \geq \|b(X)'\delta_{01}K(\frac{T-t}{h_1})^{1/2}\|_{\mathbb{P}_n,2} - \sum_{l=2}^L \|b(X)'\delta_l K(\frac{T-t}{h_1})^{1/2}\|_{\mathbb{P}_n,2}. \tag{A.2}$$

For the first term on the right hand side (r.h.s.) of (A.2), we have

$$\|b(X)'\delta_{01}K(\frac{T-t}{h_1})^{1/2}\|_{\mathbb{P}_n,2}^2$$

$$\geq \|b(X)'\delta_{01}K(\frac{T-t}{h_1})^{1/2}\|_{P,2}^2 - |(\mathbb{P}_n - \mathbb{P})(b(X)'\delta_{01})^2 K(\frac{T-t}{h_1})|$$

$$\geq \underline{C}h_1\|b(X)'\delta_{01}\|_{P,2}^2 - |(\mathbb{P}_n - \mathbb{P})(b(X)'\delta_{01})^2 K(\frac{T-t}{h_1})| \tag{A.3}$$

$$\geq \underline{C}h_1\|b(X)'\delta_{01}\|_{\mathbb{P}_n,2}^2 - \underline{C}h|(\mathbb{P}_n - \mathbb{P})(b(X)'\delta_{01})^2| - |(\mathbb{P}_n - \mathbb{P})(b(X)'\delta_{01})^2 K(\frac{T-t}{h_1})|$$

$$\geq \underline{C}h_1\|\delta_{01}\|_2^2(\kappa')^2 - \underline{C}h_1|(\mathbb{P}_n - \mathbb{P})(b(X)'\delta_{01})^2| - |(\mathbb{P}_n - \mathbb{P})(b(X)'\delta_{01})^2 K(\frac{T-t}{h_1})|$$

where the second inequality holds because

$$\mathbb{E}(b(X)'\delta_{01})^2 K(\frac{T-t}{h_1}) = h_1 \mathbb{E}(b(X)'\delta_{01})^2 \int f_{t+h_1 v}(X) K(v) dv \geq \underline{C}h_1 \mathbb{E}(b(X)'\delta_{01})^2.$$

We next bound the last term on the r.h.s. of (A.2). The second term can be bounded in the same manner. Let $\tilde{\delta}_{01} = \delta_{01}/\|\delta_{01}\|_2$. Then we have

$$|(\mathbb{P}_n - \mathbb{P})(b(X)'\delta_{01})^2 K(\frac{T-t}{h_1})| = \|\delta_{01}\|_2^2 |(\mathbb{P}_n - \mathbb{P})(b(X)'\tilde{\delta}_{01})^2 K(\frac{T-t}{h_1})|.$$

Let $\{\eta_i\}_{i=1}^n$ be a sequence of Rademacher random variables which is independent of the data and $\mathcal{F} = \{b(X)'\delta K(\frac{T-t}{h_1})^{1/2} : \|\delta\|_0 = m+s, \|\delta\|_2 = 1, t \in \mathcal{T}\}$ with envelope $F = \overline{C}_K \zeta_n (m+s)^{1/2}$. Denote π_{1n} as $(\frac{\log(p \vee n)(s+m)^2 \zeta_n^2}{nh_1})^{1/2}$ with $m = s\ell_n^{1/2}$. Then,

$$\mathbb{E} \sup_{\|\tilde{\delta}_{01}\|_0 \leq m+s, \|\tilde{\delta}_{01}\|_2=1, t \in \mathcal{T}} |(\mathbb{P}_n - \mathbb{P})(b(X)'\tilde{\delta}_{01})^2 K(\frac{T-t}{h_1})|$$

$$\leq 2\mathbb{E} \sup_{\|\tilde{\delta}_{01}\|_0 \leq m+s, \|\tilde{\delta}_{01}\|_2=1, t \in \mathcal{T}} |\mathbb{P}_n \eta (b(X)'\tilde{\delta}_{01})^2 K(\frac{T-t}{h_1})|$$

$$\leq 8\zeta_n \left(\sup_{\|\tilde{\delta}_{01}\|_0 \leq m+s, \|\tilde{\delta}_{01}\|_2=1} \|\tilde{\delta}_{01}\|_1 \right) \left(\mathbb{E} \sup_{f \in \mathcal{F}} |\mathbb{P}_n \eta f| \right)$$

$$\lesssim 8\zeta_n(m+s)^{1/2}\left[\left(\frac{\log(p\vee n)(s+m)h_1}{n}\right)^{1/2}+\frac{\overline{C}_K\zeta_n(m+s)^{1/2}\log(p\vee n)(s+m)}{n}\right]$$

$$\lesssim\left(\frac{\log(p\vee n)(s+m)^2h_1\zeta_n^2}{n}\right)^{1/2}=h_1\pi_{1n},$$

where the first inequality is by van der Vaart and Wellner (1996, Lemma 2.3.1), the second inequality is by Ledoux and Talagrand (2013, Theorem 4.12) and the remark thereafter, and the third one is by applying Corollary 5.1 of Chernozhukov et al. (2014b) with $\sigma^2=\sup_{f\in\mathcal{F}}\mathbb{E}f^2\lesssim h_1$ and, for some $A\geq e$,

$$\sup_Q N(\mathcal{F},e_Q,\varepsilon\|F\|_{Q,2})\leq\binom{p}{s+m}\left(\frac{A}{\varepsilon}\right)^{s+m}\lesssim\left(\frac{Ap}{\varepsilon}\right)^{s+m}.$$

By Assumption 2, $\pi_{1n}\to 0$. Then we have, w.p.a.1.,

$$|(\mathbb{P}_n-\mathbb{P})(b(X)'\delta_{01})^2K(\frac{T-t}{h_1})|\leq 3h_1\underline{C}(\kappa')^2\|\delta_{01}\|_2^2/8. \tag{A.4}$$

By the same token we can show that

$$\mathbb{E}\sup_{\|\tilde\delta_{01}\|_0\leq m+s,\|\tilde\delta_{01}\|_2=1,t\in\mathcal{T}}|(\mathbb{P}_n-\mathbb{P})(b(X)'\tilde\delta_{01})^2|\lesssim\sqrt{h_1}\pi_{1n}\to 0.$$

Therefore, we have, w.p.a.1.,

$$|(\mathbb{P}_n-\mathbb{P})(b(X)'\delta_{01})^2|\leq 3(\kappa')^2\|\delta_{01}\|_2^2/8. \tag{A.5}$$

Combining (A.3), (A.4), and (A.5) yields that w.p.a.1.,

$$\|b(X)'\delta_{01}K(\frac{T-t}{h_1})^{1/2}\|_{\mathbb{P}_n,2}^2\geq\|\delta_{01}\|_2^2h_1(\kappa')^2\underline{C}/4.$$

Analogously, we can show that, w.p.a.1,

$$\|b(X)'\delta_l K(\frac{T-t}{h_1})^{1/2}\|_{\mathbb{P}_n,2}^2\leq 4\|\delta_l\|_2^2\underline{C}^{-1}h_1(\kappa'')^2.$$

Following (A.2), we have, w.p.a.1,

$$\begin{aligned}\|b(X)'\delta K(\frac{T-t}{h_1})^{1/2}\|_{\mathbb{P}_n,2}&\geq h_1^{1/2}\|\delta_{01}\|_2\kappa'\underline{C}^{1/2}/2-h_1^{1/2}\sum_{l=2}^L 2\|\delta_l\|_2\kappa''\underline{C}^{-1/2}\\&\geq h_1^{1/2}\|\delta_{01}\|_2\kappa'\underline{C}^{1/2}/2-h_1^{1/2}\sum_{l=2}^L 2\kappa''\underline{C}^{-1/2}(\|\delta_{l-1}\|_1\|\delta_l\|_1)^{1/2}/\sqrt{m}\\&\geq h_1^{1/2}\|\delta_{01}\|_2\kappa'\underline{C}^{1/2}/2-2h_1^{1/2}\kappa''\underline{C}^{-1/2}\|\delta_{T^c}\|_1/\sqrt{m}\\&\geq h_1^{1/2}\|\delta_{01}\|_2\kappa'\underline{C}^{1/2}/2-2h_1^{1/2}\kappa''\underline{C}^{-1/2}c^{1/2}\|\delta_0\|_1/\sqrt{m}\\&\geq h_1^{1/2}\|\delta_{01}\|_2\kappa'\underline{C}^{1/2}/2-2h_1^{1/2}\kappa''\underline{C}^{-1/2}c^{1/2}\|\delta_0\|_2\sqrt{s}/\sqrt{m}\\&\geq h_1^{1/2}\|\delta_0\|_2\left[\kappa'\underline{C}^{1/2}/2-2\kappa''\underline{C}^{-1/2}c^{1/2}\sqrt{s}/\sqrt{m}\right],\end{aligned}$$

where the second inequality holds because, by construction, $\|\delta_l\|_2^2\leq\|\delta_{l-1}\|_1\|\delta_l\|_1/\sqrt{m}$. Since $m=s\ell_n^{1/2}$, $s/m=\ell_n^{-1/2}\to 0$, and thus, for n large enough, the constant inside the brackets is greater than $\kappa'\underline{C}^{1/2}/4$ which is independent of (t,u,n). Therefore, we can conclude that, for n large enough,

$$\inf_{(t,u)\in\mathcal{T}\mathcal{U}}\inf_{\delta\in\Delta_{2\tilde c,t,u}}\frac{\|b(X)'\delta K(\frac{T-t}{h_1})^{1/2}\|_{\mathbb{P}_n,2}}{\|\delta_{\mathcal{S}_{t,u}}\|_2\sqrt{h_1}}\geq\kappa'\underline{C}^{1/2}/4:=\underline{\kappa}.$$

This completes the proof of the lemma. ∎

We aim to prove the results with regard to $\widehat{\phi}_{t,u}(X)$ and $\hat\theta_{t,u}$ in Theorem 3.1. The derivations for the results regarding $\widetilde{\phi}_{t,u}(X)$ and $\widetilde{\theta}_{t,u}$ are exactly the same. We do not need to deal with the nonlinear logistic link function when deriving the results regarding $\widehat{v}_t(X)$, $\widetilde{v}_t(X)$, $\hat\gamma_t$, and $\tilde\gamma_t$. Therefore, the corresponding results can be shown by following the same proving strategy as below and treating $\omega_{t,u}$ defined below as 1. The proofs for results regarding $\widehat{v}_t(X)$, $\widetilde{v}_t(X)$, $\hat\gamma_t$, and $\tilde\gamma_t$ are omitted for brevity.

Let $\tilde r_{t,u}^\phi=\Lambda^{-1}(\mathbb{E}(Y_u|X,T=t))-b(X)'\theta_{t,u}$, $\delta_{t,u}=\hat\theta_{t,u}-\theta_{t,u}$, $\hat s_{t,u}=\|\hat\theta_{t,u}\|_0$, $\omega_{t,u}=\mathbb{E}(Y_u(t)|X)(1-\mathbb{E}(Y_u(t)|X))$, and $\widehat{\mathcal{S}}_{t,u}$ be the support of $\widehat\theta_{t,u}$. We need the following four lemmas, whose proofs are relegated to the online supplement.

Lemma A.1. *If Assumptions 1–4 hold, then*

$$\sup_{(t,u)\in\mathcal{TU}} \|\omega_{t,u}^{1/2} b(X)'\delta_{t,u} K(\frac{T-t}{h_1})^{1/2}\|_{\mathbb{P}_n,2} = O_p(\ell_n(\log(p\vee n)s)^{1/2}n^{-1/2})$$

and

$$\sup_{(t,u)\in\mathcal{TU}} \|\delta_{t,u}\|_1 = O_p(\ell_n(\log(p\vee n)s^2)^{1/2}(nh_1)^{-1/2}).$$

Lemma A.2. *Suppose Assumptions 1–4 hold. Let $\xi_{t,u} = Y_u - \phi_{t,u}(X)$. Then*

$$\sup_{(t,u)\in\mathcal{TU}} \left\| \widehat{\Psi}_{t,u}^{-1} \mathbb{P}_n\left[\xi_{t,u} K(\frac{T-t}{h_1})b(X)\right]\right\|_\infty = O_p((\log(p\vee n)h_1/n)^{1/2}).$$

Lemma A.3. *If the assumptions in Theorem 3.1 hold, then there exists a constant $C_\psi \in (0,1)$ such that w.p.a.1,*

$$C_\psi/2 \le \inf_{(t,u)\in\mathcal{TU},j=1,\dots,p} l^0_{t,u,j} \le \sup_{(t,u)\in\mathcal{TU},j=1,\dots,p} l^0_{t,u,j} \le 2/C_\psi. \tag{A.6}$$

For any $k = 0, 1, \dots, K$ and $\widehat{\Psi}_{t,u}^k$ defined in Algorithm 3.1, there exists a constant $C_k \in (0,1)$ such that, w.p.a.1,

$$C_k/2 \le \inf_{(t,u)\in\mathcal{TU},j=1,\dots,p} l^k_{t,u,j} \le \sup_{(t,u)\in\mathcal{TU},j=1,\dots,p} l^k_{t,u,j} \le 2C_k. \tag{A.7}$$

In addition, for any $k = 0, 1, \dots, K$ and $\widehat{\Psi}_{t,u}^k$ defined in Algorithm 3.1, there exist constants $l < 1 < L$ independent of n, (t,u), and k such that, element-wise and w.p.a.1,

$$l\widehat{\Psi}_{t,u,0} \le \widehat{\Psi}_{t,u}^k \le L\widehat{\Psi}_{t,u,0}. \tag{A.8}$$

Lemma A.4. *If the assumptions in Theorem 3.1 hold, then w.p.a.1,*

$$\sup_{t\in\mathcal{T},\|\delta\|_2=1,\|\delta\|_0\le s\ell_n} \|b(X)'\delta K(\frac{T-t}{h_1})^{1/2}\|_{\mathbb{P}_n,2} h_1^{-1/2} \le 2\underline{C}^{-1/2}\kappa''.$$

Proof of Theorem 3.1. By the mean value theorem, there exist $\underline{\theta}_{t,u} \in (\theta_{t,u}, \hat{\theta}_{t,u})$ and $\bar{r}_{t,u}^\phi \in (0, \tilde{r}_{t,u}^\phi)$ such that

$$|\phi_{t,u}(X) - \widehat{\phi}_{t,u}(X)| \le \Lambda(b(X)'\underline{\theta}_{t,u} + \bar{r}_{t,u}^\phi)(1 - \Lambda(b(X)'\underline{\theta}_{t,u} + \bar{r}_{t,u}^\phi))(b(X)'\delta_{t,u} + \tilde{r}_{t,u}^\phi),$$

where $\delta_{t,u} = \hat{\theta}_{t,u} - \theta_{t,u}$. By the proof of Lemma A.1, we have, w.p.a.1,

$$|\tilde{r}_{t,u}^\phi| \le [\underline{C}/2(1-\underline{C}/2)]^{-1}|r_{t,u}^\phi|.$$

Therefore, by Lemma A.1 and Assumptions 4 and 5, we have

$$\sup_{(t,u)\in\mathcal{TU}} |b(X)'\underline{\theta}_{t,u} + \bar{r}_{t,u}^\phi - b(X)'\theta_{t,u} - \tilde{r}_{t,u}^\phi|$$

$$\lesssim \sup_{(t,u)\in\mathcal{TU}} |b(X)'\delta_{t,u}| + \sup_{(t,u)\in\mathcal{TU}} |r_{t,u}^\phi|$$

$$\lesssim \zeta_n \sup_{(t,u)\in\mathcal{TU}} \|\delta_{t,u}\|_1 + O((\log(p\vee n)s^2\zeta_n^2/(nh_1))^{-1/2}) = o_p(1),$$

where the last equality is because $\sup_{(t,u)\in\mathcal{TU}} \|\delta_{t,u}\|_1 = O_p((\log(p\vee n)s^2)^{1/2}(nh_1)^{-1/2})$ by Lemma A.1 and $\log(p\vee n)s^2\zeta_n^2/(nh_1) \to 0$ by Assumption 5. In addition, under Assumption 3.4 we have

$$\Lambda(b(X)'\theta_{t,u} + \tilde{r}_{t,u}^\phi) = \mathbb{E}(Y_u|X, T=t) \in [\underline{C}, 1-\underline{C}].$$

Hence, there exist some positive constants c and c' only depending on $\underline{C}$ such that, w.p.a.1,

$$\Lambda(b(X)'\underline{\theta}_{t,u} + \bar{r}_{t,u}^\phi)(1 - \Lambda(b(X)'\underline{\theta}_{t,u} + \bar{r}_{t,u}^\phi)) \le c$$

and uniformly over $(t,u) \in \mathcal{TU}$,

$$|\phi_{t,u}(X) - \widehat{\phi}_{t,u}(X)| \le c(b(X)'\delta_{t,u} + \tilde{r}_{t,u}^\phi) \le c'(b(X)'\delta_{t,u} + r_{t,u}^\phi). \tag{A.9}$$

By Assumptions 3.3, 3.4, Lemma A.1, and the fact that $\omega_{t,u}$ is bounded and bounded away from zero uniformly over $\mathcal{TU}$, we have, w.p.a.1,

$$
\sup_{(t,u)\in\mathcal{TU}} \|(\phi_{t,u}(X) - \widehat{\phi}_{t,u}(X))K(\frac{T-t}{h_1})^{1/2}\|_{\mathbb{P}_n,2}
$$
$$
\leq \sup_{(t,u)\in\mathcal{TU}} c\left[\|b(X)'\delta_{t,u}K(\frac{T-t}{h_1})^{1/2}\|_{\mathbb{P}_n,2} + \|r_{t,u}^\phi K(\frac{T-t}{h_1})^{1/2}\|_{\mathbb{P}_n,2}\right]
$$
$$
= O_p(\ell_n(\log(p\vee n)s/n)^{1/2}) \tag{A.10}
$$

and

$$
\sup_{(t,u)\in\mathcal{TU}} \|\phi_{t,u}(X) - \widehat{\phi}_{t,u}(X)\|_{\mathbb{P},\infty} \lesssim \zeta_n \sup_{(t,u)\in\mathcal{TU}} \|\delta_{t,u}\|_1 + O((\log(p\vee n)s^2\zeta_n^2/(nh_1))^{1/2})
$$
$$
= O_p(\ell_n(\log(p\vee n)s^2\zeta_n^2/(nh_1))^{1/2}). \tag{A.11}
$$

Next, recall that $\lambda = \ell_n(\log(p\vee n)nh)^{1/2}$. By the first order conditions (FOC), for any $j\in\widehat{S}_{t,u}$, we have

$$
\left|\mathbb{P}_n\left[(Y_u - \Lambda(b(X)'\hat{\theta}_{t,u}))b_j(X)K(\frac{T-t}{h_1})\right]\right| = \widehat{\Psi}_{t,u,jj}\frac{\lambda}{n}.
$$

Denote $\xi_{t,u} = Y_u - \phi_{t,u}(X)$. By Lemmas A.1–A.3, for any $\varepsilon > 0$, with probability greater than $1-\varepsilon$, there exist positive constants C_λ and C, which only depend on ε and are independent of (t,u,n), such that

$$
\frac{\lambda\hat{s}_{t,u}^{1/2}}{n} = \left\|\widehat{\Psi}_{t,u}^{-1}\left\{\mathbb{P}_n\left[(Y_u - \Lambda(b(X)'\hat{\theta}_{t,u}))b(X)K(\frac{T-t}{h_1})\right]\right\}_{\widehat{S}_{t,u}}\right\|_2
$$
$$
\leq \sup_{\|\theta\|_0\leq\hat{s}_{t,u},\|\theta\|_2=1} \|\theta\|_1 \sup_{(t,u)\in\mathcal{TU}} \|\widehat{\Psi}_{t,u}^{-1}(\mathbb{P}_n\xi_{t,u}b(X)K(\frac{T-t}{h_1}))\|_\infty
$$
$$
+ \frac{\|\widehat{\Psi}_{t,u,0}^{-1}\|_\infty}{l} \sup_{\|\theta\|_0\leq\hat{s}_{t,u},\|\theta\|_2=1} \left|\left\{\mathbb{P}_n\left(\Lambda(b(X)'\hat{\theta}_{t,u}) - \Lambda(b(X)'\theta_{t,u}) - r_{t,u}^\phi\right)b(X)'\theta K(\frac{T-t}{h_1})\right\}\right|
$$
$$
\leq \frac{C_\lambda\lambda\hat{s}_{t,u}^{1/2}}{n\ell_n} + \frac{c'\|\widehat{\Psi}_{t,u,0}^{-1}\|_\infty}{l}\|(b(X)'\delta_{t,u} + r_{t,u}^\phi)K(\frac{T-t}{h_1})^{1/2}\|_{\mathbb{P}_n,2}
$$
$$
\times \sup_{\|\theta\|_0\leq\hat{s}_{t,u},\|\theta\|_2=1} \|b(X)'\theta K(\frac{T-t}{h_1})^{1/2}\|_{\mathbb{P}_n,2}
$$
$$
\leq \frac{\lambda\hat{s}_{t,u}^{1/2}}{2n} + C(\log(p\vee n)s/n)^{1/2}\phi_{max}^{1/2}(\hat{s}_{t,u})
$$
$$
\leq \frac{\lambda\hat{s}_{t,u}^{1/2}}{2n} + \frac{C\lambda s^{1/2}}{nh_1^{1/2}}\phi_{max}^{1/2}(\hat{s}_{t,u})
$$

where $\phi_{max}(s) = \sup_{\|\theta\|_0\leq s,\|\theta\|_2=1} \|b(X)'\theta K(\frac{T-t}{h_1})^{1/2}\|_{\mathbb{P}_n,2}^2$ and $r_{t,u}^\phi = r_{t,u}^\phi(X)$. This implies that there exists a constant C only depending on ε, such that, with probability greater than $1-\varepsilon$,

$$
\hat{s}_{t,u} \leq Cs\phi_{max}(\hat{s}_{t,u})/h_1. \tag{A.12}
$$

Let $\mathcal{M} = \{m \in \mathbb{Z} : m > 2Cs\phi_{max}(m)/h_1\}$. We claim that, for any $m \in \mathcal{M}$, $\hat{s}_{t,u} \leq m$. Suppose not and there exists $m_0 \in \mathcal{M}$ such that $m_0 < \hat{s}_{t,u}$. Then,

$$
\hat{s}_{t,u} \leq Cs\phi_{max}(\frac{\hat{s}_{t,u}}{m_0}m_0)/h_1 \leq \lceil\frac{\hat{s}_{t,u}}{m_0}\rceil Cs\phi_{max}(m_0)/h_1 \leq \frac{\hat{s}_{t,u}}{m_0}\left[2Cs\phi_{max}(m_0)/h_1\right] < \hat{s}_{t,u},
$$

where the second inequality holds because of Belloni and Chernozhukov (2011, Lemma 23), the third inequality holds because $\lceil a\rceil \leq 2a$ for any $a > 1$, and the last inequality holds because $m_0 \in \mathcal{M}$. Therefore we reach a contradiction. In addition, by Lemma A.4, we can choose $C_s > 4C\underline{C}^{-1}(\kappa'')^2$, which is independent of (t,u,n), such that

$$
2Cs\phi_{max}(C_s s)/h_1 \leq 4C\underline{C}^{-1}(\kappa'')^2 s < C_s s. \tag{A.13}
$$

This implies $C_s s \in \mathcal{M}$ and thus with probability greater than $1-\varepsilon$, $\hat{s}_{t,u} \leq C_s s$. This result holds uniformly over $(t,u) \in \mathcal{TU}$.

Last, we show that

$$
\sup_{(t,u)\in\mathcal{TU}} \|(\widehat{\phi}_{t,u}(X) - \phi_{t,u}(X))\|_{\mathbb{P}_n,2} = O_p(\ell_n(\log(p\vee n)s)^{1/2}(nh_1)^{-1/2}).
$$

Let $\varepsilon_n = (\log(p \vee n)s/(nh_1))^{1/2}$, $\delta_n = (\log(p \vee n)s^2 \zeta_n^2/(nh_1))^{1/2}$, and

$$\mathcal{J}_{t,u} = \left\{ \Lambda(b(x)'\theta) : \begin{array}{c} \|\theta\|_0 \leq Ms, \|(\Lambda(b(X)'\theta) - \phi_{t,u}(X))K(\frac{T-t}{h})^{1/2}\|_{\mathbb{P}_n,2} \leq M\ell_n \varepsilon_n h_1^{1/2}, \\ \|\Lambda(b(X)'\theta) - \phi_{t,u}(X)\|_{\mathbb{P},\infty} \leq M\ell_n \delta_n. \end{array} \right\}$$

By (A.10), (A.11), and (A.13), for any $\varepsilon > 0$, there exists a constant M such that, with probability greater than $1 - \varepsilon$, $\hat{\phi}_{t,u}(\cdot) \in \mathcal{J}_{t,u}$ uniformly in $(t,u) \in \mathcal{TU}$. Therefore, with probability greater than $1 - \varepsilon$,

$$\left| \mathbb{P}_n (\hat{\phi}_{t,u}(X) - \phi_{t,u}(X))^2 \left[K(\frac{T-t}{h_1}) - \mathbb{E}(K(\frac{T-t}{h_1})|X) \right] \right|$$

$$\leq \sup_{(t,u) \in \mathcal{TU}} \sup_{J \in \mathcal{J}_{t,u}} \left| \mathbb{P}_n (J(X) - \phi_{t,u}(X))^2 \left[K(\frac{T-t}{h_1}) - \mathbb{E}(K(\frac{T_i-t}{h_1})|X) \right] \right| = \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}},$$

where $\mathcal{F} = \left\{ (J(X) - \phi_{t,u}(X))^2 \left[K(\frac{T-t}{h_1}) - \mathbb{E}(K(\frac{T_i-t}{h_1})|X) \right] : J \in \mathcal{J}_{t,u}, (t,u) \in \mathcal{TU} \right\}$ with bounded envelope. Note that,

$$\sigma^2 \equiv \sup_{f \in \mathcal{F}} \mathbb{E}f^2 \leq \sup_{(t,u) \in \mathcal{TU}} \sup_{J \in \mathcal{J}_{t,u}} \mathbb{E}(J(X) - \phi_{t,u}(X))^4 K^2(\frac{T-t}{h_1})$$

$$\lesssim \ell_n^2 \delta_n^2 \sup_{(t,u) \in \mathcal{TU}} \sup_{J \in \mathcal{J}_{t,u}} \mathbb{E}(J(X) - \phi_{t,u}(X))^2 K(\frac{T-t}{h_1})$$

$$= \ell_n^2 \delta_n^2 \sup_{(t,u) \in \mathcal{TU}} \sup_{J \in \mathcal{J}_{t,u}} \mathbb{E}\|(J(X) - \phi_{t,u}(X))K^{1/2}(\frac{T-t}{h_1})\|_{\mathbb{P}_n,2}^2$$

$$\lesssim \ell_n^4 \delta_n^2 \varepsilon_n^2 h_1,$$

In addition, we note that $\mathcal{F}$ is nested by

$$\overline{\mathcal{F}} = \left\{ (\Lambda(b(X)'\theta) - \phi_{t,u}(X))^2 \left[K(\frac{T-t}{h_1}) - \mathbb{E}(K(\frac{T_i-t}{h_1})|X) \right], \|\theta\|_0 \leq Ms, (t,u) \in \mathcal{TU} \right\},$$

such that

$$\sup_Q \log N(\mathcal{F}, e_Q, \varepsilon) \lesssim s \log(p \vee n) + s \log(\frac{1}{\varepsilon}) \vee 0.$$

Therefore, by Chernozhukov et al. (2014b, Corollary 5.1), we have

$$\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \lesssim \ell_n^2 \varepsilon_n h_1^{1/2} \delta_n s^{1/2} \log^{1/2}(p \vee n) n^{-1/2} + s \log(p \vee n) n^{-1} = o_p(\ell_n^2 \varepsilon_n^2 h_1). \tag{A.14}$$

Therefore,

$$h_1 \mathbb{P}_n (\hat{\phi}_{t,u}(X) - \phi_{t,u}(X))^2$$

$$\lesssim \mathbb{P}_n h_1 \int f_{t+h_1 v}(X) K(v) dv (\hat{\phi}_{t,u}(X) - \phi_{t,u}(X))^2$$

$$= \mathbb{P}_n (\hat{\phi}_{t,u}(X) - \phi_{t,u}(X))^2 \mathbb{E}\left(K\left(\frac{T-t}{h_1} \right) |X \right)$$

$$\leq \mathbb{P}_n (\hat{\phi}_{t,u}(X) - \phi_{t,u}(X))^2 K(\frac{T-t}{h_1}) + \left| \mathbb{P}_n (\hat{\phi}_{t,u}(X) - \phi_{t,u}(X))^2 \left[K(\frac{T-t}{h_1}) - \mathbb{E}(K(\frac{T-t}{h_1})|X) \right] \right|$$

$$= O_p(\ell_n^2 \varepsilon_n^2 h_1),$$

where the last equality holds due to (A.10) and (A.14). Canceling the h_1's on both sides, we obtain the desired result. ■

Proof of Theorem 3.2. By Belloni et al. (2017, Theorem 6.2), we have

$$\sup_t \|F_t(X) - \Lambda(b(X)'\hat{\beta}_t)\|_{\mathbb{P}_n,2} \lesssim_p \sqrt{\frac{s \log(p \vee n)}{n}}$$

and

$$\sup_t \|F_t(X) - \Lambda(b(X)'\hat{\beta}_t)\|_{\mathbb{P},\infty} \lesssim_p \sqrt{\frac{\zeta_n^2 s^2 \log(p \vee n)}{n}}$$

Then, we have

$$\|\hat{f}_t(X) - f_t(X)\|_{\mathbb{P}_n,2}$$
$$\leq \left\|\frac{\Lambda(b(X)'\hat{\beta}_{t+h_1}) - F_{t+h_1}(X)}{2h_1}\right\|_{\mathbb{P}_n,2} + \left\|\frac{\Lambda(b(X)'\hat{\beta}_{t-h_1}) - F_{t-h_1}(X)}{2h_1}\right\|_{\mathbb{P}_n,2}$$
$$+ \left\|\frac{F_{t+h_1}(X) - F_{t-h_1}(X)}{2h_1} - f_t(X)\right\|_{\mathbb{P}_n,2}$$
$$\lesssim_p \frac{1}{h_1}\sqrt{\frac{s\log(p \vee n)}{n}} + h_1^2$$

and similarly,

$$\|\hat{f}_t(X) - f_t(X)\|_{\mathbb{P},\infty} \lesssim_p \frac{1}{h_1}\sqrt{\frac{\zeta_n^2 s^2 \log(p \vee n)}{n}} + h_1^2. \quad \blacksquare$$

Proof of Theorem 3.3. Let $\hat{\alpha}^\dagger(t, u) = \mathbb{P}_n \eta \Pi_{t,u}(W_u, \widehat{\phi}_{t,u}, \hat{f}_t)$ where either $\eta = 1$ or η is a random variable that has sub-exponential tails with unit mean and variance. When $\eta = 1$, $\hat{\alpha}^\dagger(t, u) = \hat{\alpha}(t, u)$, which is our original estimator. When η is random, for $\bar{\eta} = \sum_{i=1}^n \eta_i/n$,

$$\hat{\alpha}^b(t, u) = \hat{\alpha}^\dagger(t, u)/\bar{\eta}$$

is the bootstrap estimator. In the following, we establish the linear expansion of $\hat{\alpha}^\dagger(t, u)$.

Recall $\varepsilon_n = (\log(p \vee n)s/(nh_1))^{1/2}$ and $\delta_n = (\log(p \vee n)s^2 \zeta_n^2/(nh_1))^{1/2}$. By Theorems 3.1 and 3.2, for any $\varepsilon > 0$, there exists a constant M such that, with probability greater than $1 - \varepsilon$, $\hat{f}_t(\cdot) \in \mathcal{G}_t$ uniformly in $t \in \mathcal{T}$ and $\widehat{\phi}_{t,u}(\cdot) \in \mathcal{J}_{t,u}$ uniformly in $(t, u) \in \mathcal{T}\mathcal{U}$. Here, we denote

$$\mathcal{G}_t = \left\{\begin{array}{c} \tilde{f}_t(X) \equiv (\Lambda(b(X)'\beta_{t+h_1}) - \Lambda(b(X)'\beta_{t-h_1}))/(2h_1) : \\ \|\beta_{t+h_1}\|_0 + \|\beta_{t-h_1}\|_0 \leq Ms, \quad \|\tilde{f}_t(X) - f_t(X)\|_{\mathbb{P},\infty} \leq M\delta_n h_1^{-1/2}, \\ \|\tilde{f}_t(X) - f_t(X)\|_{\mathbb{P}_n,2} \leq M\varepsilon_n h_1^{-1/2} \end{array}\right\}$$

and

$$\mathcal{J}_{t,u} = \left\{\begin{array}{c} \Lambda(b(x)'\theta) : \quad \|\theta\|_0 \leq Ms, \|(\Lambda(b(X)'\theta) - \phi_{t,u}(X))\|_{\mathbb{P}_n,2} \leq M\ell_n\varepsilon_n, \\ \|\Lambda(b(X)'\theta) - \phi_{t,u}(X)\|_{\mathbb{P},\infty} \leq M\ell_n\delta_n. \end{array}\right\}$$

We focus on the case in which $(\widehat{\phi}_{t,u}, \hat{f}_t) \in \mathcal{J}_{t,u} \times \mathcal{G}_t$. Then

$$\hat{\alpha}^\dagger(t, u) - \alpha(t, u) = (\mathbb{P}_n - \mathbb{P})\eta \Pi_{t,u}(W_u, \phi_{t,u}, f_t) + (\mathbb{P}_n - \mathbb{P})\left[\eta \Pi_{t,u}(W_u, \bar{\phi}, \bar{f}) - \eta \Pi_{t,u}(W_u, \phi_{t,u}, f_t)\right]$$

$$+ \mathbb{P}\left[\eta \Pi_{t,u}(W_u, \bar{\phi}, \bar{f}) - \eta \Pi_{t,u}(W_u, \phi_{t,u}, f_t)\right] + \left[\mathbb{P}\eta \Pi_{t,u}(W_u, \phi_{t,u}, f_t) - \alpha(t, u)\right]$$

$$:= I + II + III + IV,$$

where $(\bar{\phi}, \bar{f}) = (\widehat{\phi}_{t,u}, \hat{f}_t)$.

Below we fix $(\bar{\phi}, \bar{f}) \in \mathcal{J}_{t,u} \times \mathcal{G}_t$. First,

$$\text{Term } IV = \frac{\kappa_2 h_2^2}{2}\left[\mathbb{E}\left(\partial_t^2 \phi_{t,u}(X) + \frac{2\partial_t \phi_{t,u}(X)\partial_t f_t(X)}{f_t(X)}\right)\right] + o(h_2^2) = \beta_\alpha(t, u)h_2^2 + o(h_2^2).$$

where the $o(h_2^2)$ term holds uniformly in $(t, u) \in \mathcal{T}\mathcal{U}$. For term III, uniformly over $(t, u) \in \mathcal{T}\mathcal{U}$, we have

$$\mathbb{P}\eta\left[\Pi_{t,u}(W_u, \bar{\phi}, \bar{f}) - \Pi_{t,u}(W_u, \phi_{t,u}, f_t)\right]$$

$$= \mathbb{E}\left(\bar{\phi}(X) - \phi_{t,u}(X)\right)\left(1 - \frac{\mathbb{E}(K(\frac{T-t}{h_2})|X)}{h_2 f_t(X)}\right) + \mathbb{E}\left(\frac{Y_u - \bar{\phi}(X)}{\bar{f}(X)f_t(X)}\right)\left(\frac{f_t(X) - \bar{f}(X)}{h_2}\right)K(\frac{T-t}{h_2})$$

$$= O(\ell_n\delta_n h_2^2) + \mathbb{E}\left(\frac{Y_u - \bar{\phi}(X)}{\bar{f}(X)f_t(X)}\right)\left(\frac{f_t(X) - \bar{f}(X)}{h_2}\right)K(\frac{T-t}{h_2})$$

$$= O(\ell_n\delta_n h_2^2) + \mathbb{E}\left[\frac{f_t(X) - \bar{f}(X)}{\bar{f}_t(X)f_t(X)h_2}\mathbb{E}\left((\phi_{T,u}(X) - \phi_{t,u}(X))K(\frac{T-t}{h_2})\Big|X\right)\right]$$

$$+ \mathbb{E}\left[\frac{(f_t(X) - \bar{f}_t(X))(\phi_{t,u}(X) - \bar{\phi}(X))}{\bar{f}_t(X)f_t(X)h_2}\mathbb{E}K(\frac{T-t}{h_2}|X)\right]$$

$$=O(\ell_n \delta_n h_1^{-1/2} h_2^2) + O(\|(\phi_{t,u}(X) - \overline{\phi}(X))\|_{P,2} \|(f_t(X) - \overline{f}(X))\|_{P,2})$$
$$=O(\ell_n \delta_n h_1^{-1/2} h_2^2 + \ell_n \varepsilon_n^2 h_1^{-1/2}). \tag{A.15}$$

The second equality of (A.15) follows because there exists a constant c independent of n such that

$$\sup_{(t,u) \in \mathcal{TU}} \left| 1 - \frac{\mathbb{E}(K(\frac{T-t}{h_2})|X)}{h_2 f_t(X)} \right| \leq c h_2^2$$

and then

$$\mathbb{E}\left(\overline{\phi}(X) - \phi_{t,u}(X)\right)\left(1 - \frac{\mathbb{E}(K(\frac{T-t}{h_2})|X)}{h_2 f_t(X)}\right) \leq c h_2^2 \mathbb{E}\|\overline{\phi}(X) - \phi_{t,u}(X)\|_{P,\infty} = O(\ell_n \delta_n h_2^2).$$

The third equality of (A.15) holds because $\mathbb{E}(Y_u|X,T) = \phi_{T,u}(X)$. The fourth equality of (A.15) holds by the fact that $\|\overline{f}_t(X) - f_t(X)\|_{P,\infty} = O(\delta_n h_1^{-1/2}) = o(1)$, $f_t(x)$ is assumed to be bounded away from zero uniformly over t, τ and the Cauchy inequality. The fifth inequality of (A.15) holds because

$$\|(\phi_{t,u}(X) - \overline{\phi}(X))\|_{P,2} = [\mathbb{E}\|(\phi_{t,u}(X) - \overline{\phi}(X))\|_{\mathbb{P}_n,2}^2]^{1/2} = O(\ell_n \varepsilon_n)$$

and for some constant $c > 0$ independent of (t, u, n),

$$\|(f_t(X) - \overline{f}(X))\|_{P,2} = O(\varepsilon_n h_1^{-1/2}).$$

For the term II, we have

$$\mathbb{E}(\mathbb{P}_n - \mathbb{P})\eta\left[\Pi_{t,u}(W_u, \overline{\phi}, \overline{f}) - \Pi_{t,u}(W_u, \phi_{t,u}, f_t)\right] \leq \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$$

where

$$\mathcal{F} = \cup_{(t,u) \in \mathcal{TU}} \mathcal{F}_{t,u} \quad \text{and} \quad \mathcal{F}_{t,u} = \left\{ \eta\left[\Pi_{t,u}(W_u, \overline{\phi}, \overline{f}) - \Pi_{t,u}(W_u, \phi_{t,u}, f_t)\right] : \overline{\phi} \in \mathcal{J}_{t,u}, \overline{f} \in \mathcal{G}_t \right\}.$$

Note $\mathcal{F}$ has envelope $|\frac{\eta}{h_2}|$,

$$\sigma^2 := \sup_{f \in \mathcal{F}} \mathbb{E}f^2$$

$$\lesssim \sup_{(t,u) \in \mathcal{TU}, (\overline{\phi}, \overline{f}) \in \mathcal{J}_{t,u} \times \mathcal{G}_t} \mathbb{E}\left[(\overline{\phi}(X) - \phi_{t,u}(X))^2 \left(1 - \frac{K(\frac{T-t}{h_2})}{f_t(X)h_2}\right)^2\right]$$

$$+ \sup_{(t,u) \in \mathcal{TU}, (\overline{\phi}, \overline{f}) \in \mathcal{J}_{t,u} \times \mathcal{G}_t} \mathbb{E}\left[\frac{Y_u - \overline{\phi}(X)}{\overline{f}(X)f_t(X)h_2} K(\frac{T-t}{h_2})\left(f_t(X) - \overline{f}(X)\right)\right]^2$$

$$\lesssim \sup_{(t,u) \in \mathcal{TU}, (\overline{\phi}, \overline{f}) \in \mathcal{J}_{t,u} \times \mathcal{G}_t} \mathbb{E}\left[(\overline{\phi}(X) - \phi_{t,u}(X))^2\right]\left[1 + \frac{K^2(\frac{T-t}{h_2})}{h_2^2}\right]$$

$$+ \sup_{(t,u) \in \mathcal{TU}, (\overline{\phi}, \overline{f}) \in \mathcal{J}_{t,u} \times \mathcal{G}_t} \mathbb{E}\left[f_t(X) - \overline{f}(X)\right]^2 \frac{K^2(\frac{T-t}{h_2})}{h_2^2}$$

$$\lesssim \sup_{(t,u) \in \mathcal{TU}, (\overline{\phi}, \overline{f}) \in \mathcal{J}_{t,u} \times \mathcal{G}_t} h_2^{-1}\mathbb{E}\left[(\overline{\phi}(X) - \phi_{t,u}(X))^2\right] + h_2^{-1} \sup_{(t,u) \in \mathcal{TU}, (\overline{\phi}, \overline{f}) \in \mathcal{J}_{t,u} \times \mathcal{G}_t} \mathbb{E}\left[(\overline{f}(X) - f_t(X))^2\right]$$

$$\lesssim h_2^{-1} \varepsilon_n^2 h_1^{-1}.$$

The second last inequality in the above display holds because $f_t(x)$ is bounded away from zero uniformly in (t, x), where $t = T + h_2 v$ belongs to some compact enlargement of $\mathcal{T}$. Furthermore, $\mathcal{F}$ is nested by

$$\overline{F} = \left\{ \begin{array}{c} \Pi_{t,u}(W_u, \Lambda(b(X)'\theta), b(X)'\beta) - \Pi_{t,u}(W_u, \phi_{t,u}, f_t) : (t, u) \in \mathcal{TU}, \\ \|\theta\|_0 \leq Ms, \|\beta\|_0 \leq Ms \end{array} \right\},$$

such that

$$\sup_Q \log N(\mathcal{F}, e_Q, \varepsilon) \lesssim s \log(p \vee n) + s \log(\frac{1}{\varepsilon}) \vee 0.$$

In addition, we claim $\| \max_{1 \leq i \leq n} |\eta_i/h_2| \|_{p,2} \lesssim \log(n)h_2^{-1}$. When $\eta = 1$, the above claim holds trivially. When η has sub-exponential tail, the claim holds by van der Vaart and Wellner (1996, Lemma 2.2.2). Therefore, by Chernozhukov et al. (2014b, Corollary 5.1), we have

$$\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \lesssim \varepsilon_n (nh_1 h_2)^{-1/2} s^{1/2} \log^{1/2}(p \vee n) + \log(n)(nh_2)^{-1} s \log(p \vee n).$$

Combining the bounds for *II*, *III*, and *IV*, we have

$$\hat{\alpha}^{\dagger}(t,u) - \alpha(t,u) = (\mathbb{P}_n - \mathbb{P})\eta\Pi_{t,u}(W_u, \phi_{t,u}, f_t) + \beta_{\alpha}(t,u)h_2^2 + R_n(t,u)$$

and

$$\sup_{(t,u)\in\mathcal{TU}} |R_n(t,u)| = O_p(\varepsilon_n^2(h_2^{-1/2} + \ell_n h_1^{-1/2}) + \log(n)s\log(p \vee n)(nh_2)^{-1}) + o_p(h_2^2).$$

Then, when $\eta = 1$,

$$\hat{\alpha}(t,u) - \alpha(t,u) = (\mathbb{P}_n - \mathbb{P})\Pi_{t,u}(W_u, \phi_{t,u}, f_t) + \mathcal{B}_{\alpha}(t,u)h_2^2 + R_n(t,u)$$
$$= (\mathbb{P}_n - \mathbb{P})(\Pi_{t,u}(W_u, \phi_{t,u}, f_t) - \alpha(t,u)) + \mathcal{B}_{\alpha}(t,u)h_2^2 + R_n(t,u).$$

Then, Assumption 5 implies that $\sup_{(t,u)\in\mathcal{TU}} |R_n(t,u)| = o_p((nh_2)^{-1/2})$. For the bootstrap estimator, we have

$$\hat{\alpha}^b(t,u) - \alpha(t,u) = \hat{\alpha}^{\dagger}(t,u)/\bar{\eta} - \alpha(t,u)$$
$$= (\hat{\alpha}^{\dagger}(t,u) - \alpha(t,u))/\bar{\eta} + \alpha(t,u)(1/\bar{\eta} - 1)$$
$$= (\mathbb{P}_n - \mathbb{P})\eta\Pi_{t,u}(W_u, \phi_{t,u}, f_t)/\bar{\eta} + \alpha(t,u)(1/\bar{\eta} - 1) + \mathcal{B}_{\alpha}(t,u)h_2^2/\bar{\eta} + R_n(t,u)/\bar{\eta}$$
$$= (\mathbb{P}_n - \mathbb{P})\eta\Pi_{t,u}(W_u, \phi_{t,u}, f_t - \alpha(t,u))/\bar{\eta} + \mathcal{B}_{\alpha}(t,u)h_2^2/\bar{\eta} + R_n(t,u)/\bar{\eta}$$
$$= (\mathbb{P}_n - \mathbb{P})\eta(\Pi_{t,u}(W_u, \phi_{t,u}, f_t) - \alpha(t,u)) + \mathcal{B}_{\alpha}(t,u)h_2^2 + R_n^b(t,u), \tag{A.16}$$

where $\sup_{(t,u)\in\mathcal{TU}} |R_n^b(t,u)| = O_p(\varepsilon_n^2(h_2^{-1/2} + \ell_n h_1^{-1/2}) + \log(n)s\log(p \vee n)(nh_2)^{-1}) + o_p(h_2^2)$. This is because of the fact that

$$\bar{\eta} - \mathbb{E}\eta = \bar{\eta} - 1 = O_p(n^{-1/2}),$$

$$\sup_{(t,u)\in\mathcal{TU}} |R_n(t,u)| = O_p(\varepsilon_n^2(h_2^{-1/2} + \ell_n h_1^{-1/2}) + \log(n)s\log(p \vee n)(nh_2)^{-1}) + o_p(h_2^2),$$

and the collection of functions

$$\{\eta(\Pi_{t,u}(W_u, \phi_{t,u}, f_t) - \alpha(t,u)) : (t,u) \in \mathcal{TU}\}$$

satisfies

$$\sup_{(t,u)\in\mathcal{TU}} |(\mathbb{P}_n - \mathbb{P})(\eta(\Pi_{t,u}(W_u, \phi_{t,u}, f_t) - \alpha(t,u)))| = O_p(\log^{1/2}(n)(nh_2)^{-1/2}).$$

Therefore,

$$\hat{\alpha}^b(t,u) - \hat{\alpha}(t,u) = (\mathbb{P}_n - \mathbb{P})(\eta - 1)(\Pi_{t,u}(W_u, \phi_{t,u}, f_t) - \alpha(t,u)) + R_n^b(t,u) - R_n(t,u),$$

where

$$\sup_{(t,u)\in\mathcal{TU}} |R_n^b(t,u) - R_n(t,u)| = O_p(\varepsilon_n^2(h_2^{-1/2} + \ell_n h_1^{-1/2}) + \log(n)s\log(p \vee n)(nh_2)^{-1}) + o_p(h_2^2) = o_p((nh_2)^{-1/2}). \quad \blacksquare$$

Proof of Theorem 3.4. Let $\hat{\alpha}^*(t,u)$ be either the original or the bootstrap estimator of $\alpha(t,u)$. We first derive the linear expansion of the rearrangement of $\hat{\alpha}^*(t,u)$ defined in the proof of Theorem 3.3. For $z \in (0,1)$, let

$$F(t,z) = \int_0^1 1\{\alpha(t, \psi^{\leftarrow}(v)) \leq z\}dv, \quad F(t,z|d_n) = \int_0^1 1\{\hat{\alpha}^*(t, \psi^{\leftarrow}(v)) \leq y\}dv,$$

where $\psi(\cdot)$ is defined in Section 3.3. Then, by Lemma S.1.2 in the online supplement, we have

$$\frac{F(t,z|d_n) - F(t,z)}{r_n} + \frac{d_n(t, \psi(q_z(t)))\psi'(q_z(t))}{f_{Y(t)}(q_z(t))} = o_p(\Delta_n) \tag{A.17}$$

and

$$\frac{\hat{\alpha}^{*r}(t,u) - \alpha(t,u)}{r_n} + \frac{F(t, \alpha(t,u)|d_n) - F(t, \alpha(t,u))f_{Y(t)}(u)}{r_n \psi'(u)} = o_p(\Delta_n). \tag{A.18}$$

where $r_n = (nh_2)^{-1/2}$, $d_n(t,v) = (nh_2)^{1/2}(\hat{\alpha}^*(t, \psi^{\leftarrow}(v)) - \alpha(t, \psi^{\leftarrow}(v)))$, $f_{Y(t)}(\cdot)$ is the density of $Y(t)$, $q_z(t)$ is the zth quantile of $Y(t)$, and Δ_n equals either 1 or $h_2^{1/2}$, depending on either Assumptions 5.1 or 5.2 is in place.

Combining (A.17) and (A.18), we have

$$(nh_2)^{1/2}(\hat{\alpha}^{*r}(t,u) - \alpha(t,u)) = d_n(t, \psi(u)) + o_p(\Delta_n) = (nh_2)^{1/2}(\hat{\alpha}^*(t,u) - \alpha(t,u)) + o_p(\Delta_n) \tag{A.19}$$

uniformly over $(t,u) \in \mathcal{TU}$.

We can apply Lemma S.1.2 on $\hat{\alpha}^{*r}(t, u)$ again with $J_n(t, u) = (nh_2)^{1/2}(\hat{\alpha}^{*r}(t, u) - \alpha(t, u))$, $F(t, u) = P(Y(t) \le u) = \alpha(t, u)$, $f(t, u) = f_{Y(t)}(u)$, and $F^{\leftarrow}(t, \tau) = q_\tau(t)$. Then, for Δ_n equals 1 or $h_2^{1/2}$ under either Assumptions 5.1 or 5.2, respectively, we have,

$$\frac{\hat{q}_\tau^*(t) - q_\tau(t)}{r_n} = -\frac{J_n(t, q_\tau(t))}{f_{Y(t)}(q_\tau(t))} + o_p(\Delta_n) = -\frac{(nh_2)^{1/2}(\hat{\alpha}^{*r}(t, q_\tau(t)) - \tau)}{f_{Y(t)}(q_\tau(t))} + o_p(\Delta_n) \tag{A.20}$$

uniformly over $(t, \tau) \in \mathcal{TI}$.

When $\eta = 1$, combining (A.19), (A.20), and Theorem 3.3, we have

$$\hat{q}_\tau(t) - q_\tau(t) = -(\mathbb{P}_n - \mathbb{P})\frac{\Pi_{t,u}(W_{q_\tau(t)}, \phi_{t,q_\tau(t)}, f_t)}{f_{Y(t)}(q_\tau(t))} - \frac{\beta_\alpha(t, q_\tau(t))h_2^2}{f_{Y(t)}(q_\tau(t))} + R_n(t, \tau) + o_p(\Delta_n(nh_2)^{-1/2}).$$

By taking $\Delta_n = 1$ and $\Delta_n = h_2^{1/2}$ under Assumption 5.1 and 5.2, respectively, we have established the desired results. For the bootstrap estimator, by (A.16), we have

$$\hat{q}_\tau^b(t) - q_\tau(t) = -(\mathbb{P}_n - \mathbb{P})\eta\frac{\Pi_{t,u}(W_{q_\tau(t)}, \phi_{t,q_\tau(t)}, f_t)}{f_{Y(t)}(q_\tau(t))} - \frac{\beta_\alpha(t, q_\tau(t))h_2^2}{f_{Y(t)}(q_\tau(t))} + R_n^b(t, \tau) + o_p(\Delta_n(nh_2)^{-1/2}).$$

Then,

$$\hat{q}_\tau^b(t) - \hat{q}_\tau(t)$$
$$= -(\mathbb{P}_n - \mathbb{P})(\eta - 1)\Pi_{t,u}(W_{q_\tau(t)}, \phi_{t,q_\tau(t)}, f_t)/f_{Y(t)}(q_\tau(t)) + R_n^b(t, \tau) - R_n(t, \tau) + o_p(\Delta_n(nh_2)^{-1/2}).$$

By taking $\Delta_n = 1$ and $\Delta_n = h_2^{1/2}$ under Assumption 5.1 and 5.2, respectively, we have established the linear expansion of the bootstrap estimator too. Last, note that the bootstrap estimator cannot preserve the asymptotic bias term. For the validity of bootstrap inference, we need to under-smooth and require $nh_2^5 \to 0$. This condition is assumed in Theorem 4.1. ∎

Proof of Theorem 3.5. We consider the general case in which the observations are weighted by $\{\eta_i\}_{i=1}^n$ as above. For brevity, denote $\hat{\delta} := (\hat{\delta}_0, \hat{\delta}_1)' = (\hat{\beta}_\tau^{*0}(t), \hat{\beta}_\tau^{*1}(t))'$ and $\delta := (\delta_0, \delta_1)' = (\beta_\tau^0(t), \beta_\tau^1(t))$. For any variable $R_n := R_n(\tau, t)$ and some deterministic sequence r_n, we write $R_n = O_p^*(r_n)$ (resp. $o_p^*(r_n)$) if $\sup_{(t,\tau) \in \mathcal{TI}} |R_n(\tau, t)| = O_p(r_n)$ (resp. $o_p(r_n)$). Then $\hat{\delta} = \hat{\Sigma}_2^{-1}\hat{\Sigma}_1$, where

$$\hat{\Sigma}_1 = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n K(\frac{T_i-t}{h_2})\eta_i\hat{q}_\tau^*(T_i) \\ \frac{1}{n}\sum_{i=1}^n K(\frac{T_i-t}{h_2})(T_i - t)\eta_i\hat{q}_\tau^*(T_i) \end{pmatrix}$$

and

$$\hat{\Sigma}_2 = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n K(\frac{T_i-t}{h_2})\eta_i & \frac{1}{n}\sum_{i=1}^n K(\frac{T_i-t}{h_2})(T_i - t)\eta_i \\ \frac{1}{n}\sum_{i=1}^n K(\frac{T_i-t}{h_2})(T_i - t)\eta_i & \frac{1}{n}\sum_{i=1}^n K(\frac{T_i-t}{h_2})(T_i - t)^2\eta_i \end{pmatrix}.$$

Let $\Sigma_2 = \begin{pmatrix} f(t) & 0 \\ \kappa_2 f^{(1)}(t) & \kappa_2 f(t) \end{pmatrix}$ and $G = \begin{pmatrix} h_2^{-1} & 0 \\ 0 & h_2^{-3} \end{pmatrix}$. Then we have

$$G\hat{\Sigma}_2 - \Sigma_2 = O_p^*(\log^{1/2}(n)(nh_2^3)^{-1/2}).$$

In addition, note

$$\hat{q}_\tau^*(T_i) = \delta_0 + \delta_1(T_i - t) + (q_\tau(T_i) - \delta_0 - \delta_1(T_i - t)) + (\hat{q}_\tau^*(T_i) - q_\tau(T_i))$$

and

$$\begin{pmatrix} \frac{1}{nh_2}\sum_{i=1}^n K(\frac{T_i-t}{h_2})\eta_i\left(q_\tau(T_i) - \delta_0 - \delta_1(T_i - t)\right) \\ \frac{1}{nh_2^3}\sum_{i=1}^n K(\frac{T_i-t}{h_2})(T_i - t)\eta_i\left(q_\tau(T_i) - \delta_0 - \delta_1(T_i - t)\right) \end{pmatrix}$$
$$= \begin{pmatrix} \frac{1}{2}q_\tau''(t)f_T(t)\kappa_2 h_2^2 \\ \frac{1}{6}\partial_t(q_\tau''(t)f_T(t))\kappa_4 h_2^2 \end{pmatrix} + O_p^*(\sqrt{\frac{\log(n)h_2}{n}}) + o^*(h_2^2).$$

Therefore,

$$
G\widehat{\Sigma}_1 = G\widehat{\Sigma}_2\delta + \begin{pmatrix} \frac{1}{nh_2}\sum_{i=1}^n K(\frac{T_i-t}{h_2})\Big(\hat{q}_\tau(T_i)-q_\tau(T_i)\Big)\eta_i \\ \frac{1}{nh_2^3}\sum_{i=1}^n K(\frac{T_i-t}{h_2})(T_i-t)\Big(\hat{q}_\tau(T_i)-q_\tau(T_i)\Big)\eta_i \end{pmatrix} + \begin{pmatrix} \frac{1}{2}q_\tau''(t)f_T(t)\kappa_2 h_2^2 \\ \frac{1}{6}\partial_t(q_\tau''(t)f_T(t))\kappa_4 h_2^2 \end{pmatrix}
$$
$$
+ O_p^*(\sqrt{\tfrac{\log(n)h_2}{n}}) + o^*(h_2^2). \tag{A.21}
$$

Let $E(t,\tau) = \mathbb{E}\frac{Y_{q_\tau(t),j}-\phi_{t,q_\tau(t)}(X_j)}{f_t(X_j)h_2}K(\frac{T_j-t}{h_2})+\tau$. By Theorem 3.4, we have

$$
\hat{q}_\tau^*(t) - q_\tau(t)
$$
$$
= \frac{-1}{f_{Y_t}(q_\tau(t))}\frac{1}{n}\sum_{j=1}^n \eta_j\left(\frac{Y_{q_\tau(t),j}-\phi_{t,q_\tau(t)}(X_j)}{f_t(X_j)h_2}K(\frac{T_j-t}{h_2})+\phi_{t,q_\tau(t)}(X_j)-E(t,\tau)\right)
$$
$$
- \beta_q(t,\tau)h_2^2 + o_p^*((nh_2)^{-1/2}). \tag{A.22}
$$

Let $\Upsilon_i = (Y_i, T_i, X_i, \eta_i)$. Then, by plugging (A.22) in (A.21) and noticing that

$$
\begin{Bmatrix} \sup_{t\in\mathcal{T}}\frac{1}{nh_2}\sum_{i=1}^n K(\frac{T_i-t}{h_2})\eta_i \\ \sup_{t\in\mathcal{T}}\frac{1}{nh_2^3}\sum_{i=1}^n K(\frac{T_i-t}{h_2})|T_i-t|\eta_i \end{Bmatrix} = \begin{Bmatrix} O_p(1) \\ O_p(h_2^{-1}) \end{Bmatrix},
$$

we have

$$
G\widehat{\Sigma}_1 = G\widehat{\Sigma}_2\delta - \frac{1}{n(n-1)}\sum_{i\neq j}\eta_i\eta_j\Gamma(\Upsilon_i,\Upsilon_j;t,\tau) - \begin{pmatrix} f_T(t)\beta_q(t,\tau)h_2^2 \\ f_T'(t)\beta_q(t,\tau)h_2^2 \end{pmatrix}
$$
$$
+ \begin{pmatrix} \frac{1}{2}q_\tau''(t)f_T(t)\kappa_2 h_2^2 \\ \frac{1}{6}\partial_t(q_\tau''(t)f_T(t))\kappa_4 h_2^2 \end{pmatrix} + \begin{Bmatrix} o_p^*((nh_2)^{-1/2}) \\ o_p^*((nh_2^3)^{-1/2}) \end{Bmatrix}
$$

where $\Gamma(\Upsilon_i,\Upsilon_j;t,\tau) = (\Gamma_0(\Upsilon_i,\Upsilon_j;t,\tau),\Gamma_1(\Upsilon_i,\Upsilon_j;t,\tau))'$, and

$$
\Gamma_\ell(\Upsilon_i,\Upsilon_j;t,\tau)
$$
$$
= \frac{(T_i-t)^\ell}{h_2^{1+2\ell}f_{Y_{T_i}}(q_\tau(T_i))}K(\frac{T_i-t}{h_2})\left(\frac{Y_{q_\tau(T_i),j}-\phi_{T_i,q_\tau(T_i)}(X_j)}{f_{T_i}(X_j)h_2}K(\frac{T_j-T_i}{h_2})+\phi_{T_i,q_\tau(T_i)}(X_j)-E(T_i,\tau)\right)
$$

for $\ell = 0, 1$. Let $\Gamma^s(\Upsilon_i,\Upsilon_j;t,\tau) = (\Gamma(\Upsilon_i,\Upsilon_j;t,\tau)+\Gamma(\Upsilon_j,\Upsilon_i;t,\tau))/2$. Because $nh_2^7 \to 0$, we have

$$
\hat{\beta}_\tau^{1*}(t) - \beta_\tau^1(t) = -e_2'(G\widehat{\Sigma}_2)^{-1}U_n(t,\tau) + o_p^*((nh_2^3)^{-1/2}), \tag{A.23}
$$

where $e_2 = (0,1)'$ and $U_n(t,\tau) = (C_n^2)^{-1}\sum_{1\le i<j\le n}\eta_i\eta_j\Gamma^s(\cdot,\cdot;t,\tau)$ is a U-process indexed by (t,τ). By Lemma S.1.3 in the online supplement,

$$
e_2'(G\widehat{\Sigma}_2)^{-1}U_n(t,\tau) = \frac{-1}{n}\sum_{j=1}^n \eta_j(\kappa_2 f_{Y(t)}(q_\tau(t))f_t(X_j)h_2^2)^{-1}\left[Y_{q_\tau(t),j}-\phi_{t,q_\tau(t)}(X_j)\right]\overline{K}(\frac{T_j-t}{h_2}) + o_p^*((nh_2^3)^{-1/2}). \tag{A.24}
$$

Combining (A.23) and (A.24), we have

$$
\hat{\beta}_\tau^{1*}(t) - \beta_\tau^1(t) = \frac{-1}{n}\sum_{j=1}^n \eta_j(\kappa_2 f_{Y_t}(q_\tau(t))f_t(X_j)h_2^2)^{-1}\left[Y_{q_\tau(t),j}-\phi_{t,q_\tau(t)}(X_j)\right]\overline{K}(\frac{T_j-t}{h_2}) + o_p^*((nh_2^3)^{-1/2}). \quad\blacksquare
$$

Proof of Theorem 4.1. By the proofs of Theorems 3.4 and 3.5, we have

$$
\hat{q}_\tau^b(t) - \hat{q}_\tau(t) = -(\mathbb{P}_n-\mathbb{P})(\eta-1)\left(\Pi_{t,u}(W_{q_\tau(t)},\phi_{t,q_\tau(t)},f_t)-\tau\right)/f_t(q_\tau(t)) + o_p((nh_2)^{-1/2})
$$

and

$$
\hat{\beta}_\tau^{1b}(t) - \hat{\beta}_\tau^1(t) = \frac{1}{n}\sum_{j=1}^n (\eta_j-1)(\kappa_2 f_{Y(t)}(q_\tau(t))f_t(X_j)h^2)^{-1}\left[Y_{q_\tau(t),j}-\phi_{t,q_\tau(t)}(X_j)\right]\overline{K}(\frac{T_j-t}{h_2}) + o_p^*((nh_2^3)^{-1/2}).
$$

Then, it is straightforward to show that $\sqrt{nh_2}(\hat{q}_\tau^b(t)-\hat{q}_\tau(t))$ and $(nh_2^3)^{1/2}(\hat{\beta}_\tau^{1b}(t)-\hat{\beta}_\tau^1(t))$ converge weakly to the limiting distribution of $\sqrt{nh_2}(\hat{q}_\tau(t)-q_\tau(t))$ and $(nh_2^3)^{1/2}(\hat{\beta}_\tau^1(t)-\beta_\tau^1(t))$, respectively, conditional on data in the sense of van der Vaart and Wellner (1996, Section 2.9). The desired results then follow. $\blacksquare$

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2019.06.004.

References

Altonji, J.G., Matzkin, R.L., 2005. Cross section and panel data estimators for nonseparable models with endogenous regressors. Econometrica 73 (4), 1053–1102.

Andrews, D.W., Schafgans, M.M., 1998. Semiparametric estimation of the intercept of a sample selection model. Rev. Econom. Stud. 65 (3), 497–517.

Athey, S., Imbens, G., 2016. Recursive partitioning for heterogeneous causal effects. Proc. Natl. Acad. Sci. 113 (27), 7353–7360.

Begun, J.M., Hall, W., Huang, W.-M., Wellner, J.A., 1983. Information and asymptotic efficiency in parametric-nonparametric models. Ann. Statist. 11 (2), 432–452.

Belloni, A., Chen, M., Chernozhukov, V., 2016a. Quantile graphical models: prediction and conditional independence with applications to financial risk management. arXiv:1607.00286.

Belloni, A., Chen, D., Chernozhukov, V., Hansen, C., 2012. Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80 (6), 2369–2429.

Belloni, A., Chernozhukov, V., 2011. ℓ_1-penalized quantile regression in high-dimensional sparse models. Ann. Statist. 39 (1), 82–130.

Belloni, A., Chernozhukov, V., Chetverikov, D., Wei, Y., 2018. Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework. Ann. Statist. 46 (6B), 3643–3675.

Belloni, A., Chernozhukov, V., Fernández-Val, I., Hansen, C., 2017. Program evaluation with high-dimensional data. Econometrica 85 (1), 233–298.

Belloni, A., Chernozhukov, V., Hansen, C., 2014a. High-dimensional methods and inference on structural and treatment effects. J. Econ. Perspect. 28 (2), 29–50.

Belloni, A., Chernozhukov, V., Hansen, C., 2014b. Inference on treatment effects after selection among high-dimensional controls. Rev. Econom. Stud. 81 (2), 608–650.

Belloni, A., Chernozhukov, V., Kato, K., 2019. Valid post-selection inference in high-dimensional approximately sparse quantile regression models. J. Amer. Statist. Assoc. 114 (526), 749–758.

Belloni, A., Chernozhukov, V., Wei, Y., 2016b. Post-selection inference for generalized linear models with many controls. J. Bus. Econom. Statist. 34 (4), 606–619.

Bhattacharya, D., Mazumder, B., 2011. A nonparametric analysis of black–white differences in intergenerational income mobility in the united states. Quant. Econom. 2 (3), 335–379.

Bickel, P., Klaassen, C., Ritov, Y., Wellner, J., 1993. Efficient and Adaptive Estimation for Semiparametric Models. Springer-Verlag, New York.

Bickel, P.J., Ritov, Y., Tsybakov, A.B., 2009. Simultaneous analysis of Lasso and Dantzig selector. Ann. Statist. 37 (4), 1705–1732.

Browning, M., Carro, J., 2007. Heterogeneity and microeconometrics modeling. Econom. Soc. Monogr. 43, 47.

Bühlmann, P., van de Geer, S., 2011. Statistics for High-dimensional Data: Methods, Theory and Applications. Springer Science & Business Media.

Cai, Z., Xiao, Z., 2012. Semiparametric quantile regression estimation in dynamic models with partially varying coefficients. J. Econometrics 167 (2), 413–425.

Carneiro, P., Hansen, K.T., Heckman, J.J., 2003. Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college. Internat. Econom. Rev. 71 (44), 361–422.

Cattaneo, M.D., 2010. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. J. Econometrics 155 (2), 138–154.

Cattaneo, M.D., Farrell, M.H., 2011. Efficient estimation of the dose–response function under ignorability using subclassification on the covariates. In: Missing Data Methods: Cross-Sectional Methods and Applications. Emerald Group Publishing Limited, pp. 93–127.

Cattaneo, M.D., Jansson, M., Ma, X., 2019. Two-step estimation and inference with possibly many included covariates. Rev. Econom. Stud. 86 (3), 1095–1122.

Cattaneo, M.D., Jansson, M., Newey, W.K., 2016. Alternative asymptotics and the partially linear model with many regressors. Econometric Theory 34 (2), 277–301.

Cattaneo, M.D., Jansson, M., Newey, W.K., 2018. Inference in linear regression models with many covariates and heteroskedasticity. J. Amer. Statist. Assoc. 113, 1350–1361.

Chen, X., 2007. Large sample sieve estimation of semi-nonparametric models. In: Heckman, J.J., Leamer, E.E. (Eds.), Handbook of Econometrics, vol. 6. Elsevier, pp. 5549–5632.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., 2017. Double/debiased/neyman machine learning of treatment effects. Am. Econ. Rev. 107 (5), 261–265.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters. Econom. J. 21 (1), C1–C68.

Chernozhukov, V., Chetverikov, D., Kato, K., 2014a. Anti-concentration and honest, adaptive confidence bands. Ann. Statist. 42 (5), 1787–1818.

Chernozhukov, V., Chetverikov, D., Kato, K., 2014b. Gaussian approximation of suprema of empirical processes. Ann. Statist. 42 (4), 1564–1597.

Chernozhukov, V., Fernández-Val, I., Galichon, A., 2010. Quantile and probability curves without crossing. Econometrica 78 (3), 1093–1125.

Chernozhukov, V., Fernández-Val, I., Melly, B., 2013. Inference on counterfactual distributions. Econometrica 81 (6), 2205–2268.

Chernozhukov, V., Hansen, C., Spindler, M., 2015a. Post-selection and post-regularization inference in linear models with many controls and instruments. Am. Econ. Rev. 105 (5), 486–490.

Chernozhukov, V., Hansen, C., Spindler, M., 2015b. Valid post-selection and post-regularization inference: an elementary, general approach. Annu. Rev. Econ. 7 (1), 649–688.

Chernozhukov, V., Imbens, G.W., Newey, W.K., 2007. Instrumental variable estimation of nonseparable models. J. Econometrics 139 (1), 4–14.

Chesher, A., 2003. Identification in nonseparable models. Econometrica 71 (5), 1405–1441.

Chu, T., Zhu, J., Wang, H., 2011. Penalized maximum likelihood estimation and variable selection in geostatistics. Ann. Statist. 39 (5), 2607–2625.

Cunha, F., Heckman, J.J., Schennach, S.M., 2010. Estimating the technology of cognitive and noncognitive skill formation. Econometrica 78 (3), 883–931.

Farrell, M.H., 2015. Robust inference on average treatment effects with possibly more covariates than observations. J. Econometrics 189 (1), 1–23.

Firpo, S., 2007. Efficient semiparametric estimation of quantile treatment effects. Econometrica 75 (1), 259–276.

Frölich, M., Melly, B., 2013. Unconditional quantile treatment effects under endogeneity. J. Bus. Econom. Statist. 31 (3), 346–357.

Galvao, A.F., Wang, L., 2015. Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment. J. Amer. Statist. Assoc. 110, 1528–1542.

Graham, B.S., Imbens, G.W., Ridder, G., 2014. Complementary and aggregate implications of assortative matching: a nonparametric analysis. Quant. Econ. 5 (1), 29–66.

Hahn, J., 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica 66 (2), 315–331.

Hirano, K., Imbens, G.W., 2004. The propensity score with continuous treatments. In: Gelman, A., Meng, X.-L. (Eds.), Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives, vol. 226164. Wiley, pp. 73–84.

Hirano, K., Imbens, G.W., Ridder, G., 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71 (4), 1161–1189.

Hoderlein, S., Mammen, E., 2007. Identification of marginal effects in nonseparable models without monotonicity. Econometrica 75 (5), 1513–1518.

Hoderlein, S., Su, L., White, H., Yang, T.T., 2016. Testing for monotonicity in unobservables under unconfoundedness. J. Econometrics 193 (1), 183–202.

Imbens, G.W., Newey, W.K., 2009. Identification and estimation of triangular simultaneous equations models without additivity. Econometrica 77 (5), 1481–1512. http://dx.doi.org/10.3982/ECTA7108.

Kennedy, E.H., Ma, Z., McHugh, M.D., Small, D.S., 2017. Non-parametric methods for doubly robust estimation of continuous treatment effects. J. R. Stat. Soc. Ser. B Stat. Methodol. 79 (4), 1229–1245.

Khan, S., Tamer, E., 2010. Irregular identification, support conditions, and inverse weight estimation. Econometrica 78 (6), 2021–2042.

Kong, D., Bondell, H.D., Wu, Y., 2015. Domain selection for the varying coefficient model via local polynomial regression. Comput. Stat. Aata Anal. 83, 236–250.

Ledoux, M., Talagrand, M., 2013. Probability in Banach Spaces: Isoperimetry and Processes. Springer Science & Business Media.

Lee, E.R., Mammen, E., 2016. Local linear smoothing for sparse high dimensional varying coefficient models. Electron. J. Stat. 10 (1), 855–894.

Lewbel, A., Lu, X., Su, L., 2015. Specification testing for transformation models with an application to generalized accelerated failure-time models. J. Econometrics 184 (1), 81–96.

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

Matzkin, R.L., 1994. Chapter restrictions of economic theory in nonparametric methods. In: Handbook of Econometrics, vol. 4. Elsevier, pp. 2523–2558.

Matzkin, R.L., 2003. Nonparametric estimation of nonadditive random functions. Econometrica 71 (5), 1339–1375.

Newey, W.K., 1994. The asymptotic variance of semiparametric estimators. Econometrica 1349–1382.

Ning, Y., Liu, H., 2017. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. Ann. Statist. 45 (1), 158–195.

Pfanzagl, J., 1990. Estimation in Semiparametric Models. Springer, pp. 17–22.

Powell, D., 2010. Unconditional quantile regression for panel data with exogenous or endogenous regressors. Working paper.

Robins, J.M., Li, L., Mukherjee, R., Tchetgen, E.T., van der Vaart, A., 2017. Minimax estimation of a functional on a structured high-dimensional model. Ann. Statist. 45 (5), 1951–1987.

Robins, J.M., Rotnitzky, A., 1995. Semiparametric efficiency in multivariate regression models with missing data. J. Amer. Statist. Assoc. 90 (429), 122–129.

Robins, J.M., Rotnitzky, A., 2001. Comments on inference for semiparametric models: some questions and an answer. Statist. Sinica 11, 920–936.

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70 (1), 41–55.

Sasaki, Y., 2015. What do quantile regressions identify for general structural functions? Econometric Theory 31 (05), 1102–1116.

Su, L., Chen, Q., 2013. Testing homogeneity in panel data models with interactive fixed effects. Econometric Theory 29 (06), 1079–1135.

Su, L., Hoshino, T., 2016. Sieve instrumental variable quantile regression estimation of functional coefficient models. J. Econometrics 191 (1), 231–254.

Su, L., Jin, S., Zhang, Y., 2015. Specification test for panel data models with interactive fixed effects. J. Econometrics 186 (1), 222–244.

Tan, L., Zhang, Y., 2018. Root-n consistency of the intercept of a binary response model under tail restrictions. Econometric Theory 23 (6), 1180–1206.

Tsiatis, A., 2007. Semiparametric Theory and Missing Data. Springer Science & Business Media.

van der Vaart, A.W., Wellner, J.A., 1996. Weak Convergence and Empirical Processes: With Applications to Statistics. Springer.

van der Laan, M., Dudoit, S., 2003. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. University of California, Berkeley.

van der Laan, M., Robins, J.M., 2003. Unified Methods for Censored Longitudinal Data and Causality. Springer Science & Business Media.

van der Laan, M., Rose, S., 2011. Targeted Learning: Causal Inference for Observational and Experimental Data. Springer Science & Business Media.

van der Laan, M., Rubin, D., 2006. Targeted maximum likelihood learning. Int. J. Biostat. 2 (1).

Wager, S., Athey, S., 2018. Estimation and inference of heterogeneous treatment effects using random forests. J. Amer. Statist. Assoc. 113 (523), 1228–1242.

Yu, K., Jones, M., 1998. Local linear quantile regression. J. Am. Stat. Assoc. 93 (441), 228–237.