Nonparametric Identification and Estimation of Panel Quantile Models with Sample Selection

Sungwon Lee*
Department of Economics
Sogang University
sungwon.lee83@gmail.com

First draft: September 23, 2019
This draft: February 14, 2021
[Please click here for the latest version.]

Abstract

This paper develops nonparametric panel quantile regression models with sample selection. The class of models allows the unobserved heterogeneity to be correlated with time-varying regressors in a time-invariant manner. I adopt the correlated random effects approach proposed by Mundlak (1978) and Chamberlain (1980), and the control function approach to correct the sample selection bias. The class of models is general and flexible enough to incorporate many empirical issues, such as endogeneity of regressors and censoring. Identification of the model requires that $T \geq 3$, where T is the number of time periods, and that there is an excluded variable that affects the selection probability. I also suggest semiparametric models for practical implementation of estimation. Based on the identification result, this paper proposes sieve two-step estimation to estimate the model parameters and establishes the asymptotic theory for the sieve two-step estimators, including consistency, convergence rates, and asymptotic normality of functionals. A small Monte-Carlo simulation study with a semiparametric model confirms that the estimators perform well in finite samples.

Keywords: Sample selection, panel data, quantile regression, nonseparable models, correlated random effects, control function approach, nonparametric identification, sieve two-step estimation.
JEL Classification Numbers: C14, C21, C23.

*This paper has greatly benefited from discussions with Sukjin Han. I am also very grateful to Jason Abrevaya, Timothy Christensen, Gregory Cox, Christian Dustmann, Frank Kleibergen, Patrick Kline, Andres Santos, and participants in the 2020 World Congress of the Econometric Society, the 2020 Korean Economic Review International Conference, and the seminar at Sogang University for their valuable comments and suggestions. Any errors are mine.

1 Introduction

Sample selection is prevalent in economics. Since the seminal work of Gronau (1974) and Heckman (1979), sample selection has considerably received a lot of attention from both theoretical and applied econometrics due to its relevance and importance in many empirical contexts (e.g. Ahn and Powell (1993), Donald (1995), Das et al. (2003), and Newey (2009)). At the same time, quantile regression models have become a popular alternative to conditional mean models since the seminal work of Koenker and Bassett (1978) as they allow to investigate the distribution of the outcome variable and recover heterogeneous effects. Although many papers have studied sample selection and quantile regression, the literature on the intersection of them is relatively scarce as most papers have considered sample selection issues for conditional mean regression models. In particular, sample selection issues in quantile regression models for panel data have not been well-addressed, whereas the availability of panel data has become larger.

In this paper, I develop a nonseparable panel quantile model with sample selection and study identification and estimation of the model. Specifically, I consider the following panel quantile model:

$$\begin{aligned} Y_t^* =& m(X_t, U_t), \\ Y_t =& D_t Y_t^*, \end{aligned} \tag{1}$$

where t indicates time, Y_t^* is an outcome variable of interest, X_t is a vector of time-varying covariates, U_t is an unobserved heterogeneity, and D_t is a dummy variable indicating if it is selected. The structural function m is assumed to be strictly increasing with respect to its second argument for almost all X_t.

One of distinct features of the model in (1) is nonseparability between X_t and U_t. Many papers in the literature on sample selection develop models and estimators under different sets of assumptions, but they share some common feature that they focus on additively separable models. For quantile regression in the presence of sample selection, Buchinsky (1998) considered an additively separable quantile regression model for cross-sectional data. The additive separability facilitates identification and estimation of model parameters, but it considerably restricts the type of heterogeneity that can be allowed in a model. Nonseparability is important in quantile regression as (i) it can allow for various types of heterogeneous effects and (ii) it is less vulnerable to model misspecification.[1] Nonseparable quantile regression models with sample selection have been studied quite recently by Arellano and Bonhomme (2017) and Chernozhukov et al. (2018). While their models are semiparametric and mainly for cross-sectional data, this paper focuses on nonparametric quantile regression models for panel data. To my best knowledge, this paper is the first to consider nonseparable panel quantile regression models in the presence of sample selection.

Panel data models can incorporate time-invariant heterogeneity that may be correlated with time-varying regressors. When time-invariant heterogeneity is correlated with time-varying regressors, it is called time-invariant endogeneity.[2] One can resolve time-invariant endogeneity by taking

[1] Huber and Melly (2015) point out that the additive separability may lead to inconsistency of the estimator in the linear quantile regression models and propose a test for the structure.

[2] In the standard linear panel data models, the unobserved heterogeneity U_t is decomposed into two parts: one is a time-invariant error term, and the other is an time-varying idiosyncratic error. In this paper, I do not explicitly distinguish time-invariant components in U_t, but the dependence between time-varying regressors and time-invariant components in the error term is allowed in this paper. The dependence is the main motivation of the fixed effects model where X_t and time-invariant components are correlated in an arbitrary manner.

some differencing-based approach when the model is linear or additively separable, but it is much harder to deal with time-invariant endogeneity for nonlinear or nonseparable models. To overcome this difficulty, I consider a correlated random effects (CRE) approach which was originally pioneered by Mundlak (1978) and Chamberlain (1980). The main idea of the CRE approach is to assume that the distribution of the unobserved heterogeneity depends on the whole history of the time-varying covariates. In doing so, one can allow for time-invariant endogeneity as well as improve tractability of the model.

This paper provides conditions under which the model parameters are nonparametrically identified. The main idea of the identification strategy in this paper is to utilize variation in some excluded variables. Note that the model in this paper contains two types of endogeneity - time-invarant endogeneity and endogenous selection. Therefore, it is expected to have at least two excluded variables for identification. I show that one can use the rich information in panel data to deal with the time-invariant endogeneity, and this feature of the identification strategy requires that the number of time periods be greater than or equal to 3 and covariates have enough variation. On the other hand, I make use of a control function approach to correct for the selection bias, and this requires for an instrumental variable that varies the selection probability but does not directly affect the outcome. An exclusion restriction associated with the instrument, together with a conditional independence assumption, allows to resolve the endogenous selection, and this is a generalization of the approach of Heckman (1979). Under these standard identification conditions, the structural function of the outcome variable, which is denoted by $m(\cdot,\cdot)$, and the conditional distribution of the unobserved error term for the selected are nonparametrically identified. I also consider several extensions of the model to address some important empirical issues such as time-varying endogeneity and censoring. It is shown that the model in this paper can easily be extended to incorporate those issues, and therefore the class of models in this paper is very general and flexible.

While the fully nonparametric models are robust to model misspecification, they may not be tractable in estimation. In this regard, I propose two classes of semiparametric models: (i) semiparametric index models and (ii) additively separable models. These classes of models are very useful in a sense that one can reduce the dimension of some nonparametric object. Then, I provide conditions under which the parameters of the models are identified.

The identification result suggests a nonlinear optimization problem for estimation that the selection probability enters as a control function. Based on the identification result, I propose two-step nonparametric sieve estimation. The method of sieves provides a very flexible and general way to estimate semi-nonparametric or nonparametric models. The sieve method is also easy to implement in practice, and therefore it has been widely used. This paper provides the asymptotic theory for two-step nonparametric sieve estimators, including consistency, convergence rates, and asymptotic normality of smooth functionals.

Unlike the cross-sectional or time-series data, there are multiple types of data in terms of the number of individuals and the number of time periods, which are denoted by n and T, respectively, for panel data models. The relative magnitude between these two quantities defines the data structure, and this feature of the data structure is very important for panel data models as they are related to estimation of models. In this paper, I consider a fixed T-panel data model, and the fixed-T framework renders the model fit into data where T is much smaller than n. The large-T framework is frequently used in the literature on nonlinear fixed effects panel models to handle the incidental parameter problem (Neyman and Scott (1948)).[3] For panel quantile models with fixed effects, Koenker (2004), Canay (2011), Kato et al. (2012), and Besstremyannaya and Golovan (2019) make use of the large-T framework.[4] To adopt the large-T framework, however, the number of time periods in data should be larger than the number of individuals, and this requirement may not be appropriate to or suitable for some datasets, especially microdatasets or short panel datasets. In addition, not only the asymptotic properties of estimators, but finite-sample performances also depend on the magnitudes of n and T.[5] In this regard, estimators based on the large-T framework may be sensitive to the model specification and nature of data. On the other hand, I consider the fixed-T framework while incorporating time-invariant endogeneity, and this allows for a much wider applicability of the model in this paper.

I conduct a Monte-Carlo simulation study with a semiparametric model to examine the performance of estimators in finite samples. The results show that the semiparametric estimators have negligible biases and small standard deviations, which suggest that they perform well in finite samples.

Literature This paper is related to the literature on the panel data models with sample selection (e.g. Wooldridge (1995); Kyriazidou (1997); Semykina and Wooldridge (2010, 2013)).[6] For panel data in the presence of sample selection, Wooldridge (1995) and Kyriazidou (1997) propose estimators for panel data models where the outcome variable equation is linear in parameters. Wooldridge (1995) adopts the Mundlak-Chamberlain device (Mundlak (1978) and Chamberlain (1980)) to handle the time-invariant unobserved heterogeneity and uses the control function approach in the same spirit of Heckman (1979). The idea of Wooldridge (1995) is extended by Semykina and Wooldridge (2010) and Semykina and Wooldridge (2013) to incorporate time-varying endogeneity and dynamic panel data models, respectively. These papers, however, hugely rely on (semi-) parametric assumptions as well as the additive separability of the error term. Kyriazidou (1997) considers the conditional exchangeability assumption and the additively separable structure of the model. However, not only the additively separable error structure, but the conditional exchangeability condition may also fail to hold in some cases.[7] This paper differs from the aforementioned papers in that it considers nonseparable quantile regression models for panel data, whereas Wooldridge (1995) and Semykina and Wooldridge (2010) consider linear conditional mean models.

This paper is also related to the studies in the literature on (nonparametric) identification and estimation with endogeneity. This literature is too large to list all related papers, and one may refer to Matzkin (2007) for a comprehensive review. Focusing on sample selection, this paper is closely related to, for example, Buchinsky (1998), Das et al. (2003), and Newey (2009). The model varies across them, but they make use of the control function approach to correct for the sample selection

[3] Fernández-Val and Weidner (2018) provide a comprehensive review on the literature on large-T panel data models.

[4] Canay (2011) originally imposed a condition that $n/T^s \to 0$ for some $s > 1$ to establish consistency and asymptotic normality. Under this condition, one may use short panel data where n grows faster than T. Besstremyannaya and Golovan (2019), however, point out that the rate condition is not sufficient for existence of a limiting distribution or for zero mean of a limiting distribution. This result in Besstremyannaya and Golovan (2019) suggests that the estimator of Canay (2011) does not fit into short panel data.

[5] The simulation results in Kato et al. (2012) show that the root mean squared error is quite large when T is small in the location-scale shift model.

[6] One can refer to Dustmann and Rochina-Barrachina (2007) for comparison of some estimators including Wooldridge (1995) and Kyriazidou (1997).

[7] A related discussion can be found in Altonji and Matzkin (2005).

bias. This paper shares some common with them as the identification strategy in this paper also utilizes a control function, but differs from them as the model in this paper is nonseparable and for panel data. As mentioned earlier, Arellano and Bonhomme (2017) and Chernozhukov et al. (2018) consider nonseparable quantile regression models with sample selection, but their models are more fitting into cross-sectional data. Furthermore, they consider semiparametric model specifications for estimation. In contrast, this paper considers a class of nonparametric models with sample selection for panel data, and the identification or estimation strategy developed in this paper does not impose such distributional assumptions. As a result, this paper extends Arellano and Bonhomme (2017) and Chernozhukov et al. (2018) to nonparametric quantile regression models for panel data.

The literature on (nonlinear) panel data models is another area that this paper is closely related to. One of the features of the model in this paper is that I adopt the CRE approach, and this approach is also widely used in the literature to address time-invariant endogeneity. Abrevaya and Dahl (2008) make use of the CRE approach for linear panel quantile regression, but this paper differs from them as it considers nonparametric nonseparable models. Bester and Hansen (2009) study identification of the marginal effects in general panel data models with the CRE approach with a focus on identification of marginal effects, and therefore this paper is different from theirs in terms of the parameter of interest and identification/estimation strategy. Arellano and Bonhomme (2016) recently consider CRE specifications and develop a class of tractable nonlinear panel models, but their identification strategy relies on a high-level condition called injectivity, which is related to the completeness condition. On the other hand, this paper adopts a control function approach for identification. For general nonlinear panel data models, Altonji and Matzkin (2005) study identification and estimation of local average responses (LARs) and structural functions under an assumption called exchangeability. While the exchangeability condition generally implies some shape restrictions on the distribution of the unobserved error term, I circumvent to restrict the shape of the distribution of the error term by taking the CRE approach. Hoderlein and White (2012), Chernozhukov et al. (2013), and Chernozhukov et al. (2015) study identification of average structural functions and quantile structural functions, but this paper considers identification and estimation of the structural functions. Evdokimov (2010) studies identification and estimation of a class of panel data models, but his identification is based on deconvolution. Therefore, the identification strategy in this paper is completely different from his. More importantly, none of them address sample selection issues which are the main focus of this paper.

Outline The rest of this paper is organized as follows. In section 2, I introduce the model and consider some extensions of the model. Section 3 considers identification of the model, and section 4 presents several semiparametric models. Section 5 proposes two-step sieve estimation based on the identification result. Section 6 establishes the asymptotic theory for the nonparametric sieve two-step estimators. Section 8 concludes and discusses future work. All mathematical proofs for the asymptotic theory are presented in the appendix.

Notation I introduce some notation. For a vector A, $A^{'}$ denotes the transpose of A. For a generic random variable A_t, the support of A_t is denoted by $Supp(A_t)$. Let $\mathbf{A} \equiv (A_1, A_2, ..., A_T)^{'}$ be the random vector consisting of A_t's from time period 1 to T. and let $\mathbf{A}_{-t} \equiv (A_1, ..., A_{t-1}, A_{t+1}, ..., A_T)^{'}$ be the random vector consisting of A_t's from time period 1 to T but not t. I use notation $\mathbf{A}_{-t,s}$ to denote the random vector consisting of A_t's from time period 1 to T but not t and s. Realizations of A and $\mathbf{A}$ are denoted by a and $\mathbf{a}$, respectively.[8] For two random variables A and B and for any $u \in (0,1)$, $Q_{A|B}(u|b)$ indicates the u-th conditional quantile of A on $B = b$, and $F_{A|B}(a|b)$ is the conditional distribution function of A given $B = b$. $\mathbb{E}[\cdot]$ is the expectation operator.

2 The Model

I consider the following general non-separable panel data model:

$$Y_t^* = m(X_t, U_t), \tag{2}$$

where $Y_t^* \in \mathbb{R}$ is an outcome variable of interest, $X_t \in \mathbb{R}^{d_x}$ is a vector of time-varying covariates, and $U_t \in \mathbb{R}$ is an unobserved error term. I assume that $m(x, \cdot)$ is strictly increasing for almost all $x \in Supp(X_t)$ for all $t = 1, 2, ..., T$ and that $\{U_t : t = 1, 2, ..., T\}$ is stationary. Since the quantile operator is preserved under a monotone transformation, it is straightforward to see that for any $u \in \mathcal{U} \subseteq (0, 1)$,

$$Q_{Y_t^*|\mathbf{X}}(u|\mathbf{X}) = m(X_t, Q_{U_t|\mathbf{X}}(u|\mathbf{X}); u). \tag{3}$$

Note that the structural function m is allowed to vary across quantile levels.

It is common to assume that the unobserved error term U_t can be decomposed into time-invariant individual heterogeneity and time-varying idiosyncratic terms and that the time-invariant individual heterogeneity may be correlated with X_t. For the standard linear panel data model, such time-invariant heterogeneity can be eliminated by taking difference. For nonlinear models, however, the approach based on differencing does not work in general.

To overcome the difficulty in identification and estimation of the model with short panels, I adopt the CRE approach. Specifically, I assume that the conditional quantile function of U_t given $\mathbf{X}$ is an unknown function of $\mathbf{X}$. This is motivated by the CRE approach which was pioneered by Mundlak and Chamberlain (Mundlak (1978); Chamberlain (1980, 1982)). The CRE approach provides an effective way to deal with the unobserved heterogeneity in nonlinear panel models and it has been widely considered in the literature. Abrevaya and Dahl (2008) propose a linear panel quantile model, and Bester and Hansen (2009) investigate identification of marginal effects in a class of nonseparable panel models.[9] Both of them utilize some CRE approach to handle the unobserved individual effects with short panels. Arellano and Bonhomme (2016) recently develop a tractable estimation strategy for nonseparable panel data models based on the CRE approach.

The class of models in this paper is also related to the correlated random coefficient models in the literature (e.g. Arellano and Bonhomme (2012); Graham and Powell (2012); Laage (2019)). For quantile regression, Graham et al. (2018) consider linear panel quantile models with random coefficients, building upon Graham and Powell (2012). However, the model of this paper differs from those in that I consider a nonparametric structural function m with a scalar error term, whereas they consider a parametric structural function for m with a multi-dimensional error structure.

[8] Note that, however, I use u for the quantile level index throughout the paper, and thus u is not a realization of the random variable U_t in (1).

[9] The class of models considered in this paper encompasses the linear panel quantile regression models in Abrevaya and Dahl (2008) as a special case, and thus it can be viewed as a nonparametric generalization of the linear panel quantile models with correlated random effects.

Below I present some illustrative examples that fit into the class of CRE models in (3).

Example 2.1 (Random Coefficient Model). *Suppose that* $Supp(X_t) = \mathbb{R}$ *and that the data generating process is as follows:*

$$Y_t^* = \exp(X_t)U_t.$$

It is obvious that $Q_{Y_t^*|\mathbf{X}}(u|\mathbf{X}) = \exp(X_t) \cdot Q_{U_t|\mathbf{X}}(u|\mathbf{X})$, *and hence the structural function* $m(x, \gamma; u) = x \cdot \gamma$ *for all* $u \in \mathcal{U} \subseteq (0, 1)$.

Example 2.2 (Linear Panel Quantile Model). *Abrevaya and Dahl (2008) propose a class of linear panel quantile models as follows:*

$$Y_t^* = X_t^{'}\beta(u) + \alpha(u) + \epsilon_t(u),$$
$$\alpha(u) = \mathbf{X}^{'}\delta(u) + c(u),$$

where $\alpha(u)$ *is an unobserved time-invariant heterogeneity,* $c(u)$ *is an unobserved error term, and* $Q_{c(u)+\epsilon_t(u)}(u|\mathbf{X}) = 0$. *It is straightforward to see that* $Q_{Y_t^*|\mathbf{X}}(u|\mathbf{X}) = X_t^{'}\beta(u) + \mathbf{X}^{'}\delta(u)$ *under the restriction on the model. This class of models is a special case of (3). Specifically, one can set* $U_t = \alpha + \epsilon_t$ *and* $m(X_t, \gamma; u) = X_t^{'}\beta(u) + \gamma$. *The conditional quantile of* U_t *given* $\mathbf{X}$ *is equal to* $\mathbf{X}^{'}\delta(u)$.

Example 2.3 (Panel Quantile Model). *Arellano and Bonhomme (2016) consider the following model as an example:*

$$Y_t^* = X_t^{'}\beta(\epsilon_t) + \alpha\delta(\epsilon_t),$$
$$\alpha = \mathbf{X}^{'}\mu(V),$$

where for all $t = 1, 2, ..., T$, ϵ_t *and* V *are uniformly distributed conditional on* $\mathbf{X}$ *and* α *is an unobserved time-invariant heterogeneity. As pointed out in Arellano and Bonhomme (2016), this model is a generalization of the standard linear quantile models of Koenker and Bassett (1978) to panel data. Assuming that the map* $u \mapsto X_t^{'}\beta(u) + \mathbf{X}_i^{'}\mu(u) \cdot \delta(u)$ *is strictly increasing and that* ϵ_t *and* V *are comonotonic, it can be shown that* $Q_{Y_t^*|\mathbf{X}}(u|\mathbf{X}) = X_t^{'}\beta(u) + \mathbf{X}^{'}\theta(u)\delta(u)$.[10] *Letting* $U_t \equiv X_t^{'}\{\beta(\epsilon_t) - \beta(u)\} + \mathbf{X}^{'}\mu(V) \cdot \delta(\epsilon_t)$, $m(x, \gamma; u) = x^{'}\beta(u) + \gamma$ *and* $Q_{U_t|\mathbf{X}}(u) = \mathbf{X}^{'}\mu(u)\delta(u) \equiv \mathbf{X}^{'}\theta(u)$.

In examples 2.2 and 2.3, although there are two unobserved error terms, they can be collapsed into a scalar error term. While additivity plays the role of putting them together in example 2.2, comonotonicity of ϵ_t and V enables to collapse the error terms into a scalar error in example 2.3 where the unobserved error terms are nonlinearly enter. Therefore, the class of generalized CRE models in (3) is quite flexible and general.

Based on (2) and (3), I develop a panel quantile model with sample selection. Let $\Pr(D_t = 1|X_t = x, Z_t = z) \equiv h_t(x, z)$ be the propensity score (or selection probability), where $Z_t \in \mathbb{R}^{d_z}$ is a vector of excluded variables and $\mathbf{Z} = (Z_1, Z_2, ..., Z_T)^{'}$. The selection probability conditioning

[10] For the definition of comonotonicity, one may refer to Koenker (2005, p.60).

on X_t and Z_t is denoted by P_t (i.e. $P_t \equiv h_t(X_t, Z_t)$). The random vector $\mathbf{W} \equiv (\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{D})$ is observed from the data. In the presence of sample selection, it is well-known that using only selected observations usually yields a sample selection bias, and thus it is necessary to correct such a bias. In this paper, I adopt the control function approach to correct the sample selection bias. The control function approach to sample selection was originally proposed by Heckman (1979), and it has been adapted to various models. Specifically, I impose the following assumption:

Assumption 1. *Let $u \in \mathcal{U}$ be given. For all $t \in \{1, 2, ., T\}$,*

$$Q_{U_t|\mathbf{X},Z_t,D_t=1}(u|\mathbf{X}, Z_t, D_t = 1) = r(\mathbf{X}, h_t(X_t, Z_t); u), \tag{4}$$

where r is an unknown measurable function.

Note that r is allowed to take a different form across the quantile level u, and thus one can infer the conditional distribution function of U_t from the conditional quantile function of U_t, $r(\mathbf{X}, P_t; u)$. As a consequence, the way to correct the sample selection bias in this paper is to implicitly modify the conditional distribution function of the unobserved error term, and this is similar to that of Buchinsky (1998). However, it is different from the way that is considered in Arellano and Bonhomme (2017) or Chernozhukov et al. (2018) in that I do not impose any parametric or semiparametric structure on the conditional distribution of U_t. In sum, I consider the following model in this paper:

$$\begin{aligned} &Y_t^* = m(X_t, U_t), \\ &Y_t = D_t Y_t^*, \\ &Q_{U_t|\mathbf{X},Z_t,D_t=1}(u|\mathbf{X}, Z_t, D_t = 1) = r(\mathbf{X}, h_t(X_t, Z_t); u). \end{aligned} \tag{5}$$

From (5), it is straightforward to see that

$$\begin{aligned} Q_{Y_t|\mathbf{X},Z_t,D_t=1}(u|\mathbf{X}, Z_t, D_t = 1) &= m(X_t, Q_{U_t|\mathbf{X},Z_t,D_t=1}(u|\mathbf{X}, Z_t, D_t = 1)) \\ &= m(X_t, r(\mathbf{X}, h_t(X_t, Z_t); u); u). \end{aligned} \tag{6}$$

Assumption 1 shares a common feature with the models of Buchinsky (1998), Das et al. (2003) and Newey (2009) in that the selection bias is adjusted by including a control function, and this can be considered as a generalization of the control function approach in Heckman (1979). This paper, however, differs from Buchinsky (1998) in that neither any parametric restriction nor additivity of the error term is imposed on model (5).[11] Therefore, this paper extends the additive semiparametric quantile models for cross-sectional data in Buchinsky (1998) to nonseparable quantile regression models for panel data. Das et al. (2003) and Newey (2009) study identification and estimation of nonparametric sample selection models with an additive error term. This paper differs from them in that model (5) does not impose such an additive separability and thus it allows much various types of heterogeneity. In addition, this paper considers quantile regression, whereas they focus on conditional mean functions.

The structural functions m and r are related to many objects of interest. I first define the local structural function (LQSF) in time t as follows:

[11] Buchinsky (1998) considers a class of models where m is characterized by some finite-dimensional parameter. Specifically, his model is written as $m(X, \epsilon) = X'\beta + \epsilon$.

Definition 2.1 (Local Quantile Structural Function (LQSF)). *The local u-th quantile structural function (LQSF) at* $(X_t, r(\mathbf{X}, P_t)) = (x, \gamma)$ *in time t is*

$$q_t^{local}(u, x, \gamma) \equiv Q_{Y_t^*|X_t, r(\mathbf{X}, P_t)}(u|X_t = x, r(\mathbf{X}, P_t) = \gamma).$$

Note that the definition of the LQSF is similar to, but slightly different from that of Fernández-Val et al. (2019). It is straightforward to see that the structural function $m(x, \gamma; u)$ in this paper corresponds to the LQSF. Related to the LQSF, one can consider the quantile structural function (QSF) which was introduced by Imbens and Newey (2009). The following definition of the QSF is a generalization of the QSF in Imbens and Newey (2009) to that for panel data:

Definition 2.2 (Quantile Structural Function (QSF)). *The u-th quantile structural function (QSF) in time t evaluated at* $X_t = x$ *is*

$$q_t(u, x) \equiv \mathbb{E}[q_t^{local}(u, x, r(\mathbf{X}, P_t))].$$

The LQSF and QSF are parameters of interest in many empirical analyses and closely related to the (local) quantile treatment effect of changing X_t. To make it concrete, I provide the definitions of the local and the average quantile treatment effects of X_t below:

Definition 2.3 (Local Quantile Treatment Effect (LQTE)). *The u-th local quantile treatment effect in time t of changing* X_t *from* x_0 *to* x_1 *at* $r(\mathbf{X}, P_t) = \gamma$ *is*

$$LQTE_t(u, x_0, x_1, \gamma) \equiv q_t^{local}(u, x_1, \gamma) - q_t^{local}(u, x_0, \gamma)$$

Definition 2.4 (Quantile Treatment Effect (QTE)). *The u-th average conditional quantile effect in time t of changing* X_t *from* x_0 *to* x_1 *is*

$$\begin{aligned} QTE_t(u, x_0, x_1) &\equiv \int q_t^{local}(u, x_1, r(\mathbf{x}, h(x, z))) - q_t^{local}(u, x_0, r(\mathbf{x}, h(x, z))) dF_{\mathbf{X}, Z_t}(\mathbf{x}, z) \\ &= q_t(u, x_1) - q_t(u, x_0). \end{aligned}$$

If X_t is continuous, then the LQTE and QTE can be interpreted as the local and average marginal effects, respectively. Many objects that are similar to the LQTE or the QTE are considered in the literature on nonseparable panel data models (e.g. Altonji and Matzkin (2005); Bester and Hansen (2009); Imbens and Newey (2009); Hoderlein and White (2012); Chernozhukov et al. (2013, 2015)). It is clear to see that the LQTE and QTE are functionals of the structural functions m and r from the definitions.

3 Nonparametric Identification

3.1 Main Results

In this section, I consider identification of the model parameters. The identification strategy is based on the model implication in (6), and the main objects of interest in (5) are $m(\cdot, \cdot; u)$ and $r(\cdot, \cdot; u)$. Note that the conditional selection probability at time t, $h_t(x, z)$, is identified from the data. As shown earlier, one can answer many questions that are empirically relevant, such as the marginal effect of X_t on the conditional quantile of Y_t^*, through identification of these objects. To achieve identification of $m(\cdot,\cdot;u)$ and $r(\cdot,\cdot;u)$, I impose the following assumption:

Assumption 2. *Let $T \geq 3$. For any $u \in \mathcal{U} \subseteq (0,1)$, the following conditions hold:*

(i) For each $t = 1,2,...,T$, there exists a known value $\bar{x}(u) \in Supp(X_t) \subseteq \mathbb{R}^{d_x}$ such that $m(\bar{x}(u),\gamma;u) = \gamma$;

(ii) Let $x \in Supp(X_t)$ and $\gamma \in Supp(r(\mathbf{X}, h(X_t, Z_t)))$ be given. For any $t,s \in \{1,2,...,T\}$ with $t \neq s$, there exists a non-empty subset $\tilde{\mathcal{X}}_{-t,s}(x,\bar{x}(u))$ of $Supp(\mathbf{X}_{-t,s}|(X_t,X_s) = (x,\bar{x}(u)))$ and such that, for any $\mathbf{x}_{-t,s} \in \tilde{\mathcal{X}}_{-t,s}(x,\bar{x}(u))$, $r(\mathbf{x}_0,p) = \gamma$ for some $p \in Supp(h(x,Z_t)|\mathbf{X} = \mathbf{x}_0)$ and $\Pr(\mathbf{X}_{-t,s} \in \tilde{\mathcal{X}}_{-t,s}(x,\bar{x}(u))) > 0$, where $\mathbf{x}_0 = (X_t = x, X_s = \bar{x}(u)), \mathbf{X}_{-t,s} = \mathbf{x}_{-t,s})$;

(iii) For any $t = 1,2,...,T$ and for any $x \in Supp(X_t)$ and $z \in Supp(Z_t) \subseteq \mathbb{R}^{d_z}$, there exists a non-empty set $\mathcal{Z}_s(z) \subseteq Supp(Z_s)$ for some $s \in \{1,2,...,T\}$ such that, for any $\tilde{z} \in \mathcal{Z}_s(z)$, $h_t(x,z) = h_s(\bar{x}(u),\tilde{z})$ and $\Pr(Z_s \in \mathcal{Z}_s(z)) > 0$.

Condition (i) is a normalization. Theorem 3.1 in Matzkin (2007) implies that it is necessary to impose a normalization to identify function $m(\cdot,\cdot;u)$. Note that the value $\bar{x}(u)$ may differ across the quantile indices, but I assume that $\bar{x}(u)$ remains the same across $u \in \mathcal{U}$ for simplicity.

Condition (ii) is implied by sufficient variation in $\mathbf{X}_{-t,s}$. This variable can be viewed as an excluded variable that provides a source of exogenous variation to r while fixing X_t and X_s for some t and s. To illustrate how this condition is used for identification, consider the linear panel quantile model in Example 2.2 with assuming that $T = 3$ and $d_x = 1$. In addition, I ignore the sample selection issues, and therefore Y_t^* is observed for everyone, to elucidate the role of condition (ii) in identification analysis. Note that $\bar{x}(u) = 0$, $m(x,\gamma) = x\beta(u) + \gamma$, and $r(\mathbf{x}) = \mathbf{x}'\delta(u)$. Let $x \in Supp(X_1)$ be given. Then, one can show that $Q_{Y_1^*|\mathbf{X}}(u|\mathbf{X}_0) = x\beta(u) + r(\mathbf{X}_0)$ and that $Q_{Y_2^*|\mathbf{X}}(u|\mathbf{X}_0) = r(\mathbf{X}_0) = x\delta_1(u) + X_3\delta_3(u)$, where $\delta(u) = (\delta_1(u),\delta_2(u),\delta_3(u))'$ and $\mathbf{X}_0 = (x,0,X_3)'$. Condition (ii) ensures that one can find a set of values of X_3 such that for a given $\gamma \in Supp(r(\mathbf{X}))$, $x\delta_1(u) + X_3\delta_3(u) = \gamma$. Therefore, a necessary condition in this illustration that guarantees condition (ii) in Assumption 2 is that $\delta_3(u) \neq 0$. If X_3 has enough variation conditioning on $X_1 = x$ and $X_2 = 0$ and $\delta_3(u) \neq 0$, then condition (ii) for this example is satisfied. This is similar to Assumption 2 in Imbens and Newey (2009) that a large support condition for the excluded variable is satisfied. Since the model in this paper allows for time-invariant endogeneity and it is captured by the CRE specification, X_t can be considered endogenous (in the time-invariant manner), and the covariates in other time periods are used as an excluded variable that helps resolve the time-invariant endogeneity. It is worth pointing out that more than three time periods gives additional exogenous variation that can be used to identify the structural functions and therefore having more than 3 time periods provide additional identification power.

A related assumption to condition (ii) is exchangeability considered by Altonji and Matzkin (2005). Exchangeability typically places some restriction on the admissible class of functions for r, it may help weaken assumptions on variation in the excluded variable. Compared to the exchangeability assumption, condition (ii) is likely to require stronger conditions on the support of $\mathbf{X}_{-t,s}$, but it does not impose any shape restrictions on the class of functions that r belongs to. More importantly, exchangeability may not be plausible to be assumed with panel data where t indicates time. Exchangeability is related to symmetry of the effects of covariates on the distribution of the unobserved error term, and thus the effect of a change in X_s is the same (or similar) to that of a change in X_t for some $t \neq s$. In this regard, exchangeability may be consistent with some variants of the form in Mundlak (1978) in a sense that the correlated random effects specification of Mundlak (1978) is the average of X_t's over time and therefore the effects of X_t and X_s with $t \neq s$ are symmetric. On the other hand, I do not impose such restrictions on the model so that one can consider more flexible specifications for r. One can refer to Altonji and Matzkin (2005, pp.1062-1066) for further discussion on the exchangeability condition and CRE approach.

Condition (iii) requires variation in the excluded variable Z_t, which is an instrumental variable. This condition also requires that the excluded variable Z_t affect the selection probability, so one can use the variation in Z_t and Z_s to match the selection probabilities in time periods t and s. This condition is needed to deal with the endogenous selection. For illustration, suppose that $D_t = \mathbf{1}(X_t\zeta + Z_t\pi \geq \nu_t)$, where $\nu_t \sim N(0,1)$, $(X_t, Z_t) \perp \nu_t$ and $d_x = d_z = 1$. Then, for given $x \in \mathit{Supp}(X_t)$, $\Pr(D_t = 1|X_t = x, Z_t) = \Phi(x\zeta + Z_t\pi)$ and $\Pr(D_s = 1|X_s = \bar{x}(u), Z_s) = \Phi(\bar{x}(u)\zeta + Z_s\pi)$, where Φ is the standard normal distribution function. In this case, condition (iii) in Assumption 2 is satisfied if $\pi \neq 0$ and variation in either Z_t or Z_s is large enough. The former condition $\pi \neq 0$ corresponds to the standard relevance condition for instrumental variables, and such relevance conditions are usually required for nonparametric identification with endogeneity. The latter condition which is about variation in Z_t (or Z_s) is similar to the large support condition in Imbens and Newey (2009). Similar assumptions to condition (iii) can be found in, for example, Altonji and Matzkin (2005) and Vytlacil and Yildiz (2007).

An informal description of the identification strategy in this paper is as follows: Fixing $X_t = x$, the information in time period s, together with the normalization, is used to derive an expression for r at $(X_t', X_s', \mathbf{X}_{-t,s}')' = (x', \bar{x}(u)', \mathbf{X}_{-t,s}')'$ and $h_s(\bar{x}(u), Z_s)$. Then, one can utilize the variation in $\mathbf{X}_{-t,s}$, Z_t, and Z_s to find values $\mathbf{x}_0$ and p such that $r(\mathbf{x}_0, p) = \gamma \in \mathit{Supp}(r(\mathbf{X}, h(X_t, Z_t)))$. Taking average over such values yields identification of $m(x, \gamma; u)$ over $\mathit{Supp}(X_t, r(\mathbf{X}, h_t(X_t, Z_t)))$ for each $t = 1, 2, ..., T$. Then, one can identify $r(\mathbf{x}, p)$ for all $(\mathbf{x}, p) \in \mathit{Supp}(\mathbf{X}, h_t(X_t, Z_t))$ by taking average of the inverse map of m conditional on $\mathbf{X} = \mathbf{x}$ and $h_t(X_t, Z_t) = p$. The following theorem demonstrates that Assumption 2 are sufficient for identification of $m(\cdot, \cdot; u)$ and $r(\cdot, \cdot; u)$ in (6).

Theorem 3.1. *Let $u \in \mathcal{U}$ be given. Suppose that Assumptions 1 and 2 hold. Then, for each $t = 1, 2, ..., T$, $m(\cdot, \cdot; u)$ is over $\mathit{Supp}(X_t, r(\mathbf{X}, h_t(X_t, Z_t)))$. Furthermore, $r(\cdot, \cdot; u)$ is identified over the set $\cup_t^T \mathit{Supp}(\mathbf{X}, h_t(X_t, Z_t))$.*

Proof. I drop u in functions m and r for simplicity of notation. Let $\mathbf{X}_0 = (X_t = x, X_s = \bar{x}, \mathbf{X}_{-t,s})$. Note that $h(X_t, Z_t)$ is directly identified from the data. Then, one can show that under Assumption 2,

$$
\begin{aligned}
Q_{Y_t|\mathbf{X}, Z_t, D_t=1}(u|\mathbf{X}_0, Z_t) &= m(x, r(\mathbf{X}_0, H_{t0})), \\
Q_{Y_s|\mathbf{X}, Z_s, D_s=1}(u|\mathbf{X}_0, Z_s) &= r(\mathbf{X}_0, H_{s0}),
\end{aligned}
$$

where $H_{t0} \equiv \Pr(D_t = 1|X_t = x, Z_t)$ and $H_{s0} \equiv \Pr(D_s = 1|X_s = \bar{x}, Z_s)$. Let $\gamma \in \mathbb{R}$ be given, then it is straightforward to see that

$$
m(x, \gamma) = \mathbb{E}[Q_{Y_t|\mathbf{X}, Z_t, D_t=1}(u|\mathbf{X}_0, Z_t)|Q_{Y_s|\mathbf{X}, Z_s, D_s=1}(u|\mathbf{X}_0, Z_s) = \gamma, H_{t0} = H_{s0}]. \tag{7}
$$

The conditioning event in (7) has a positive measure by conditions (ii) and (iii) in Assumption 2, and therefore $m(x,\gamma)$ is identified.

Since $m(\cdot,\cdot;u)$ is assumed to be strictly monotone in its second argument, there exists the inverse mapping with respect to the second argument. From (6), one obtains that

$$r(\mathbf{X}, P_t) = m^{-1}(X_t, Q_{Y_t|\mathbf{X},Z_t,D_t=1}(u|\mathbf{X}, Z_t)),$$

where $m^{-1}(x,y)$ is the inverse mapping of $m(x,e)$ with respect to e that is identified.[12] For any $(\mathbf{x}, p) \in Supp(\mathbf{X}, P_t)$, one obtains that

$$r(\mathbf{x}, p) = \mathbb{E}[m^{-1}(x, Q_{Y_t|X,Z,D_t=1}(u|\mathbf{X}, \mathbf{Z}))|\mathbf{X} = \mathbf{x}, h_t(x, Z_t) = p],$$

and this establishes identification of $r(\mathbf{x}, p)$ over $Supp(\mathbf{X}, h_t(X_t, Z_t))$ for each $t = 1, 2, ..., T$. ■

The LQTE and QTE are objects that may be important and relevant to policy evaluation. These objects are functionals of the structural functions m, r, and h, and thus they are identified once those structural functions are identified. The following corollary demonstrates that the LQTE and QTE are identified under the same set of conditions for identification of the structural functions.

Corollary 3.2. *Suppose that the conditions in Theorem 3.1 hold. Let $t \in \{1, 2, ..., T\}$ and $x_0, x_1 \in Supp(X_t)$ be given. Then, for any $\gamma \in Supp(r(\mathbf{X}, h_t(X_t, Z_t); u))$, $LQTE_t(u, x_0, x_1, \gamma)$ is identified. In addition, $QTE(u, x_0, x_1)$ is also identified.*

Proof. Recall that $m(x, \gamma; u) = q_t^{local}(u, x, \gamma)$. Since m and r are identified by Theorem 3.1, one obtains that

$$\begin{aligned} LQTE_t(u, x_0, x_1, \gamma) &\equiv q_t^{local}(u, x_1, \gamma) - q_t^{local}(u, x_0, \gamma) \\ &= \mathbb{E}[m(x_1, r(\mathbf{X}, h(X_t, Z_t); u); u) - m(x_0, r(\mathbf{X}, h(X_t, Z_t); u); u)|r(\mathbf{X}, h_t(X_t, Z_t); u) = \gamma]. \end{aligned}$$

Note that the conditioning event is of a positive probability because $\gamma \in Supp(r(\mathbf{X}, h_t(X_t, Z_t); u))$, and thus $LQTE_t(u, x_0, x_1, \gamma)$ is identified. Similarly, it follows from the definition of $QTE_t(u, x_0, x_1)$ that

$$QTE_t(u, x_0, x_1) = \mathbb{E}[m(x_1, r(\mathbf{X}, h_t(X_t, Z_t); u); u) - m(x_0, r(\mathbf{X}, h_t(X_t, Z_t); u); u)],$$

where the expectation is taken over $Supp(r(\mathbf{X}, h_t(X_t, Z_t); u))$. Therefore, $QTE_t(u, x_0, x_1)$ is also identified. ■

The CRE approach was also adopted by Bester and Hansen (2009) and Arellano and Bonhomme (2016), and they require that T to be greater than or equal to 3 for identification. The identification strategy in this paper, however, is different from theirs. Specifically, Bester and Hansen (2009) focus on the marginal effects of continuous covariates without completely specifying the data generating process. They use a derivative argument for identification of the marginal effects. In contrast, I focus on identification of the structural functions with specifying the data generating process (equation (1)), and the identification strategy in this paper is to use variation in excluded variables. The marginal effects in this paper are also identified as a by-product (Corollary 3.2).

[12] $y = m(x, e)$ if any only if $e = m^{-1}(x, y)$.

Arellano and Bonhomme (2016) consider nonparametric identification of structural functions, but the identification strategy in Theorem 3.1 is different from that of Arellano and Bonhomme (2016). Specifically, Arellano and Bonhomme (2016) use a high-level assumption, called an injectivity condition, and this condition resembles completeness conditions that are commonly used in the literature on nonparametric identification (e.g. Newey and Powell (2003) and Blundell et al. (2007)). The injectivity condition, however, is relatively difficult to interpret and verify in practice. More importantly, estimation and inference may suffer from an ill-posed inverse problem which leads to a slower convergence rate. On the other hand, the identification strategy in this paper does not rely on completeness conditions, and hence it is not subject to an ill-posed inverse problem.

The identification strategy in Theorem 3.1 does not require to specify the distribution of the unobserved error term. In contrast, Arellano and Bonhomme (2017) and Chernozhukov et al. (2018) consider some semiparametric specification of the joint distribution of Y^* and D. Furthermore, both papers focus on quantile regression models for cross-sectional data, whereas this paper considers models for panel data.

The implication of model (equation (6)) is similar to that of Lewbel and Linton (2007) or Escanciano et al. (2016), and hence the identification strategy of this paper shares some common with their strategies. However, the models of the papers are different from that of this paper. Specifically, the model of Lewbel and Linton (2007) differs from (6) in that they assume that X_t is excluded from $\mathbf{X}$ and that the selection probability does not depend on X_t. Therefore, their identification strategy cannot be directly applied to identify m and r in (6). Escanciano et al. (2016) study a class of models where there are two index functions and m relates these two index functions.[13] The focus of Escanciano et al. (2016) is on identification and estimation of the finite-dimensional parameter in one of the index functions, but this paper studies nonparametric identification and estimation of (6) without specifying an index function for X_t, which allows for more flexibility of the model.

While the linear correlated random coefficients models allow for multi-dimensional error terms, the identification comes at cost of a larger (but fixed) number of time periods (e.g. Arellano and Bonhomme (2012); Graham and Powell (2012); Graham et al. (2018); Laage (2019)). In contrast, this paper imposes a scalar error term, but identification requires T be greater than equal to 3 with some support condition. The requirement for T is much weaker than that in the correlated random coefficients models where T should be greater than or equal to the number of covariates. In addition, the model in this paper is completely nonparametric, whereas most of correlated random coefficients models are parametric or semiparametric.

3.2 Extensions

In this section, I discuss some extensions of the panel quantile models with sample selection in Section 2. I consider (i) endogeneity of X_t, (ii) censoring, and (iii) dynamic panel data models, and they are useful and relevant to many empirical situations. I show that model (5) can be easily extended to incorporate these issues.

[13] One of the index functions is a linear-index function, and the other one is a known function from data.

3.2.1 Endogenous Regressors

Endogeneity issues are prevalent in many empirical questions. The CRE specification effectively captures "time-invariant" endogeneity, but some regressors may exhibit "time-varying" endogeneity.[14] The model implication in (6) is closely related to the control function approach to deal with sample selection bias, and it can be extended to allow for endogeneity of X_t. To make it concrete, suppose that $X_t = (X_t^{e'}, \tilde{X}_t')'$, where X_t^e is a vector of endogenous regressors and $\tilde{X}_t$ is a vector of exogenous regressors. For brevity of the model, I assume that $X_t^e \in \mathbb{R}$, but it can be easily extended to the case where X_t^e is a vector. Assume that and $Z_t = (Z_{1t}', Z_{2t}')'$, and consider the following class of models:

$$\begin{aligned} Y_t^* &= m(X_t, U_t), : \\ Y_t &= D_t Y_t^*, \\ X_t^e &= q(Z_{2t}, V_t), \\ Q_{U_t|\mathbf{X},\mathbf{Z},D_t=1}(u|\mathbf{X}, Z_t, D_t = 1) &= r^e(\tilde{\mathbf{X}}, V_t, h_t(X_t, Z_{1t}); u), \end{aligned} \tag{8}$$

where $V_t \in \mathbb{R}$ is unobserved and independent of Z_{2t}, and $q(z_2, v)$ is a non-trivial function of z_2 and strictly increasing in v for all z_2. Without loss of generality, V_t is assumed to be uniformly distributed on the unit interval, conditional on Z_{2t}. Model (8) is closely related to the sample selection model with endogeneity that is studied by Das et al. (2003). The conditional quantile restriction on U_t in model (8) implies that, conditional on V_t, X_t^e is no longer endogenous, and hence the following model implication is obtained:

$$Q_{Y_t|\mathbf{X},\mathbf{Z},D_t=1}(u|\mathbf{X}, \mathbf{Z}, D_t = 1) = m(X_t, r^e(\tilde{\mathbf{X}}, V_t, h_t(X_t, Z_{1t}); u); u). \tag{9}$$

This extends the control function approach to handle the sample selection bias that is presented in (6) to a more general case where some regressors are endogenous. The variable V_t plays the role of a control function to handle endogeneity of X_t^e, and needs to be estimated in the first-stage. Since the model implication (6) suggests that the selection probability $h_t(X_t, Z_t)$ plays the role of control function to correct selection bias, the roles of $h_t(X_t, Z_t)$ and V_t are almost the same. Similar approaches for cross-sectional data models are considered by, for example, Newey et al. (1999), Lee (2007), Imbens and Newey (2009), and Chernozhukov et al. (2015). For panel data models, Semykina and Wooldridge (2010) develop a class of models that is similar to (9), but their focus is on the conditional mean function with additively separable error terms.

3.2.2 Censoring

Censoring is an issue that empirical researchers frequently face. I consider the following censoring rule with sample selection:

$$\begin{aligned} Y_t^* &= m(X_t, U_t), \\ Y_t &= D_t \cdot \max(Y_t^*, C_t), \end{aligned}$$

[14] As mentioned earlier, the CRE specification is closely related to the dependence between the regressors and time-invariant unobserved heterogeneity, which is commonly assumed in fixed effects models. On the other hand, I use the term time-varying endogeneity to allow for dependence between the regressors and time-varying components of the error term U_t. See also Laage (2019).

where C_t's are fixed constants. Since the quantile operator is preserved under monotone transformations, quantile regression models can easily incorporate censored data in a similar fashion of Powell (1986). Specifically, one can show that

$$Q_{Y_t|\mathbf{X},Z_t,D_t=1}(u|\mathbf{X},Z_t,D_t=1) = \max(Q_{Y_t^*|\mathbf{X},Z_t,D_t=1}(u|\mathbf{X},Z_t,D_t=1),C_t),$$

and the conditional quantile function of Y_t^* is the same as (6). It is worth noting that one can simultaneously incorporate endogeneity and censoring, as in Chernozhukov et al. (2015). This result can be further extended to the Tobit type-3 model considered by Fernández-Val et al. (2019). The model of Fernández-Val et al. (2019) is different from that of this paper in that the selection rule in their model is not binary and that the error term in the outcome equation can be multi-dimensional, and their identification strategy relies on the control function approach of Imbens and Newey (2009). Nevertheless, the model in this paper can incorporate such a class of selection rules with the control function approach as shown earlier.

4 Semiparametric Models

While fully nonparametric models are attractive as they are robust to model misspecification, one important and practical issue is that it is difficult to estimate parameters in them when the dimension of covariate is large. Although the CRE approach allows us to consider flexible and general models, the number of covariates involved in estimating parameters can be very large and thus the fully nonparametric model presented in the previous section may not be practically useful. To address such issues, I propose some semiparametric models and study identification of parameters in those models.

4.1 Index Models

One can impose an index structure on the structural function r , and this is originally motivated by the original CRE approach of Mundlak (1978) and Chamberlain (1980). Specifically, I impose the following assumption.

Assumption 1′. *Let $u \in \mathcal{U}$ be given. For all $t \in \{1,2,...,T\}$,*

$$Q_{U_t|\mathbf{X},Z_t,D_t=1}(u|\mathbf{X},Z_t,D_t=1) = r(\mathbf{X}^{'}\delta(u),P_t), \tag{10}$$

where $\delta = (\delta_1^{'},\delta_2^{'},...,\delta_T^{'})^{'}$ and $\delta_t \in \mathbb{R}^{d_x}$ for all t.

The index structure is consistent with the CRE specification of Chamberlain (1980) and reduces the dimension of the structural function r, while allowing for nonseparability between the index and P_t. This semiparametric specification requires additional assumptions for identification of the index coefficient vector δ, and these assumptions depend on X_t being continuous or discrete. To make the role of each type of regressor, I assume that $X_t = (X_t^{c'},X_t^{d'})^{'}$, where $X_t^c \in \mathbb{R}^{d_{xc}}$ is a vector of continuous regressors and $X_t^d \in \mathbb{R}^{d_{xd}}$ is a vector of discrete regressors. Consequently, I partition the coefficient δ into two parts. That is, I assume that $\delta_t = (\delta_t^{c'},\delta_t^{d'})^{'}$, where $\delta_t^c = (\delta_{1t}^c,\delta_{2t}^c,...,\delta_{d_{xc}t}^c)^{'}$ and $\delta_t^d = (\delta_{1t}^d,\delta_{2t}^d,...,\delta_{d_{xd}t}^d)^{'}$, and impose the following assumption.

Assumption 2$'$. *Let $T \geq 3$. For any $u \in \mathcal{U} \subseteq (0,1)$, the following conditions hold:*

(i) For each $t = 1,2,...,T$, there exists a known value $\bar{x}(u) \in Supp(X_t) \subseteq \mathbb{R}^{d_x}$ such that $m(\bar{x}(u), \gamma; u) = \gamma$;

(ii) Let $x \in Supp(X_t)$ and $\gamma \in Supp(r(\mathbf{X}^{'}\delta, h(X_t, Z_t)))$ be given. For any $t, s \in \{1,2,...,T\}$ with $t \neq s$, there exists a non-empty subset $\tilde{\mathcal{X}}^S_{-t,s}(x, \bar{x}(u))$ of $Supp(\mathbf{X}^{'}_{-t,s}\delta_{-t,s}|(X_t, X_s) = (x, \bar{x}(u)))$ and such that, for any $\mathbf{x}_{-t,s} \in \tilde{\mathcal{X}}^S_{-t,s}(x, \bar{x}(u))$, $r(\mathbf{x}_0^{'}\delta, p) = \gamma$ for some $p \in Supp(h(x, Z_t)|\mathbf{X} = \mathbf{x}_0)$ and $\Pr(\mathbf{X}_{-t,s} \in \tilde{\mathcal{X}}_{-t,s}(x, \bar{x}(u))) > 0$, where $\mathbf{x}_0 = (X_t = x, X_s = \bar{x}(u)), \mathbf{X}_{-t,s} = \mathbf{x}_{-t,s}$);

(iii) For any $t = 1,2,...,T$ and for any $x \in Supp(X_t)$ and $z \in Supp(Z_t) \subseteq \mathbb{R}^{d_z}$, there exists a non-empty set $\mathcal{Z}_s(z) \subseteq Supp(Z_s)$ for some $s \in \{1,2,...,T\}$ such that, for any $\tilde{z} \in \mathcal{Z}_s(z)$, $h_t(x,z) = h_s(\bar{x}(u), \tilde{z})$ and $\Pr(Z_s \in \mathcal{Z}_s(z)) > 0$.

(iv) $\delta^c_{1t} = 1$ for all t.

(v) $m(\cdot,\cdot)$ is differentiable with respect to the second argument, and $r(\cdot,\cdot)$ is differentiable with respect to the first argument.

(vi) $r(\cdot,\cdot)$ is invertible with respect to the first argument.

Conditions (i), (ii), and (iii) in Assumption 2$'$ are almost the same as conditions (i), (ii), and (iii) in Assumption 2, respectively. Condition (iv) – (vi) in Assumption 2 are additionally imposed to identify the finite-dimensional parameter δ. Condition (iv) is a normalization, which is very standard in the literature (e.g. Escanciano et al. (2016)). It requires that there exist at least on continuous regressor whose the coefficient is nonzero. Condition (v) imposes some smoothness on m and r, and this condition allows one to identify δ^c_t's. It is worth pointing out that if X_t consists only of continuous regressors, the coefficients δ^c_t's are be identified without condition (vi) in Assumption 2$'$. Condition (vi) can be implied by strict monotonicity of $r(\cdot,\cdot)$ with respect to its first argument, and Escanciano et al. (2016) also impose a similar condition to identify the coefficients on discrete regressors. To motivate this assumption, consider the linear panel quantile model in Example 2.2. It can easily be shown that $Q_{Y_t|\mathbf{X},Z_t,D_t=1}(u|\mathbf{X}, Z_t, D_t = 1) = X_t\beta + \mathbf{X}\delta + g(P_t)$ for some unknown function $g(\cdot)$ and that $r(a,p) = a + g(p)$. In this case, the structural function r is strictly increasing in its first argument, and therefore condition (vi) is satisfied. In the wage equation example, the $\mathbf{X}^{'}\delta$ can be considered as the ability of individual, and it is natural to assume that the structural function r is monotonically increasing in $\mathbf{X}^{'}\delta$. A similar assumption is made by Evdokimov (2010), without considering the CRE approach.

The following theorem demonstrates that the parameters of the semiparametric model in (10) are identified under Assumptions 1$'$ and 2$'$.

Theorem 4.1. *Let $u \in \mathcal{U}$ be given and Assumption 1$'$ hold. Suppose that conditions (i) – (v) in Assumption 2$'$ are satisfied. Then, for each $t = 1,2,...,T$, $m(\cdot,\cdot;u)$ and $r(\cdot,\cdot;u)$ are identified over $Supp(X_t, r(\mathbf{X}^{'}\delta, h_t(X_t, Z_t)))$ and the set $\cup_t^T Supp(\mathbf{X}^{'}\delta, h_t(X_t, Z_t))$, respectively, and δ^c_t's are also identified. If condition (vi) in Assumption 2$'$ additionally holds, then δ^d_t's are also identified.*

Proof. Let $t \in \{1,2,...,T\}$ be given. Under conditions (i) through (iii) in Assumption 2$'$, the structural function m is identified over it support and one can show that

$$r(\mathbf{X}^{'}\delta, P_t) = m^{-1}(X_t, Q_{Y_t|\mathbf{X},Z_t,D_t=1}(u|\mathbf{X}, Z_t)) \tag{11}$$

by using the same argument of the proof of Theorem 3.1. In addition, the structural function r is also identified from (11) over its support in a similar way to the proof of Theorem 3.1.

I first identify the coefficients on the continuous regressors δ_t^c's. Choose $s \in \{1, ..., t-1, t+1, ..., T\}$. Taking derivative with respect to X_{1s}^c, one obtains that

$$r_1(\mathbf{X}^{'}\delta, P_t) = m_2^{-1}(X_t, Q_{Y_t|\mathbf{X},Z_t,D_t=1}(u|\mathbf{X}, Z_t)) \frac{\partial Q_{Y_t|\mathbf{X},Z_t,D_t=1}(u|\mathbf{X}, Z_t)}{\partial X_{1s}^c}. \tag{12}$$

Pick any $k \in \{2, 3, ..., d_x\}$. Taking derivative with respect to X_{ks}^c yields that

$$r_1(\mathbf{X}^{'}\delta, P_t)\delta_{ks}^c = m_2^{-1}(X_t, Q_{Y_t|\mathbf{X},Z_t,D_t=1}(u|\mathbf{X}, Z_t)) \frac{\partial Q_{Y_t|\mathbf{X},Z_t,D_t=1}(u|\mathbf{X}, Z_t)}{\partial X_{ks}^c}. \tag{13}$$

As a result, one can see that δ_{ks}^c is identified by the ratio between (12) and (13), and therefore one can identify $\delta_t^{'c}$ for all $t \in \{1, 2, ..., T\}$. Note that to identify δ_t^c, one can consider the model restriction in (11) for some different time period s and use the same argument.

By the invertibility condition (condition (vi) in Assumption 2′), one obtains that

$$\mathbf{X}^{'}\delta = r^{-1}(m^{-1}(X_t, Q_{Y_t|\mathbf{X},Z_t,D_t=1}(u|\mathbf{X}, Z_t)), P_t).$$

Using the variation in the discrete regressor X_t^d, one can identify δ_t^d's. ■

One can also consider an index structure for structural function m, and this is in particular useful when the dimension of X_t is large. Specifically, if it is assumed that $m(X_t, \gamma) = m(X_t^{'}\beta, \gamma)$, then one can use a similar argument in the proof of Theorem 4.1 under similar conditions for m to those for r. These conditions include (i) the differentiability and invertibility of m with respect to its first argument and (ii) a normalization condition for β.

4.2 Additively Separable Models

Additively separable models are very popular in empirical studies as they are very tractable. In particular, one can use a location-scale model for quantile regression to allow for general and flexible specifications even with additive separability. The following assumption imposes additive separability between X_t and r as well as a parametric specification for structural function m.

Assumption 1″. *Let $u \in \mathcal{U}$ be given. For all $t \in \{1, 2, ..., T\}$,*

$$Q_{Y_t|\mathbf{X},Z_t,D_t=1}(u|\mathbf{X}, Z_t, D_t = 1) = X_t^{'}\beta(u) + \mathbf{X}^{'}\delta(u) + g(P_t; u). \tag{14}$$

Equation (14) is corresponding to the model considered in Example 2.2, and it is a special case of the nonparametric model in (5). The quantile restriction in Assumption 1″ is similar to that of Buchinsky (1998) and the conditional mean restriction of Das et al. (2003).

While it is evident that Assumption 1″ restricts the type of heterogeneity that can be allowed in the model, it provides substantial identifying power. If Assumption 1″ holds and some rank condition is satisfied, one can weaken the condition on the number of time periods that is needed for identification. The following assumption provides a set of identification conditions for model (14).

Assumption 2″. *Let $T \geq 2$. For any $u \in \mathcal{U} \subseteq (0,1)$, the following conditions hold:*

(i) For any $t \neq s$, $\Pr\left(rank\left((X_t - X_s)\cdot(X_t - X_s)^{'}\right) = d_x\right) = 1$;

(ii) For any $t \neq s$ and $(x_t^{'}, x_s^{'})^{'} \in Supp(X_t, X_s)$, there exists $\mathcal{Z}((x_t^{'}, x_s^{'})^{'}) \subseteq Supp(Z_t, Z_s)$ such that $\Pr(\mathcal{Z}((x_t^{'}, x_s^{'})^{'})) > 0$ and $h_t(x_t, z_t) = h_s(x_s, z_s)$ for all $(z_t^{'}, z_s^{'})^{'} \in \mathcal{Z}((x_t^{'}, x_s^{'})^{'})$;

Condition (i) in Assumption 2″ is a rank condition. It requires that the time-varying covariate X_t have sufficient variation across time periods, and this condition is standard in the literature on panel data models. It is worth noting that this condition rules out the case where X_t contains some time-invariant regressors, such as a constant regressor. Condition (ii) is similar to condition (iii) in Assumption 2, which requires the excluded variable Z_t to have sufficient variation. If Z_t has a large support and the selection is determined by a threshold crossing equation model, then this condition is likely to be met. Condition (iii) is likely to be stronger than the relevance condition of the excluded variable proposed by Das et al. (2003). Their identification condition, however, requires a location normalization for g as it is identified up to an additive constant. On the other hand, condition (iii) allows to identify g without imposing such a normalization.

The next theorem establishes the identification of β, δ, and $g(\cdot)$ under these conditions.

Theorem 4.2. *Let $u \in \mathcal{U}$ be given and Assumptions 1″ and 2″ hold. Then, $\beta(u)$ and $\delta(u)$ are identified. Moreover, $g(\cdot; u)$ is identified over $\cup_{t=1}^{T} Supp(P_t)$.*

Proof. For simplicity of notation, I assume that X_t is a continuous random variable.

Identification of δ is from the following derivative:

$$\delta_s = \frac{\partial Q_{Y_t|\mathbf{X}, Z_t, D_t=1}(u|\mathbf{X}, Z_t, D_t = 1)}{\partial X_s},$$

where $t \neq s$. Since $\delta \in \mathbb{R}^T$ and there are $T \times (T-1)$ equations, one can identify δ.

For identification of β, pick any $t, s \in \{1, 2, ..., T\}$ such that $t \neq s$. Then,

$$Q_{Y_t|\mathbf{X}, Z_t, D_t=1}(u|\mathbf{X}, Z_t, D_t = 1) - Q_{Y_s|\mathbf{X}, Z_s, D_s=1}(u|\mathbf{X}, Z_s, D_s = 1) = (X_t - X_s)^{'}\beta + g(P_t) - g(P_s).$$

Taking conditional expectation on $(Z_t^{'}, Z_s^{'})^{'} \in \mathcal{Z}((X_t^{'}, X_s^{'})^{'})$, one obtains that

$$\begin{aligned}&\mathbb{E}[Q_{Y_t|\mathbf{X}, Z_t, D_t=1}(u|\mathbf{X}, Z_t, D_t = 1) - Q_{Y_s|\mathbf{X}, Z_s, D_s=1}(u|\mathbf{X}, Z_s, D_s = 1)|X_t, X_s, (Z_t^{'}, Z_s^{'})^{'} \in \mathcal{Z}((X_t^{'}, X_s^{'})^{'})]\\ &= (X_t - X_s)^{'}\beta\end{aligned}$$

and this condition expectation is well-defined by condition (ii) in Assumption 2″. Therefore, by condition (i) in Assumption 2″, one can show that

$$\begin{aligned}\beta = &\left((X_t - X_s)\cdot(X_t - X_s)^{'}\right)^{-1}.(X_t - X_s)\\ &\times \mathbb{E}[Q_{Y_t|\mathbf{X}, Z_t, D_t=1}(u|\mathbf{X}, Z_t, D_t = 1) - Q_{Y_s|\mathbf{X}, Z_s, D_s=1}(u|\mathbf{X}, Z_s, D_s = 1)|X_t, X_s, (Z_t^{'}, Z_s^{'})^{'} \in \mathcal{Z}((X_t^{'}, X_s^{'})^{'})].\end{aligned}$$

Since β and δ are identified, $g(\cdot)$ is also identified over $\cup_{t=1}^{T} Supp(P_t)$ from equation (14). ∎

5 Estimation

Let m_0, r_0, and $h_{t,0}$ be the true parameter values for m, r, and h_t, respectively. For a generic random vector A, $vec(A)$ denotes the vectorization of A. Let $\{\mathbf{W}_i \equiv (\mathbf{Y}_i, \mathbf{X}_i^{'}, \mathbf{Z}_i^{'}, \mathbf{D}_i^{'})^{'} : i = 1, 2, ...N\}$ be the data, where $\mathbf{X}_i \equiv (vec(X_{i1})^{'}, ..., vec(X_{iT})^{'})^{'}$ and $\mathbf{Z}_i \equiv (vec(Z_{i1})^{'}, ..., vec(Z_{iT})^{'})^{'}$. The model implication (6) suggests estimation strategies based on the standard quantile regression. That is, assuming that the selection probability $P_{it} = h_{t,0}(X_{it}, Z_{it})$ is observed, one can estimate functions m and r by solving the following minimization problem:

$$\max_{(m,r)\in\mathcal{M}\times\mathcal{R}} \frac{1}{nT} \sum_{i:D_{it}=1}^{n} \sum_{t}^{T} \rho_u(Y_{it} - m(X_{it}, r(\mathbf{X}_i, P_{it}; u); u)), \tag{15}$$

where $\rho_u(x) = x(\mathbf{1}(x < 0) - u)$, n is the number of individuals who are selected, and $\mathcal{M}$ and $\mathcal{R}$ are classes of admissible functions for $m(\cdot, \cdot; u)$ and $r(\cdot, \cdot; u)$, respectively. Note that the observations that are used for estimation are those who are selected. I assume that $n/N \to n^* \in (0, 1)$. Under this assumption, the asymptotic analysis can be based on n.

Since the selection probability P_{it} is not directly observed from the data, the maximization problem in (15) is infeasible. I propose a two-step estimation procedure. Recall that the selection probability is a function of $(X_{it}^{'}, Z_{it}^{'})^{'}$, $h_{t,0}(X_{it}, Z_{it})$, and I assume that $h_{t,0} \in \mathcal{H}$ for some space of functions for all $t = 1, 2, ..., T$. This selection probability for each t is estimated in the first step, and the estimate is denoted by $\hat{P}_{it} = \hat{h}_{t,n}(X_{it}, Z_{it})$, where $\hat{h}_{t,n}(\cdot, \cdot)$ is an estimator of $h_{t,0}(\cdot, \cdot)$. There are several methods to estimate the selection probability, including some parametric estimators, the semiparametric estimators of Klein and Spady (1993), and nonparametric estimators such as the series logit estimators in Hirano et al. (2003). Then, one solves (15) with replacing P_{it} with $\hat{P}_{it}$ to estimate m_0 and r_0 in the second step.

I use sieve methods to estimate functions m and r. Sieve estimation is useful to impose additional structures of the model, such as additive separability, and easy to implement in practice. Specifically, let $\mathcal{M}_n$ and $\mathcal{R}_n$ be appropriate sieve spaces for $\mathcal{M}$ and $\mathcal{R}$, respectively. Note that, since one can consider the sieve methods to estimate $h_{t,0}$, I allow the parameter space $\mathcal{H}$ to depend on the sample size n, and denote it by $\mathcal{H}_n$.[15] Then, a feasible sieve estimator for $(m(\cdot, \cdot; u), r(\cdot, \cdot; u))$, denoted by $(\hat{m}_n(\cdot, \cdot; u), \hat{r}_n(\cdot, \cdot; u))$, is defined as follows:

$$(\hat{m}_n(\cdot, \cdot; u), \hat{r}_n(\cdot, \cdot; u)) \equiv \arg \max_{(m,r)\in\mathcal{M}_n\times\mathcal{R}_n} \frac{1}{nT} \sum_{i:D_{it}=1}^{n} \sum_{t}^{T} \rho_u(Y_{it} - m(X_{it}, r(\mathbf{X}_i, \hat{P}_{it}; u); u)). \tag{16}$$

The choice of sieve spaces depends on the class of functions and support of unknown function. I introduce one of the most popular classes of functions, which is called the Hölder class. Let $f : \mathbb{D} \to \mathbb{R}$ where $\mathbb{D} \subseteq \mathbb{R}^{d_x}$ for some integer $d_x \geq 1$. Let $\omega = (\omega_1, ..., \omega_{d_x})$ be a d_x-tuple of nonnegative integers, and define the differential operator as $\nabla^\omega f \equiv \frac{\partial^{|\omega|}}{\partial x_1^{\omega_1} \partial x_2^{\omega_2} \cdots \partial x_{d_x}^{\omega_{d_x}}} f(x)$, where $x = (x_1, x_2, ..., x_{d_x}) \in \mathbb{D}$ and $|\omega| \equiv \sum_{i=1}^{d_x} \omega_i$. Let $[p]$ be the integer part of $p \in \mathbb{R}_+$, then a function $f : \mathcal{X} \to \mathbb{R}$ is called p-smooth if it is $[p]$ times continuously differentiable on $\mathcal{X}$ and for all ω such that $|\omega| = [p]$ and for some $\nu \in (0, 1]$ and constant $c > 0$, $|\nabla^\omega f(x) - \nabla^\omega f(y)| \leq c \cdot ||x - y||_E^\nu$ for

[15] $\mathcal{H}_n$ may not vary across the sample size in some cases. In particular, $\mathcal{H}_n = \mathcal{H}$ for all $n \geq 1$ for parametric estimation.

all $x, y \in \mathcal{X}$, where $||\cdot||_E$ is the Euclidean norm. Let $\mathcal{C}^{[p]}(\mathcal{X})$ denote the space of all $[p]$ times continuously differentiable real-valued functions on $\mathcal{X}$. A Hölder ball with smoothness p is defined as follows:

$$\Lambda_C^p(\mathcal{X}) \equiv \{f \in \mathcal{C}^{[p]}(\mathcal{X}) : \sup_{|\omega| \leq [p]} \sup_{x \in \mathcal{X}} |\nabla^\omega f(x)| \leq C, \sup_{|\omega|=[p]} \sup_{x,y \in \mathcal{X}, x \neq y} \frac{|\nabla^\omega f(x) - \nabla^\omega f(y)|}{||x-y||_E^\nu} \leq C\},$$

where C is a positive finite constant. I assume that $\mathcal{M}$ and $\mathcal{R}$ are Hölder balls with possibly different degrees of smoothness.

When an unknown function is in a Hölder ball and its support is the unit interval, one can use polynomial, triometric polynomial, or spline sieve spaces. If the support is unbounded, then Hermite polynomial sive spaces can be used. For the detailed discussion on the choice of sieve spaces, one can refer to Chen (2007).

It is worth noting that none of the structural functions depend on endogenous regressors, and therefore it does not suffer from an ill-posed inverse problem. In addition, it is also possible to incorporate penalty as in Chen and Pouzo (2012), but estimation with penalization is not considered in this paper.[16]

6 Asymptotic Theory for Nonparametric Two-Step Estimators

In this section, I provide the asymptotic theory for the sieve estimator of (m_0, r_0). The objective function of the maximization problem (16) contains (possibly nonparametrically) generated regressors. I adapt the approach proposed by Hahn et al. (2018a) who consider nonparametric two-step sieve estimation. All mathematical proofs for theorems in this section are presented in the appendix.

I assume that $h_{1,0} = h_{2,0} = ... = h_{T,0} \equiv h_0$ for simplicity in establishing the asymptotic theory, but the asymptotic theory developed in this paper can allow the selection probability function to vary across time. Let $\theta = (m, r)'$ and $\alpha = (m, r, h)'$. The parameter spaces for θ, h, and α are denoted by Θ, $\mathcal{H}$, and $\mathcal{A}$, respectively. Let $l_u(\mathbf{W}_i, \theta, h) \equiv \sum_t D_{it} \rho_u(\mathbf{Y}_{it} - m(X_{it}, r(\mathbf{X}_i, P_{it}; u); u))$. Then, $L_{u,n}(\mathbf{W}, \alpha) = \frac{1}{n} \sum_i l_u(\mathbf{W}_i, \alpha)$ and $L_{u,0}(\alpha) = \mathbb{E}[L_{u,n}(\mathbf{W}, \alpha)]$. I assume that for each t, the conditional distribution function of Y_t^* on $\mathbf{X}$ and Z_t is absolutely continuous with respect to the Lebesgue measure, so that it admits its density function $f_{Y_t^*|\mathbf{X}, Z_t}$.

I define several norms that are used in this paper. Let $||\cdot||_\infty$ and $||\cdot||_2$ denote the supremum-norm and L_2-norm on a function space, respectively. For any $\theta, \tilde{\theta} \in \Theta$, define $d_{\Theta,\infty}(\theta, \tilde{\theta}) \equiv ||m(x,\gamma) - \tilde{m}(x,\gamma)||_\infty + ||r(\mathbf{x}, p) - \tilde{r}(\mathbf{x}, p)||_\infty$ and $||\theta - \tilde{\theta}||_{\Theta,2}^2 \equiv ||m(x,\gamma) - \tilde{m}(x,\gamma)||_2^2 + ||r(\mathbf{x}, p) - \tilde{r}(\mathbf{x}, p)||_2^2$. For any $\alpha, \tilde{\alpha} \in \mathcal{A}$, let $d_{\mathcal{A},\infty}(\alpha, \tilde{\alpha}) \equiv d_{\Theta,\infty}(\theta, \tilde{\theta}) + ||h(x,z) - \tilde{h}(x,z)||_\infty$ and $||\alpha - \tilde{\alpha}||_{\mathcal{A},2}^2 \equiv ||\theta - \tilde{\theta}||_{\Theta,2}^2 + ||h - \tilde{h}||_2^2$. Similarly, I define the Euclidean norm on $\mathcal{A}$ as $||\alpha - \tilde{\alpha}||_{\mathcal{A},E}^2 \equiv |m - \tilde{m}|^2 + |r - \tilde{r}|^2 + |h - \tilde{h}|^2$.

6.1 Consistency and Convergence Rates

I first show the consistency of the sieve estimator for θ_0, $\hat{\theta}_n$, with respect to the sup-norm. To establish consistency, I consider the following assumptions.

[16] Related to penalization, Chen and Pouzo (2012) argue that estimation without penalty can be applied when the parameters of interest are some smooth functions.

Assumption 3. *(i) The data* $\{\mathbf{W}_i = (W_{i1}, ..., W_{iT})' : i = 1, 2, ...n\}$ *are i.i.d across* i*, and* $\{W_{it} : t = 1, 2, ..., T\}$ *is stationary for each* $i = 1, 2, ..., n$*; (ii) for any* $t = 1, 2, ..., T$*, the conditional distribution of* Y_t^* *on* $\mathbf{X}$ *and* Z_t *is absolutely continuous with respect to the Lebesgue measure, and its conditional density function* $f_{Y_t^*|\mathbf{X},Z_t}$ *satisfies that* $0 < \inf_{y\in\mathbb{R},(\mathbf{x},z)\in Supp(\mathbf{X},Z_t)} f_{Y_t^*|\mathbf{X},Z_t}(y|\mathbf{x}, z) < \sup_{y\in\mathbb{R},(\mathbf{x},z)\in Supp(\mathbf{X},Z_t)} f_{Y_t^*|\mathbf{X},Z_t}(y|\mathbf{x}, z) < \infty$*; (iii) for any* $t = 1, 2, ..., T$*,* $\mathbb{E}[|Y_t^*|]$ *and* $\mathbb{E}[|q_u(\mathbf{X}_i, Z_{it})|]$ *are uniformly bounded; (iv) for each* $t = 1, 2, ..., T$*, the supports of* X_{it} *and* Z_{it} *are compact.*

Assumption 4. *(i)* $m_0 \in \mathcal{M} \equiv \Lambda_{c_m}^{p_m}(\mathbb{M})$ *and* $r_0 \in \mathcal{R} \equiv \Lambda_{c_r}^{p_r}(Supp(\mathbf{X}, h_0(X_t, Z_t))$ *with* $p_m > 1$ *and* $p_r > 1$*,* $m_0(x, \gamma)$ *and* $r_0(\mathbf{x}, p)$ *are continuously differentiable with respect to* γ *and* p*, respectively, and the derivatives are uniformly bounded; (ii) there exist measurable functions* $\overline{\mathbb{A}(\mathbf{x}, z_t)}$ *and* $\underline{\mathbb{A}(\mathbf{x}, z_t)}$ *such that for any* $\alpha, \tilde{\alpha} \in \mathcal{A}$ *and for any* $\mathbf{x} = (x_t, \mathbf{x}_{-t}) \in Supp(\mathbf{X})$ *and* $z_t \in Supp(Z_t)$*,*

$$\underline{\mathbb{A}(\mathbf{x}, z_t)}||\alpha - \tilde{\alpha}||_{\mathcal{A},E}^2 \le \{m(x_t, r(\mathbf{x}, h(x_t, z_t))) - \tilde{m}(x_t, \tilde{r}(\mathbf{x}, \tilde{h}(x_t, z_t)))\}^2 \le \overline{\mathbb{A}(\mathbf{x}, z_t)}||\alpha - \tilde{\alpha}||_{\mathcal{A},E}^2$$

and $\mathbb{E}[\underline{\mathbb{A}(\mathbf{X}, Z_t)}^2], \mathbb{E}[\overline{\mathbb{A}(\mathbf{X}, Z_t)}^2] < \infty$*.*

Assumption 5. *(i)* $\mathcal{M}_n = \{m_n(x, \gamma) = \phi_m^{k_{m,n}}(x, \gamma)'\beta_{m,n} : \sup_{x,\gamma} |m_n(x, \gamma)| \le c_m\}$*,* $\mathcal{R}_n = \{r_n(\mathbf{x}, p) = \phi_r^{k_{r,n}}(\mathbf{x}, p)'\beta_{r,n} : \sup_{\mathbf{x},p} |r_n(\mathbf{x}, p)| \le c_r\}$*, where* $k_{m,n}$ *and* $k_{r,n}$ *are some positive non-decreasing integer sequences such that* $k_{m,n}, k_{r,n} \to \infty$*,* $\max(k_{m,n}, k_{r,n}) = o(n)$*; (ii) let* $\mathbb{Q}_{m,t} \equiv \mathbb{E}[\phi_m^{k_{m,n}}(x_t, r_n(\mathbf{X}, P_t)) \cdot \phi_m^{k_{m,n}}(x_t, r_n(\mathbf{X}, P_t))']$ *and* $\mathbb{Q}_{r,t} \equiv \mathbb{E}[\phi_r^{k_{r,n}}(\mathbf{X}, P_t) \cdot \phi_r^{k_{r,n}}(\mathbf{X}, P_t)']$*, then for any* t*, the eigenvalues of* $\mathbb{Q}_{m,t}$ *and* $\mathbb{Q}_{r,t}$ *are bounded above and away from zero uniformly over all* n*.*

Assumption 3 imposes conditions on the data generating process. Note that the first condition of Assumption 3 allows for serial correlation as it only requires the data be i.i.d. across the individuals. Condition (ii) in Assumption 3 is standard for quantile regression models. Condition (iii) is a mild condition on moments of the dependent variable and conditional quantile function.

Assumption 4 specifies the parameter space for the structural functions m and r. Condition (ii) is implied by some smoothness conditions on m and r. This can be considered as a variant of an assumption that is imposed for nonlinear quantile regression.[17] This condition implies that $\mathbb{E}[\{m(X_{it}, r(\mathbf{X}_i, h(X_{it}, Z_{it}))) - \tilde{m}(X_{it}, \tilde{r}(\mathbf{X}_i, \tilde{h}(X_{it}, Z_{it})))\}^2] \asymp ||\alpha - \tilde{\alpha}||_{\mathcal{A},2}^2$.

Assumption 5 defines sieve spaces for $\mathcal{M}$ and $\mathcal{R}$. The choice of sieve spaces depends on the parameter spaces and support conditions. When the parameters of interest belong to a Hölder space and the supports are compact, one can use finite-dimensional linear sieve spaces, such as polynomial, trigonometric, or B-spline sieve spaces. Assumption 5, together with Assumption 4-(i) ensures that the sieve spaces approximate the parameter spaces well. Condition (ii) of Assumption 5 is standard in the literature on sieve or series estimation (cf. Newey (1997) and Chen and Christensen (2018)). Note that the sieve spaces are linear finite-dimensional, and thus the maximization problem in (16) becomes a finite-dimensional optimization problem.

Assumptions 4 and 5 together imply that there exist $\{\beta_{m,n}^*\}_{n=1}^\infty$ and $\{\beta_{r,n}^*\}_{n=1}^\infty$ such that $\sup |m_0(x, \gamma) - \phi_m^{k_{m,n}}(x, \gamma)'\beta_{m,n}^*| = O(k_{m,n}^{-\sigma_m})$ and $\sup |r_0(\mathbf{x}, p) - \phi_r^{k_{m,n}}(\mathbf{x}, p)'\beta_{r,n}^*| = O(k_{r,n}^{-\sigma_r})$ for some $\sigma_m, \sigma_r > 0$. When the polynomial or spline sieve spaces are used for m_0 and r_0, Assumption 4-(i) implies that $\sigma_m = p_m/(d_x + 1)$ and $\sigma_r = p_r/(Td_x + 1)$ (Newey (1997)). I denote the sequences of functions $\{\phi_m^{k_{m,n}}(x, \gamma)'\beta_{m,n}^*\}_{n=1}^\infty$ and $\{\phi_r^{k_{r,n}}(\mathbf{x}, p)'\beta_{r,n}^*\}_{n=1}^\infty$ by $\{\pi_n m_0\}_n$ and $\{\pi_n r_0\}_n$, respectively. It is also worth mentioning that Assumption 4-(i) and Assumption 5 together imply

[17] See, for example, Koenker (2005, p.124).

that $\mathcal{M}_n \subseteq \mathcal{M}_{n+1}$ and $\mathcal{R}_n \subseteq \mathcal{R}_{n+1}$ for all $n \geq 1$ and that $\overline{\cup_{n=1}^{\infty}\mathcal{M}_n} = \mathcal{M}$ and $\overline{\cup_{n=1}^{\infty}\mathcal{R}_n} = \mathcal{R}$, where, for a set A, $\overline{A}$ is the closure of A.

The next assumption imposes conditions on the first-step estimator. Before providing the assumption, I introduce additional notations. For a (d_x+1) tuple ω and a (Td_x+1) tuple $\omega^{'}$ of nonnegative integers, let $\zeta_{\kappa,m}(k_{m,n}) \equiv \sup_{x,\gamma} ||\nabla^{\omega}\phi_m^{k_{m,n}}(x,\gamma)||_E$ and $\zeta_{\kappa,m}(k_{r,n}) \equiv \sup_{\mathbf{x},p} ||\nabla^{\omega}\phi_r^{k_{r,n}}(\mathbf{x},p)||_E$. Bounds on these quantities depend on $k_{m,n}$, $k_{r,n}$, and sieves. For example, it is well-known that $\zeta_{\kappa,m}(k_{m,n}) = k_{m,n}^{1+2\kappa}$ for polynomial sieves and $\zeta_{\kappa,m}(k_{m,n}) = k_{m,n}^{\frac{1}{2}+\kappa}$ for spline sieves (cf. Newey (1997)).

Assumption 6. *(i) $h_0 \in \mathcal{H} \equiv \Lambda_{c_h}^{p_h}(Supp(X_t, Z_t))$ with $p_h > \frac{d_x+d_z}{2}$; (ii) there exists a positive non-increasing sequence $\delta_{h,n}^*$ such that $\delta_{h,n}^* \downarrow 0$ and $||\hat{h}_n - h_0||_2 = O_p(\delta_{h,n}^*)$; (iii) $\zeta_{1,m}(k_{m,n}) \cdot \zeta_{1,r}(k_{r,n}) \cdot \delta_{h,n}^* \log(\log(n)) = o(1)$.*

Condition (i) in Assumption 6 defines the parameter space for h_0. Condition (ii) in Assumption 6 requires that the first-step estimator converge at a suitable rate with respect to the L_2-norm. This condition is a high-level condition but easy to verify in practice. For example, $\delta_{h,n}^* = n^{-1/2}$ with standard parametric estimators of the selection probability. For series estimation of $h_0(x,z)$, the convergence rate depends on the number of series terms to approximate h_0 and some smoothness conditions (e.g. Newey (1997)). Condition (iii), along with Assumption 5, further restricts the rates $k_{m,n}$ and $k_{r,n}$. Once the convergence rate of the first-step estimator is established, this can be satisfied with a proper choice of $k_{m,n}$ and $k_{r,n}$. The terms $\zeta_{1,m}(k_{m,n})$ and $\zeta_{1,r}(k_{r,n})$ appear in the condition due to the nonlinearity of the conditional quantile function with respect to the selection probability.

The following theorem establishes the consistency of sieve estimator $\hat{\theta}_n$ with respect to the sup-norm $d_{\Theta,\infty}(\cdot,\cdot)$.

Theorem 6.1. *Suppose that Assumptions 1, 2, and 3 – 6 are satisfied. Then, $d_{\Theta,\infty}(\hat{\theta}_n, \theta_0) = o_p(1)$.*

Now I establish the convergence rate of the sieve estimator $\hat{\theta}_n$ with respect to L_2-norm. The following theorem s provides the L_2-convergence rate of the sieve estimator $\hat{\theta}_n$ under the same conditions in Theorem 6.1.

Theorem 6.2. *Suppose that the conditions in Theorem 6.1 are satisfied. Then,*

$$||\hat{\theta}_n - \theta_0||_{\Theta,2} = O_p\left(\sqrt{\frac{k_{m,n}}{n}} + k_{m,n}^{-(p_m/(d_x+1))} + \sqrt{\frac{k_{r,n}}{n}} + k_{r,n}^{-(p_r/(Td_x+1))} + \zeta_{1,m}(k_{m,n}) \cdot \zeta_{1,r}(k_{r,n}) \cdot \delta_{h,n}^*\right).$$

The L_2-convergence rate of the sieve estimator $\hat{\theta}_n$ can be divided into three components. The first component $\sqrt{\frac{k_{m,n}}{n}} + \sqrt{\frac{k_{r,n}}{n}}$ reflects the convergence rate of the variance term of the estimator $\hat{\theta}_n$. The rate increases as $k_{m,n}$ or $k_{r,n}$ increases. The second component $k_{m,n}^{-(p_m/(d_x+1))} + k_{r,n}^{-(p_r/(Td_x+1)}$ is the convergence rate of the approximation error, which decreases as $k_{m,n}$ or $k_{r,n}$ increases. These are consistent with the standard convergence rate of sieve or series estimators (e.g. Newey (1997) and Chen (2007)). The last component $\zeta_{1,m}(k_{m,n}) \cdot \zeta_{1,r}(k_{r,n}) \cdot \delta_{h,n}^*$ reflects the effect of the first-step estimator on the convergence rate of $\hat{\theta}_n$, and similar results are found in, for example, Newey et al. (1999) and Imbens and Newey (2009). The term $\zeta_{1,m}(k_{m,n}) \cdot \zeta_{1,r}(k_{r,n})$ appears due to nonlinearity of the conditional quantile function, which is consistent with the result in Hahn et al. (2018b, (Lemma 2.2)). It is worth pointing out that the sieve estimator $\hat{\theta}_n$ does not suffer from an ill-posed inverse problem, so it is not required to take the degree of ill-posedness into account to obtain the convergence rate.[18]

The result in Theorem 6.2 is useful to derive the convergence rate of the sieve estimator of the conditional quantile function $g_0(\mathbf{X}_i, p) \equiv m_0(X_{it}, r_0(\mathbf{X}_i, p))$. Let $\hat{g}_n(\mathbf{X}_i, p) \equiv \hat{m}_n(X_{it}, \hat{r}_n(\mathbf{X}_i, p))$ be the sieve estimator of g_0. Then, one obtains the following result:

Corollary 6.3. *Suppose that the conditions in Theorem 6.2 hold. Then,*

$$||\hat{g}_n - g_0||_2 = O_p\left(\sqrt{\frac{k_{m,n}}{n}} + k_{m,n}^{-(p_m/(d_x+1))} + \sqrt{\frac{k_{r,n}}{n}} + k_{r,n}^{-(p_r/(Td_x+1))} + \zeta_{1,m}(k_{m,n}) \cdot \zeta_{1,r}(k_{r,n}) \cdot \delta_{h,n}^*\right).$$

6.2 Asymptotic Normality for Regular Functionals

In this section, I establish the asymptotic normality of regular functionals. I adopt the approach of Hahn et al. (2018a) who provide a set of conditions under which the two plug-in sieve estimator of a (regular) functional is asymptotically normal. To this end, I assume that the first-step estimator is estimated by the sieve (or series) method. Since $\mathbb{E}[D_{it}|X_{it}, Z_{it}] = h_0(X_{it}, Z_{it})$, one can use the series method. Alternatively, one can use the series logit estimator in Hirano et al. (2003).

Many functionals of interest are functionals of the conditional quantile function, not just of the structural functions m and r. Therefore, I consider $\hat{g}_n$ itself as the second-step estimator and focus on the functionals of (g_0, h_0). For simplicity of notation, I redefine the true parameter value in terms of g_0 and h_0 (i.e. $\alpha_0 \equiv (g_0, h_0)' \in \mathcal{A}$), and a generic element in $\mathcal{A}$ is denoted by α. I denote the convergence rate of $\hat{g}_n$ provided in Corollary 6.3 by $\delta_{g,n}^*$. Define $\delta_{h,n} \equiv \delta_{h,n}^* \cdot \log(\log n)$ and $\delta_{g,n} \equiv \delta_{g,n}^* \cdot \log(\log n)$, and it is assumed that $\delta_{h,n}$ and $\delta_{g,n}$ are $o(1)$. Let $\mathcal{N}_h \equiv \{h \in \mathcal{H} : ||h - h_0||_2 \leq \delta_{h,n}\}$ and $\mathcal{N}_g \equiv \{g \in \mathcal{G} : ||g - g_0||_2 \leq \delta_{g,n}\}$ be shrinking neighborhoods of h_0 and g_0, respectively, with $\mathcal{G}$ is the parameter space of g_0. Then, define $\mathcal{N}_{h,n} \equiv \mathcal{N}_h \cap \mathcal{H}_n$ and $\mathcal{N}_{g,n} \equiv \mathcal{N}_g \cap \mathcal{G}_n$ where $\mathcal{G}_n$ the sieve space for $\mathcal{G}$. The sieve space $\mathcal{G}_n$ can be taken as $\mathcal{M}_n \circ \mathcal{R}_n$ where $\circ$ means the composition operator. Then, $(\hat{g}_n, \hat{h}_n) \in \mathcal{N}_{g,n} \times \mathcal{N}_{h,n} \equiv \mathcal{N}_{\alpha,n}$ with probability approaching to one (wpa1). Let $\mathcal{N}_\alpha \equiv \{\alpha \in \mathcal{A} : ||\alpha - \alpha_0||_{\mathcal{A},2} \leq \delta_{\alpha,n}\}$ where $\delta_{\alpha,n} = \max(\delta_{g,n}, \delta_{h,n})$.

Suppose that $\mathcal{L}_{1,n}(\mathbf{W}_i, h) - \mathcal{L}_{1,n}(\mathbf{W}_i, h_0)$ is approximated by $\Delta_1(\mathbf{W}_i, h_0)[h - h_0] \equiv \sum_t (D_{it} - h_0(X_{it}, Z_{it})) \cdot [h - h_0]$ which is linear in $[h - h_0]$. For any $h_1, h_2 \in \mathcal{N}_{h,n}$, define

$$||h_1 - h_2||_{1,\Delta}^2 \equiv -\frac{\partial \mathbb{E}\left[\Delta_1(\mathbf{W}, h_0 + \tau[h_1 - h_2])[h_1 - h_2]\right]}{\partial \tau},$$

and this is a norm on $\mathcal{N}_h$. Let $h_{o,n}$ be the projection of h_0 on $\mathcal{H}_n$ under the norm $|| \cdot ||_{1,\Delta}$. Let $\mathbb{V}_{1,n}$ be the closed linear span of $\mathcal{N}_{h,n} - \{h_{o,n}\}$ under the norm $|| \cdot ||_{1,\Delta}$, and let

$$< v_{h_1}, v_{h_2} >_{1,\Delta} \equiv -\frac{\partial \mathbb{E}\left[\Delta_1(\mathbf{W}, h_0 + \tau v_{h_2})[v_{h_1}]\right]}{\partial \tau}\bigg|_{\tau=0}$$

denote the inner product on $\mathbb{V}_{1,n}$. It is clear that this inner product induces the norm $|| \cdot ||_{1,\Delta}$,

[18] For nonparametric instrumental variables (NPIV) regression, the convergence rates are usually slower than the standard convergence rate in the literature on nonparametric estimation because of ill-posed inverse problems. For sieve estimation, this is reflected by the sieve measure of ill-posedness. See, for example, Blundell et al. (2007) and Chen and Pouzo (2012).

and therefore $(\mathbb{V}_{1,n}, ||\cdot||_{1,\Delta})$ is a Hilbert space. In addition, $\mathbb{V}_1$ denotes the closed linear span of $\mathcal{N}_h - \{h_0\}$ under the norm $||\cdot||_{1,\Delta}$.

Let $g(\mathbf{X}, Z_t; h) \equiv m(X_t, r(\mathbf{X}, h(X_t, Z_t)))$. Note that, by Knight's identity, for any $g \in \mathcal{N}_{g,n}$, $l_u(\mathbf{W}, g, h_0) - l_u(\mathbf{W}, g_0, h_0) = \Delta_2(\mathbf{W}, g_0, h_0)[g - g_0] + R_2(\mathbf{W}, g, g_0)$, where

$$\Delta_2(\mathbf{W}, g_0, h_0)[g - g_0] \equiv \sum_t [\{g(\mathbf{X}_i, Z_{it}; h_0) - g_0(\mathbf{X}_i, Z_{it}; h_0)\} \cdot \{u - \mathbf{1}(Y_{it} \leq g_0(\mathbf{X}_i, Z_{it}; h_0)\}],$$
$$R_2(\mathbf{W}, g, g_0) \equiv -\sum_t \int_0^{g(\mathbf{X}_i, Z_{it}; h_0) - g_0(\mathbf{X}_i, Z_{it}; h_0)} \{\mathbf{1}(Y_{it} \leq g_0(\mathbf{X}_i, Z_{it}; h_0) + s) - \mathbf{1}(Y_{it} \leq g_0(\mathbf{X}_i, Z_{it}; h_0))\} ds. \tag{17}$$

Then, $\Delta_2(\mathbf{W}, g_0, h_0)[g - g_0]$ is linear in $[g - g_0]$. Let

$$||g_1 - g_2||^2_{2,\Delta} \equiv -\left.\frac{\partial \mathbb{E}\left[\Delta_2(\mathbf{W}, g_0 + \tau[g_1 - g_2], h_0)][g_1 - g_2]\right]}{\partial \tau}\right|_{\tau=0}$$
$$= \mathbb{E}\left[\sum_t f_{Y_t^*|\mathbf{X}, Z_t}(g_0(\mathbf{X}_i, Z_{it}; h_0))|g_1 - g_2|^2\right]$$

be a norm on $\mathcal{N}_g$ and denote $g_{o,n}$ be the projection of g_0 on $\mathcal{G}_n$ under the norm $||\cdot||_{2,\Delta}$. Let $\mathbb{V}_{2,n}$ be the closed linear span of $\mathcal{N}_{g,n} - \{g_{o,n}\}$ under the norm $||\cdot||_{2,\Delta}$, and for any $v_{g_1}, v_{g_2} \in \mathbb{V}_{2,n}$, let

$$< v_{g_1}, v_{g_2} >_{2,\Delta} \equiv -\left.\frac{\partial \mathbb{E}\left[\Delta_2(\mathbf{W}, g_0(\mathbf{X}, Z_t; h_0) + \tau v_{g_2}, h_0)[v_{g_1}]\right]}{\partial \tau}\right|_{\tau=0}$$

be the corresponding inner product on $\mathbb{V}_{2,n}$. Then, $(\mathbb{V}_{2,n}, ||\cdot||_{2,\Delta})$ is a Hilbert space. Denote the closed linear span of $\mathcal{N}_g - \{g_0\}$ under $||\cdot||_{2,\Delta}$ by $\mathbb{V}_2$.

Let $f : \mathcal{A} \to \mathbb{R}$ be a functional and define for each $v \in \mathbb{V}_{1,n}$ and $\tilde{v} \in \mathbb{V}_{2,n}$,

$$\partial_h f(\alpha_0)[v] \equiv \left.\frac{\partial f(g_0, h_0 + \tau v)}{\partial \tau}\right|_{\tau=0},$$
$$\partial_g f(\alpha_0)[\tilde{v}] \equiv \left.\frac{\partial f(g_0 + \tau \tilde{v}, h_0)}{\partial \tau}\right|_{\tau=0}.$$

I assume that $\partial_h f(\alpha_0)[\cdot] : \mathbb{V}_1 \to \mathbb{R}$ and $\partial_g f(\alpha_0)[\cdot] : \mathbb{V}_2 \to \mathbb{R}$ are linear functionals. Since e $(\mathbb{V}_{1,n}, ||\cdot||_{1,\Delta})$ and $(\mathbb{V}_{2,n}, ||\cdot||_{2,\Delta})$ are finite-dimensional Hilbert spaces, every linear functional is continuous and hence bounded. By the Riesz representation theorem, there exists $v^*_{h_n} \in \mathbb{V}_{1,n}$ such that for any $v \in \mathbb{V}_{1,n}$,

$$\partial_h f(\alpha_0)[v] =< v^*_{h_n}, v >_{1,\Delta} \text{ and } ||v^*_{h_n}||^2_{1,\Delta} = \sup_{v \in \mathbb{V}_{1,n} - \{0\}} \frac{|\partial_h f(\alpha_0)[v]|^2}{||v||^2_{1,\Delta}} < \infty. \tag{18}$$

Similarly, there exists $v^*_{g_n} \in \mathbb{V}_{2,n}$ such that for any $v \in \mathbb{V}_{2,n}$,

$$\partial_g f(\alpha_0)[v] =< v^*_{g_n}, v >_{2,\Delta} \text{ and } ||v^*_{g_n}||^2_{2,\Delta} = \sup_{v \in \mathbb{V}_{2,n} - \{0\}} \frac{|\partial_g f(\alpha_0)[v]|^2}{||v||^2_{2,\Delta}} < \infty. \tag{19}$$

For any $v = (v_h, v_g) \in \mathbb{V} \equiv \mathbb{V}_1 \times \mathbb{V}_2$, define

$$\partial_\beta f(\alpha_0)[v] \equiv \partial_h f(\alpha_0)[v_h] + \partial_g f(\alpha_0)[v_g].$$

I impose the following assumption on the functional f:

Assumption 7. $\lim_{n\to\infty} ||v^*_{h_n}||^2_{1,\Delta} < \infty$ *and* $\lim_{n\to\infty} ||v^*_{g_n}||^2_{2,\Delta} < \infty$.

Assumption 7 means that the functional is regular. Since $\mathbb{V}_1$ and $\mathbb{V}_2$ are infinite-dimensional Hilbert spaces, one cannot directly invoke the Riesz representation theorem for $\partial_h f(\alpha_0)[\cdot]$ and $\partial_g f(\alpha_0)[\cdot]$ on $\mathbb{V}_1$ and $\mathbb{V}_2$, respectively. On the other hand, Assumption 7 imposes that the linear functionals $\partial_h f(\alpha_0)[\cdot]$ and $\partial_g f(\alpha_0)[\cdot]$ on $\mathbb{V}_1$ and $\mathbb{V}_2$, respectively, are bounded, and thus one can apply the Riesz representation theorem for $\partial_h f(\alpha_0)[\cdot]$ and $\partial_g f(\alpha_0)[\cdot]$ on $\mathbb{V}_1$ and $\mathbb{V}_2$, respectively.

For any $v_g \in \mathbb{V}_2$, let $\Gamma_g(\alpha_0)[v_g] \equiv -\sum_t \mathbb{E}[F_{Y^*_t|\mathbf{X},Z_t}(m_0(X_{it}, r_0(\mathbf{X}_i, h_0(X_{it}Z_{it}))))v_g]$ and define

$$\begin{aligned}
\Gamma(\alpha_0)[v_h, v_g] &\equiv \left.\frac{\partial \Gamma_g(g_0, h_0 + \tau v_h)[v_g]}{\partial \tau}\right|_{\tau=0} \\
&= -\sum_t \mathbb{E}[f_{Y^*_t|\mathbf{X},Z_t}(m_0(X_{it}, r_0(\mathbf{X}_i, h_0(\mathbf{X}_i Z_{it})))) \cdot \partial_\gamma m_0(X_{it}, r_0(\mathbf{X}_i, h_0(\mathbf{X}_i Z_{it}))) \\
&\quad \times \partial_p r_0(\mathbf{X}_i, h_0(X_{it}Z_{it})) \cdot v_g \cdot v_h]
\end{aligned} \tag{20}$$

for any $v_h \in \mathbb{V}_1$. Then, $\Gamma(\alpha_0)[\cdot,\cdot]$ is a bilinear functional on $\mathbb{V}$. Define $v^*_{\Gamma_n} \in \mathbb{V}_{1,n}$ as

$$\Gamma(\alpha_0)[v_h, v^*_{g_n}] = < v_h, v^*_{\Gamma_n} >_{1,\Delta}$$

for all $v_h \in \mathbb{V}_{1,n}$, where $v^*_{g_n}$ is the same as in (19). Let

$$||v^*_n||_{sd} \equiv Var\left[n^{-1/2}\sum_i^n \left\{\Delta_1(\mathbf{W}_i, h_0)[v^*_{h_n}] + \Delta_1(\mathbf{W}_i, h_0)[v^*_{\Gamma_n}] + \Delta_2(\mathbf{W}_i, \beta_0)[v^*_{g_n}]\right\}\right].$$

To establish the asymptotic normality of $f(\hat{\alpha}_n)$, I impose additional assumptions.

Assumption 8. *The following conditions hold:*

(i) $\liminf_n ||v^*_n||_{sd} > 0$*; (ii) the functional* $f(\cdot)$ *satisfies*

$$\sup_{\alpha\in\mathcal{N}_{\alpha,n}} \left|\frac{f(\alpha) - f(\alpha_0) - \partial_g f(\alpha_0)[\alpha - \alpha_0]}{||v^*_n||_{sd}}\right| = o\left(n^{-1/2}\right);$$

(iii) the following condition holds:

$$\frac{1}{||v^*_n||_{sd}} \max\left\{\left|\partial_h f(\alpha_0)[h_{o,n} - h_0]\right|, \left|\partial_g f(\alpha_0)[g_{o,n} - g_0]\right|\right\} = o\left(n^{-1/2}\right).$$

Condition (i) implies that the sieve variance is asymptotically bounded away from zero. Condition (ii) is implied if the functional is well-approximated uniformly on the neighborhood of α_0. Condition (iii) is an overfitting condition that guarantees the sieve approximation errors converge to zero at a faster rate than a certain rate ($n^{-1/2}||v^*_n||_{sd}$). Assumption 8 can be verified once a functional of interest is given.

Let $(u^*_{h_n}, u^*_{g_n}, u^*_{\Gamma_n}) \equiv ||v_n||^{-1}_{sd}(v^*_{h_n}, v^*_{g_n}, v^*_{\Gamma_n})$ and $\mu_n[f] \equiv \frac{1}{n}\sum_i^n[f(\mathbf{W}_i) - \mathbb{E}[f(\mathbf{W}_i)]]$ be an empirical process indexed by f.

Assumption 9. *(i) The following condition is satisfied:*

$$\left| < \hat{h}_n - h_{o,n}, u^*_{h_n} + u^*_{\Gamma_n} >_{1,\Delta} -\mu_n \left[\Delta_1(\mathbf{W}, h_0)[u^*_{h_n} + u^*_{\Gamma_n}]\right] \right| = O_p(\kappa_n);$$

(ii) the following convergence in distribution holds:

$$\frac{1}{\sqrt{n}}\sum_i^n \left\{\Delta_1(\mathbf{W}_i, h_0)[u^*_{h_n} + u^*_{\Gamma_n}] + \Delta_2(\mathbf{W}_i, g_0, h_0)[u^*_{g_n}]\right\} \xrightarrow{d} N(0,1);$$

(iii) $\kappa_n \cdot \delta^{*^{-1}}_{g,n} = o(1)$ *with* $\kappa_n = o(n^{-1/2})$ *and* $||u^*_{g_n}||_{2,\Delta} = O(1)$.

Assumption 9 corresponds to Assumption 3.3 in Hahn et al. (2018a). The first condition is a high-level condition, and it can be verified once the first-stage estimation method is chosen. The second condition can be verified by using some appropriate central limit theorem. The last condition is mild as $\delta^*_{g,n}$ is a convergence rate of a nonparametric estimator, which is not faster than $\sqrt{n}$.

Assumption 10. *(i) For all* $t = 1, 2, ..., T$, *the conditional density function* $f_{Y_t^*|\mathbf{X},Z_t}(y|\mathbf{x}, z_t)$ *is continuously differentiable with respect to* y, *and its derivative* $f'_{Y_t|\mathbf{X},Z_t}(y|\mathbf{x}, z_t)$ *is uniformly bounded; (ii) the functions* $m_0(x,\gamma)$ *and* $r_0(\mathbf{x}, p)$ *are twice-continuously differentiable with respect to* γ *and* p, *respectively, and the second-order derivatives are uniformly bounded.*

Assumption 11. *Let* $\sup_{h\in\mathcal{N}_{h,n}} |h(x,z) - h_0(x,z)| = O(\delta^{*\,\sup}_{h,n})$ *and* $\delta^{\sup}_{h,n} \equiv \delta^{*\,\sup}_{h,n}\log(\log(n)) = o(1)$. *Then, the following conditions hold:*

(i) $\zeta_{1,m}(k_{m,n})^2 \cdot \zeta_{1,r}(k_{r,n})^2 \cdot \delta^2_{h,n}(\zeta_{0,m}(k_{m,n}) + \zeta_{0,r}(k_{r,n})) \cdot \delta_{\theta,n} = o(n^{-1})$;
(ii) $\delta^3_{\theta,n} \cdot (\zeta_{0,m}(k_{m,n}) + \zeta_{0,r}(k_{r,n})) = o(n^{-1})$;
(iii) $\zeta_{1,m}(k_{m,n})^3 \cdot \zeta_{1,r}(k_{r,n})^3 \cdot \delta^2_{h,n}\delta^{\sup}_{h,n} = o(n^{-1})$;
(iv) $\delta^2_{\theta,n} \cdot \zeta_{1,m}(k_{m,n}) \cdot \zeta_{1,r}(k_{r,n}) \cdot \delta^{\sup}_{h,n} = o(n^{-1})$;
(v) $\left\{\zeta_{2,m}(k_{m,n}) \cdot \zeta_{1,r}(k_{r,n})^2 + \zeta_{1,m}(k_{m,n}) \cdot \zeta_{2,r}(k_{r,n})\right\}\delta^2_{h,n} = o(n^{-1/2})$;
(iv) $\zeta_{1,r}(k_{r,n})\delta_{\theta,n}\delta_{h,n} = o(n^{-1/2})$.

Assumption 10 strengthens the smoothness conditions on the conditional density function and the structural functions m_0 and r_0. Assumption 11 restricts the rates of $k_{m,n}$, $k_{r,n}$ and possibly $k_{h,n}$ where $k_{h,n}$ is the number of approximating functions for h_0. Once the sieve space is chosen and the convergence rates are derived, one can set the rates on $k_{m,n}$, $k_{r,n}$ and $k_{h,n}$ that satisfy Assumption 11. The following theorem establishes the asymptotic normality of $f(\hat{\alpha}_n)$:

Theorem 6.4. *Suppose that conditions in Theorem 6.2 hold. If Assumptions 7–11 additionally hold, then*

$$\sqrt{n}\frac{f(\hat{\alpha}_n) - f(\alpha_0)}{||v^*_n||_{sd}} \xrightarrow{d} N(0,1).$$

7 Simulation

In this section, I present results of Monte-Carlo simulations to examine the finite-sample performance of the estimators. To this end, I consider the following model with $T = 3$:

$$Q_{Y_t|\mathbf{X},Z_t,D_t=1}(u|\mathbf{X}, Z_t, D_t = 1) = X_t\beta(u) + \mathbf{X}^{'}\delta(u) + Q_{U_t|\mathbf{X},Z_t,D_t=1}(u|\mathbf{X}, Z_t, D_t = 1),$$
$$D_t = \mathbf{1}(X_t\xi + Z_t\gamma \geq V_t),$$

where $(U_t, V_t) \sim BVN\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.3\\0.3 & 1\end{pmatrix}\right)$ and X_t and Z_t are drawn from normal distributions. The true parameter values are $\beta(u) = 1 + 0.5 \times Q_{U_t}(u)$, $\delta_t(u) = 1 + 0.2 \times Q_{U_t}(u)$, $\xi = -1$, and $\gamma = 1$. In the simulation, I assume that the first-stage equation is correctly specified, and therefore the parameters ξ and γ are estimated using a probit model.

I consider two sample sizes, $N \in \{500, 1000\}$, and I use the Hermite polynomial sieve spaces to approximate the unknown function $Q_{U_t|\mathbf{X},Z_t,D_t=1}(u|\mathbf{X}, Z_t, D_t = 1)$, and the order is set to be proportional to $N^{1/7}$.[19] I focus on the root mean squared errors (RMSEs) of $\beta(u)$ and $\delta(u)$ for $u \in \{0.25, 0.5, 0.75\}$ to investigate the finite-sample performance, and obtain the results from 2000 iterations.

Tables 1 and 2 present the simulation results with sample sizes 500 and 1000, respectively. For all quantile levels, the biases of the estimators are negligible and the standard deviations are also small. Therefore, the results suggest that the semiparametric estimators of $\beta(u)$ and $\delta(u)$ perform well in finite samples.

Table 1: Semiparametric Model, $N = 500$, $T = 3$

	$u = 0.25$			$u = 0.5$			$u = 0.75$		
	Bias	S.D.	RMSE	Bias	S.D.	RMSE	Bias	S.D.	RMSE
$\beta(u)$	-0.0011	0.0778	0.0778	0.0029	0.1104	0.1105	0.0018	0.0786	0.0787
$\delta_1(u)$	0.0025	0.0511	0.0511	0.0056	0.0448	0.0451	-0.0020	0.0510	0.0510
$\delta_2(u)$	0.0032	0.0503	0.0504	0.0025	0.0441	0.0441	-0.0033	0.0507	0.0508
$\delta_3(u)$	-0.0054	0.0574	0.0576	-0.0055	0.0501	0.0504	0.0040	0.0533	0.0534

Table 2: Semiparametric Model, $N = 1000$, $T = 3$

	$u = 0.25$			$u = 0.5$			$u = 0.75$		
	Bias	S.D.	RMSE	Bias	S.D.	RMSE	Bias	S.D.	RMSE
$\beta(u)$	0.0006	0.0665	0.0665	0.0016	0.0486	0.0486	-0.0006	0.0545	0.0545
$\delta_1(u)$	0.0036	0.0354	0.0355	0.0037	0.0315	0.0317	-0.0016	0.0354	0.0354
$\delta_2(u)$	0.0027	0.0356	0.0357	0.0023	0.0327	0.0328	-0.0018	0.0360	0.0360
$\delta_3(u)$	-0.0056	0.0397	0.0401	-0.0085	0.0348	0.0358	0.0026	0.0388	0.0389

[19] The order of the Hermite polynomial is 4 when $N = 500$, whereas it is 5 when $N = 1000$.

8 Conclusion

In this paper, I develop a nonparametric panel quantile regression model with sample selection. The model is nonseparable and allows for time-invariant endogeneity in a similar spirit of the fixed effects models. To resolve the time-invariant endogeneity of the regressors and the sample selection bias, I adopt the CRE and control function approaches. In doing so, I avoid imposing any parametric or semiparametric restrictions on the distribution of the unobserved error terms, except for a conditional independence condition. The class of models is general and flexible enough to be extended to address many empirical issues about data, such as time-varying endogeneity and censoring. I study identification of the structural functions of the model. Identification requires that the number of time periods be greater than or equal to 3 ($T \geq 3$) and that there exist excluded variables that affect the selection probability. For practically tractable estimation, I also suggest some semiparametric models and present a set of identification conditions for the semiparametric models. Based on the identification result, I use a two-step nonparametric sieve method to estimate the model parameters. I establish the consistency and convergence rates of the two-step sieve estimators under low-level conditions. I also provide a set of conditions under which the plug-in estimate of a smooth functional of the parameter is asymptotically normal. A small Monte-Carlo study with semiparametric models confirms that the estimators perform well in finite samples.

References

Abrevaya, J. and C. M. Dahl (2008). The effects of birth inputs on birthweight: evidence from quantile estimation on panel data. *Journal of Business & Economic Statistics 26*(4), 379–397. 1, 2, 9, 2.2

Ahn, H. and J. L. Powell (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics 58*(1-2), 3–29. 1

Altonji, J. G. and R. L. Matzkin (2005). Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica 73*(4), 1053–1102. 7, 1, 2, 3.1

Arellano, M. and S. Bonhomme (2012). Identifying distributional characteristics in random coefficients panel data models. *The Review of Economic Studies 79*(3), 987–1020. 2, 3.1

Arellano, M. and S. Bonhomme (2016). Nonlinear panel data estimation via quantile regressions. *The Econometrics Journal 19*(3), C61–C94. 1, 2, 2.3, 3.1

Arellano, M. and S. Bonhomme (2017). Quantile selection models with an application to understanding changes in wage inequality. *Econometrica 85*(1), 1–28. 1, 1, 2, 3.1

Besstremyannaya, G. and S. Golovan (2019). Reconsideration of a simple approach to quantile regression for panel data. *The Econometrics Journal 22*(3), 292–308. 1, 4

Bester, C. A. and C. Hansen (2009). Identification of marginal effects in a nonparametric correlated random effects model. *Journal of Business & Economic Statistics 27*(2), 235–250. 1, 2, 2, 3.1

Blundell, R., X. Chen, and D. Kristensen (2007). Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica 75*(6), 1613–1669. 3.1, 18

Buchinsky, M. (1998). The dynamics of changes in the female wage distribution in the USA: a quantile regression approach. *Journal of Applied Econometrics 13*(1), 1–30. 1, 1, 2, 2, 11, 4.2

Canay, I. A. (2011). A simple approach to quantile regression for panel data. *The Econometrics Journal 14*(3), 368–386. 1, 4

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *The Review of Economic Studies 47*(1), 225–238. (document), 1, 1, 2, 4.1, 4.1

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics 18*(1), 5–46. 2

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics 6B*, 5549–5632. 5, 6.1, A.3

Chen, X. and T. M. Christensen (2018). Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression. *Quantitative Economics 9*(1), 39–84. 6.1, A.3

Chen, X. and D. Pouzo (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica 80*(1), 277–321. 5, 16, 18

Chen, X. and X. Shen (1998). Sieve extremum estimates for weakly dependent data. *Econometrica 66*(2), 289–314. A.1

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013). Average and quantile effects in nonseparable panel models. *Econometrica 81*(2), 535–580. 1, 2

Chernozhukov, V., I. Fernández-Val, S. Hoderlein, H. Holzmann, and W. Newey (2015). Nonparametric identification in panels using quantiles. *Journal of Econometrics 188*(2), 378–392. 1, 2

Chernozhukov, V., I. Fernández-Val, and A. E. Kowalski (2015). Quantile regression with censoring and endogeneity. *Journal of Econometrics 186*(1), 201–221. 3.2.1, 3.2.2

Chernozhukov, V., I. Fernández-Val, and S. Luo (2018). Distribution regression with sample selection, with an application to wage decompositions in the UK. *arXiv preprint arXiv:1811.11603*. 1, 1, 2, 3.1

Das, M., W. K. Newey, and F. Vella (2003). Nonparametric estimation of sample selection models. *The Review of Economic Studies 70*(1), 33–58. 1, 1, 2, 3.2.1, 4.2, 4.2

Donald, S. G. (1995). Two-step estimation of heteroskedastic sample selection models. *Journal of Econometrics 65*(2), 347–380. 1

Dustmann, C. and M. E. Rochina-Barrachina (2007). Selection correction in panel data models: an application to the estimation of females' wage equations. *The Econometrics Journal 10*(2), 263–293. 6

Escanciano, J. C., D. Jacho-Chávez, and A. Lewbel (2016). Identification and estimation of semiparametric two-step models. *Quantitative Economics 7*(2), 561–589. 3.1, 4.1

Evdokimov, K. (2010). Identification and estimation of a nonparametric panel data model with unobserved heterogeneity. *Department of Economics, Princeton University*. 1, 4.1

Fernández-Val, I., F. Peracchi, A. van Vuuren, and F. Vella (2019). Decomposing real wage changes in the United States. *arXiv preprint arXiv:1901.00419v2*. 2, 3.2.2

Fernández-Val, I. and M. Weidner (2018). Fixed effects estimation of large-T panel data models. *Annual Review of Economics 10*, 109–138. 3

Graham, B. S., J. Hahn, A. Poirier, and J. L. Powell (2018). A quantile correlated random coefficients panel data model. *Journal of Econometrics 206*(2), 305–335. 2, 3.1

Graham, B. S. and J. L. Powell (2012). Identification and estimation of average partial effects in "irregular" correlated random coefficient panel data models. *Econometrica 80*(5), 2105–2152. 2, 3.1

Gronau, R. (1974). Wage comparisons–A selectivity bias. *Journal of Political Economy 82*(6), 1119–1143. 1

Hahn, J., Z. Liao, and G. Ridder (2018a). Nonparametric two-step sieve M estimation and inference. *Econometric Theory 34*(6), 1281–1324. 6, 6.2, 6.2, A.3, B, B.3, B.3

Hahn, J., Z. Liao, and G. Ridder (2018b). Supplemental appendix for "Nonparametric two-step sieve M estimation and inference". *Econometric Theory 34*(6). 6.1, B, B.1, B.1, B.2, B.2, B.2

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica 47*(1), 153–161. 1, 1, 1, 2, 2

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189. 5, 6.2

Hoderlein, S. and H. White (2012). Nonparametric identification in nonseparable panel data models with generalized fixed effects. *Journal of Econometrics 168*(2), 300–314. 1, 2

Huber, M. and B. Melly (2015). A test of the conditional independence assumption in sample selection models. *Journal of Applied Econometrics 30*(7), 1144–1168. 1

Imbens, G. W. and W. K. Newey (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica 77*(5), 1481–1512. 2, 2, 3.1, 3.2.1, 3.2.2, 6.1

Kato, K., A. F. Galvao Jr, and G. V. Montes-Rojas (2012). Asymptotics for panel quantile regression models with individual effects. *Journal of Econometrics 170*(1), 76–91. 1, 5

Klein, R. W. and R. H. Spady (1993). An efficient semiparametric estimator for binary response models. *Econometrica 61*(2), 387–421. 5

Knight, K. (1998). Limiting distributions for L_1 regression estimators under general conditions. *The Annals of Statistics 26*(2), 755–770. A.1

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis 91*(1), 74–89. 1

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press. 10, 17

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica 46*(1), 33–50. 1, 2.3

Kyriazidou, E. (1997). Estimation of a panel data sample selection model. *Econometrica 65*(6), 1335–1364. 1, 6

Laage, L. (2019). A correlated random coefficient panel model with time-varying endogeneity. *Working Paper*. 2, 3.1, 14

Lee, S. (2007). Endogeneity in quantile regression models: A control function approach. *Journal of Econometrics 141*(2), 1131–1158. 3.2.1

Lewbel, A. and O. Linton (2007). Nonparametric matching and efficient estimators of homothetically separable functions. *Econometrica 75*(4), 1209–1227. 3.1

Matzkin, R. L. (2007). Nonparametric identification. *Handbook of Econometrics 6B*, 5307–5368. 1, 3.1

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica 46*(1), 69–85. (document), 1, 1, 2, 3.1, 4.1

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics 79*(1), 147–168. 6.1, 6.1, 6.1

Newey, W. K. (2009). Two-step series estimation of sample selection models. *The Econometrics Journal 12*, S217–S229. 1, 1, 2

Newey, W. K. and J. L. Powell (2003). Instrumental variable estimation of nonparametric models. *Econometrica 71*(5), 1565–1578. 3.1

Newey, W. K., J. L. Powell, and F. Vella (1999). Nonparametric estimation of triangular simultaneous equations models. *Econometrica 67*(3), 565–603. 3.2.1, 6.1

Neyman, J. and E. L. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica 16*(1), 1–32. 1

Powell, J. L. (1986). Censored regression quantiles. *Journal of Econometrics 32*(1), 143–155. 3.2.2

Semykina, A. and J. M. Wooldridge (2010). Estimating panel data models in the presence of endogeneity and selection. *Journal of Econometrics 157*(2), 375–380. 1, 3.2.1

Semykina, A. and J. M. Wooldridge (2013). Estimation of dynamic panel data models with sample selection. *Journal of Applied Econometrics 28*(1), 47–61. 1

Van de Geer, S. A. (2000). *Empirical processes in M-estimation*. Cambridge University Press. A.1, A.2, A.2

Van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge University Press. A.2, A.2

Van der Vaart, A. W. and J. A. Wellner (1996). *Weak convergence and empirical processes*. Springer. A.1, A.2, A.2, A.3

Vytlacil, E. and N. Yildiz (2007). Dummy endogenous variables in weakly separable models. *Econometrica 75*(3), 757–779. 3.1

Wooldridge, J. M. (1995). Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics 68*(1), 115–132. 1, 6

A Proofs of the Results in Section 6

In this section, I provide mathematical proofs of the main results in Section 6. I introduce notation that will be used in the proofs. For any positive real sequences $\{a_n\}$ and $\{b_n\}$, $a_n \lesssim b_n$ means that there exist a finite constant $C > 0$ and $N \in \mathbb{N}$ such that $a_n \leq Cb_n$ for all $n \geq N$. If $a_n \lesssim b_n$ and $b_n \lesssim a_n$, it is denoted by $a_n \asymp b_n$. Let $(\mathcal{F}, ||\cdot||_{\mathcal{F}})$ be a metric space of real valued function $f : \mathcal{X} \to \mathbb{R}$. The covering number $N(\epsilon, \mathcal{F}, ||\cdot||_{\mathcal{F}})$ is the minimum number of $||\cdot||_{\mathcal{F}}$ ϵ-balls that cover $\mathcal{F}$. The entropy is the logarithm of the covering number. An ϵ-bracket in $(\mathcal{F}, ||\cdot||_{\mathcal{F}})$ is a pair of functions $l, u \in \mathcal{F}$ such that $||l||_{\mathcal{F}}, ||u||_{\mathcal{F}} < \infty$ and $||u - l||_{\mathcal{F}} \leq \epsilon$. The covering number with bracketing $N_{[]}(\epsilon, \mathcal{F}, ||\cdot||_{\mathcal{F}})$ is the minimum number of $||\cdot||_{\mathcal{F}}$ ϵ-brackets that cover $\mathcal{F}$. The entropy with bracketing is the logarithm of the covering number with bracketing. The bracketing integral is defined as $\int_0^\delta \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, ||\cdot||_{\mathcal{F}})}d\epsilon$, and it is denoted by $J_{[]}(\delta, \mathcal{F}, ||\cdot||_{\mathcal{F}})$. Let $\mu_n[f] \equiv \frac{1}{n}\sum_i \{f(\mathbf{W}_i) - \mathbb{E}[f(\mathbf{W}_i)]\}$ be an empirical process indexed by $f \in \mathcal{F}$ for some space $\mathcal{F}$. $\mathbb{G}_n[f]$ is an empirical process such that $\sqrt{n}\mu_n[f]$ (i.e. $\mathbb{G}_n[f] = \frac{1}{\sqrt{n}}\sum_i \{f(\mathbf{W}_i) - \mathbb{E}[f(\mathbf{W}_i)]\}$). Recall that $||\hat{h}_n - h_0||_2 = O_p(\delta^*_{h,n})$ by Assumption 6. Let C denote a generic positive and finite constant. It can be different across where it appears.

Some empirical processes may not be measurable, and thus the expectation operator cannot be applied to those processes. In such a case, one can replace the expectation operator with the outer expectation operator. I use the notation $\mathbb{E}[\cdot]$ mainly to indicate the expectation operator, but it may also stand for the outer expectation if its argument is not measurable.

A.1 Proof of Theorem 6.1

Proof. Let $\tilde{B}_n(h_0) \equiv \{h \in \mathcal{H}_n : ||h - h_0||_2 \leq \delta_{h,n}\}$ with $\delta_{h,n} \equiv \delta^*_{h,n} \cdot \log(\log(n))$. I verify the condition in Proposition B.1 in Section B. For Assumption C.1, recall that $l_u(\mathbf{W}_i, \alpha) = \sum_t \rho_u(\mathbf{Y}_{it} - m(X_t, r(\mathbf{X}_i, h(X_{it}, Z_{it}); u); u))$. For simplicity of notation, let $q_u(\mathbf{X}_i, Z_{it}) \equiv m_0(X_t, r_0(\mathbf{X}_i, h_0(X_{it}, Z_{it}); u); u)$. Then,

$$
\begin{aligned}
&|\mathbb{E}[l_u(\mathbf{W}_i, \theta_0, h_0)|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{D}_i]| \\
=&|\sum_t \mathbb{E}[\rho_u(\mathbf{Y}^*_{it} - q_u(\mathbf{X}_i, Z_{it}))|\mathbf{X}_i, \mathbf{Z}_i]| \\
\leq&\sum_t [|\{u \int_{q_u(\mathbf{X}_i, Z_{it})}^{\infty} \{y - q_u(\mathbf{X}_i, Z_{it})\} dF_{Y^*_t|\mathbf{X}_i, Z_{it}}(y|\mathbf{X}_i, Z_{it})| \\
&+ |(1-u)\int_{-\infty}^{q_u(\mathbf{X}_i, Z_{it})} \{y - q_u(\mathbf{X}_i, Z_{it})\} dF_{Y^*_t|\mathbf{X}_i, Z_{it}}(y|\mathbf{X}_i, Z_{it})\}|] \\
\lesssim&\mathbb{E}[|Y^*_{it} - q_u(\mathbf{X}_i, Z_{it})||\mathbf{X}_i, Z_{it}],
\end{aligned}
$$

where the first inequality holds by the triangle inequality and the second inequality is derived by Jensen's inequality and the fact that $\max(u, 1-u) \leq 1$. By condition (iii) in Assumption 3, one obtains that $|\mathbb{E}[l_u(\mathbf{W}_i, \theta_0, h_0)]| < \infty$. Now I check condition (ii) in Assumption C.1. Take any $\epsilon > 0$. Following Knight's identity (Knight (1998)), one can obtain that

$$\rho_u(w) - \rho_u(w - v) = -v(u - \mathbf{1}(w \leq 0)) + \int_0^v \{\mathbf{1}(w \leq t) - \mathbf{1}(w \leq 0)\}dt.$$

Applying the above identity with $w_{it} = Y_{it} - q_u(\mathbf{X}_i, Z_{it})$ and $v_{it} = m(X_{it}, r(\mathbf{X}_i, h_0(X_{it}, Z_{it}))) - q_u(\mathbf{X}_i, Z_{it})$ such that $d_{\Theta,\infty}(\theta, \theta_0) \geq \epsilon$ results in that

$$\begin{aligned}
&\mathbb{E}[l_u(\mathbf{W}_i, \theta_0, h_0) - l_u(\mathbf{W}_i, \theta, h_0)|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{D}_i] \\
&= \sum_t \mathbb{E}[\rho_u(Y_{it} - q_u(\mathbf{X}_i, Z_{it})) - \rho_u(Y_{it} - m(X_{it}, r(\mathbf{X}_i, Z_{it})))|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{D}_i] \\
&= \sum_t \mathbb{E}[\int_0^{v_{it}} \{\mathbf{1}(w_{it} \leq \tau) - \mathbf{1}(w_{it} \leq 0)\} d\tau |\mathbf{X}_i, \mathbf{Z}_i, \mathbf{D}_i] \\
&= \sum_t \int_0^{v_{it}} F_{Y_t^*|\mathbf{X}_i, Z_{it}}(q_u(\mathbf{X}_i, Z_{it}) + \tau) - F_{Y_t^*|\mathbf{X}_i, Z_{it}}(q_u(\mathbf{X}_i, Z_{it})) d\tau \\
&\geq \sum_t \int_0^{v_{it}} \tau d\tau \underline{f}^* \geq \underline{f}^* \sum_t v_{it}^2,
\end{aligned} \tag{21}$$

where $\underline{f}^* \equiv \min_t \inf_{y \in \mathbb{R}, (\mathbf{x}, p) \in Supp(\mathbf{X}, Z_t)} f_{Y_t^*|\mathbf{X}, Z_t}(y|\mathbf{x}, z) > 0$ by Assumption 3-(ii). It is straightforward to see that $\mathbb{E}[l_u(\mathbf{W}_i, \theta_0, h_0) - l_u(\mathbf{W}_i, \theta, h_0)] \gtrsim ||\theta - \theta_0||_{\Theta,2}^2$ by Assumptions (2) and 4-(ii). By Lemma 2 in Chen and Shen (1998), $||\theta - \theta_0||_{\Theta,2} \gtrsim d_{\Theta,\infty}(\theta, \theta_0)^{1 + \frac{1}{2p_\theta}}$ and it leads to that for all $n \geq 1$, $c_n(\epsilon) = C\epsilon^{2 + \frac{1}{p_\theta}}$ for some constant $C > 0$. With this construction, $c_n(\epsilon)$ is a positive non-increasing sequence and it obviously satisfies the condition $\liminf_n c_n(\epsilon) > 0$. Therefore Assumption C.1 is satisfied.

Now I verify Assumption C.2. Note that for given $u \in \mathcal{U}$,

$$\begin{aligned}
|\mathbb{E}[l_u(\mathbf{W}_i, \pi_n\theta_0, h_0) - l_u(\mathbf{W}_i, \theta_0, h_0)]| &\leq \mathbb{E}|l(\mathbf{W}_i, \pi_n\theta_0, h_0) - l(\mathbf{W}_i, \theta_0, h_0)| \\
&\leq \max(u, 1-u)\mathbb{E}|\pi_n\theta_0 - \theta_0| \\
&\lesssim \sup_{x,\gamma} |\pi_n m_0(x, \gamma) - m_0(x, \gamma)| + \sup_{\mathbf{x}, p} |\pi_n r_0(\mathbf{x}, p) - r_0(\mathbf{x}, p)| \\
&= O\left(k_{m,n}^{-p_m/(d_x+1)}\right) + O\left(k_{r,n}^{-p_r/(T \cdot d_x+1)}\right),
\end{aligned}$$

where the third line holds because $\sup_{x,\gamma} |m_{0,\gamma}(x, \gamma)| < \infty$, where $m_{0,\gamma}(x, \gamma) \equiv \frac{\partial m_0(x,\gamma)}{\partial \gamma}$, and the last equality holds by Assumption 4 and 5. Since $k_{m,n}, k_{r,n} \to \infty$ by Assumption 5, one can set $\eta_{2,n} \equiv \max\left(k_{m,n}^{-p_m/(d_x+1)}, k_{r,n}^{-p_r/(T \cdot d_x+1)}\right) = o(1)$.

Lastly, I verify Assumption C.3. Define $\mathcal{L}_{u,n} \equiv \{l_u(\mathbf{W}_i, \theta, h) : \theta \in \Theta_n, h \in \tilde{B}_n(h_0)\}$. I derive the convergence rate $\eta_{0,n}$ by applying Theorem 2.14.2 in Van der Vaart and Wellner (1996). Note that one can take a constant function as an envelope function of $\mathcal{L}_{u,n}$ by Assumption 4-(i). It remains to calculate the bracketing integral of $\mathcal{L}_{u,n}$. Note that

$$\mathbb{E}|l_u(\mathbf{W}_i, \theta, h) - l_u(\mathbf{W}_i, \tilde{\theta}, \tilde{h})|^2 \leq \max(u, 1-u) \sum_t \mathbb{E}|\theta(W_{it}, h) - \tilde{\theta}(W_{it}, \tilde{h})|^2 \lesssim ||\alpha - \tilde{\alpha}||_2^2$$

by Assumption 4-(ii). Applying Theorem 2.7.11 in Van der Vaart and Wellner (1996) results in $N_{[]}(C_\alpha\epsilon, \mathcal{L}_{u,n}, ||\cdot||_2) \leq N(\epsilon, \tilde{\mathcal{A}}_n, ||\cdot||_2)$ where $\tilde{\mathcal{A}}_n \equiv \mathcal{M}_n \times \mathcal{R}_n \times \tilde{B}_n(h_0)$. It can also be easily shown that

$$\log N(\epsilon, \mathcal{A}_n, ||\cdot||_2) \leq \log N(\frac{\epsilon}{4}, \mathcal{M}_n, ||\cdot||_2) + \log N(\frac{\epsilon}{4}, \mathcal{R}_n, ||\cdot||_2) + \log N(\frac{\epsilon}{2}, \tilde{B}_n(h_0), ||\cdot||_2)$$

by the definition of the covering number. Therefore, for any $\delta > 0$, one obtains that

$$J_{[]}(1, \mathcal{L}_{u,n}, ||\cdot||_2) \leq \int_0^1 \sqrt{1 + \log N(\frac{\epsilon}{4}, \mathcal{M}_n, ||\cdot||_2) + \log N(\frac{\epsilon}{4}, \mathcal{R}_n, ||\cdot||_2) + \log N(\frac{\epsilon}{2}, \tilde{B}_n(h_0), ||\cdot||_2)} d\epsilon.$$

By Lemma 2.5 in Van de Geer (2000), $\log N(\frac{\epsilon}{4}, \mathcal{M}_n, d) \leq k_{n,m} \log\left(1 + \frac{16c_m}{\epsilon}\right)$ and $\log N(\frac{\epsilon}{4}, \mathcal{R}_n, d) \lesssim k_{r,n} \log\left(1 + \frac{16c_r}{\epsilon}\right)$. To get a bound on $N(\frac{\epsilon}{2}, \tilde{B}_n, ||\cdot||_2)$, I apply Corollary 2.7.4 in Van der Vaart and Wellner (1996). Since $N(\epsilon, \mathcal{F}, ||\cdot||) \leq N_{[]}(\epsilon, \mathcal{F}, ||\cdot||)$ for any class of real valued functions $\mathcal{F}$ and any norm $||\cdot||$, it is enough to bound $N_{[]}(\frac{\epsilon}{2}, B_n(h_0), ||\cdot||_2)$. Since $\mathcal{H}_n \subseteq \mathcal{H}$, $\tilde{B}_n(h_0) \subseteq B_n(h_0) \equiv \{h \in \mathcal{H} : ||h - h_0||_2 \leq \delta_{h,n}\}$. Note that, since the support of (X_t', Z_t') is compact by Assumption 3, one can set the $M_j = 0$ for j's such that I_j^1's are outside the support of (X_t', Z_t'), where I_j^1's and M_j's are the same as defined in Corollary 2.7.4 in Van der Vaart and Wellner (1996). Pick any $V \in \mathbb{R}$ such that $(d_x + d_z)/p_h \leq V < 2$, then it follows that

$$\log N_{[]}(\frac{\epsilon}{2}, B_n(h_0), ||\cdot||_2) \lesssim \epsilon^{-V} \delta_{h,n}^V$$

by the definition of $B_n(h_0)$. This bound results in that $\int_0^1 \sqrt{\log N(\frac{\epsilon}{2}, B_n(h_0), ||\cdot||_2)} d\epsilon \lesssim \delta_{h,n}^{V/2} \int_0^1 \epsilon^{-V/2} d\epsilon = O(\delta_{h,n}^{V/2})$ by the condition on V. Therefore,

$$\begin{aligned} J_{[]}(1, \mathcal{L}_{u,n}, ||\cdot||_2) &\lesssim \int_0^1 \sqrt{\max(k_{m,n}, k_{r,n}) \log(1 + \frac{C}{\epsilon})} + \int_0^1 \sqrt{\log N(\frac{\epsilon}{2}, B_n(h_0), ||\cdot||_2)} d\epsilon \\ &\lesssim \sqrt{\max(k_{m,n}, k_{r,n})} + \sqrt{\delta_{h,n}^V}. \end{aligned}$$

Applying Theorem 2.14.2 in Van der Vaart and Wellner (1996) yields that

$$\mathbb{E}\left[\sup_{\theta \in \Theta_n, h \in \tilde{B}_n(h_0)} |\mu_n[l_u(\mathbf{W}_i, \theta, h)]|\right] \lesssim \frac{1}{\sqrt{n}} J_{[]}(1, \mathcal{L}_{u,n}, ||\cdot||_2) = O\left(\sqrt{\frac{\max(k_{m,n}, k_{r,n})}{n}} + \sqrt{\frac{\delta_{h,n}^V}{n}}\right),$$

and thus one can set $\eta_{0,n} = \sqrt{\frac{\max(k_{m,n}, k_{r,n})}{n}} + \sqrt{\frac{\delta_{h,n}^V}{n}}$ by the Markov inequality. By Assumption 5, one obtains that $\eta_{0.n} \to 0$. For the second condition in Assumption C.3, note that

$$\begin{aligned} &\sup_{\theta \in \Theta_n, h \in \tilde{B}_n(h_0)} |\mathbb{E}[l_u(\mathbf{W}_i, \theta, h) - l_u(\mathbf{W}_i, \theta, h_0)]| \\ &\leq \sup_{\theta \in \Theta_n, h \in \tilde{B}_n(h_0)} \max(u, 1-u) \sum_t \mathbb{E}|m(X_{it}, r(\mathbf{X}_i, h(X_{it}, Z_{it}))) - m(X_{it}, r(\mathbf{X}_i, h_0(X_{it}, Z_{it},)))| \\ &\leq O\left(\zeta_{1,m}(k_{m,n}) \cdot \zeta_{1,r}(k_{r,n}) \cdot \delta_{h,n}\right) \end{aligned}$$

and that $\zeta_{1,m}(k_{m,n}) \cdot \zeta_{1,r}(k_{r,n}) \cdot \delta_{h,n} = o(1)$ by Assumption 6. Letting $\eta_{1,n} = \zeta_{1,m}(k_{m,n}) \cdot \zeta_{1,r}(k_{r,n}) \cdot \delta_{h,n}$ implies Assumption C.3. In all, all conditions in Proposition B.1 are satisfied, and hence $d_{\Theta.\infty}(\hat{\theta}_n, \theta_0) = o_p(1)$. ■

A.2 Proof of Theorem 6.2

Let $\delta_{h,n}^*$ be the L_2-convergence rate of $\hat{h}_n$. Define $B_{1,K_1,n}(h_0) \equiv \{h \in \mathcal{H} : ||h - h_0||_2 \leq K_1 \delta_{h,n}^*\} \cap \mathcal{H}_n$, $B_{2,K_2,n}(\theta_0) \equiv \{\theta \in \Theta : ||\theta - \theta_0||_{\Theta,2} \leq K_2\} \cap \Theta_n$, and $B_{2,K_2,0}(\theta_0) \equiv \{\theta \in \Theta : ||\theta - \theta_0||_{\Theta,2} \leq K_2\}$.

Proof. Since Assumptions 2–6 are sufficient for Assumptions C.1–C.3, it remains to show that Assumption C.4 holds. For condition (i) in Assumption C.4, observe that

$$
\begin{aligned}
&\sup_{h\in B_{1,K_1,n}(h_0)} |\mathbb{E}[l_u(\mathbf{W},\pi_n\theta_0,h)-l_u(\mathbf{W},\theta_0,h)]|^2\\
\lesssim &\sup_{h\in B_{1,K_1,n}(h_0)} \max(u,1-u)\mathbb{E}\sum_t |\pi_n m_0(X_{it},\pi_n r_0(\mathbf{X}_i,h(X_{it},Z_{it})))-m_0(X_{it},r_0(\mathbf{X}_i,h(X_{it},Z_{it})))|^2\\
\lesssim &O\left(k_{m,n}^{-2(p_m/(d_x+1))}\right)+O\left(k_{r,n}^{-2(p_r/(Td_x+1))}\right)
\end{aligned}
$$

by Assumptions 4 and 5. Thus, one can set $\delta_{2,n}^2 \asymp k_{m,n}^{-2(p_m/(d_x+1))}+k_{r,n}^{-2(p_r/(Td_x+1))}$, and it is obvious that $\delta_{2n}=o(1)$.

For condition (ii) in Assumption C.4, pick any $\theta\in B_{2,K_2,n}(\theta_0)$ and $\delta,\tilde{\delta}>0$ such that $\tilde{\delta}<||\theta-\theta_0||_{\Theta,2}<\delta$. Note that

$$
\mathbb{E}[l_u(\mathbf{W},\theta,h)-l_u(\mathbf{W},\theta_0,h)]=\mathbb{E}[l_u(\mathbf{W},\theta,h)-l_u(\mathbf{W},\theta_0,h_0)+l_u(\mathbf{W},\theta_0,h_0)-l_u(\mathbf{W},\theta_0,h)]
$$

and that

$$
\sup_{h\in B_{1,K_1,n}(h_0)} \mathbb{E}[l_u(\mathbf{W},\theta,h)-l_u(\mathbf{W},\theta_0,h_0)]\leq \sup_{h\in B_{1,K_1,n}(h_0)}\ \sup_{\{\theta\in B_{2,K_2,0}(\theta_0):\tilde{\delta}<||\theta-\theta_0||_{\Theta,2}<\delta\}} \mathbb{E}[l_u(\mathbf{W},\theta,h)-l_u(\mathbf{W},\theta_0,h_0)].
$$

Then, by applying Knight's identity with Assumption 3, there exists a finite constant $\tilde{C}>0$ such that

$$
\sup_{\{\theta\in B_{2,K_2,0}(\theta_0):\tilde{\delta}<||\theta-\theta_0||<\delta\}} \mathbb{E}[l_u(\mathbf{W},\theta,h)-l_u(\mathbf{W},\theta_0,h_0)]\leq -\tilde{C}||\alpha-\alpha_0||^2_{\mathcal{A},2}\leq -\tilde{C}||\theta-\theta_0||^2_{\Theta,2},
$$

where the first inequality holds by Assumption 4 and the second inequality holds by definition of the norms. Since $0<\tilde{\delta}<\delta$, there exists a finite constant $C_\delta>0$ such that $\tilde{\delta}>\delta/C_\delta$. Therefore, $\mathbb{E}[l_u(\mathbf{W},\theta,h)-l_u(\mathbf{W},\theta_0,h_0)]\leq -C\delta^2$ with $C=\tilde{C}/C_\delta>0$. Next, letting $v_{it}=m_0(X_{it},r_0(\mathbf{X}_i,h_0(X_{it},Z_{it})))-m(X_{it},r(\mathbf{X}_i,h(X_{it},Z_{it})))$, it can be shown that by Assumption 3-(ii),

$$
\mathbb{E}[l_u(\mathbf{W},\theta_0,h_0)-l_u(\mathbf{W},\theta_0,h)]\leq \bar{f}\sum_t \mathbb{E}v_{it}^2\lesssim ||h_0-h||_2^2\leq K_1\delta_{h,n}^{*^2}
$$

uniformly over $B_{1,K_1,n}(h_0)$. Therefore,

$$
\sup_{h\in B_{1,K_1,n}(h_0)} \mathbb{E}[l_u(\mathbf{W},\theta,h)-l_u(\mathbf{W},\theta_0,h)]\leq K_1\delta_{h,n}^{*^2}-C\delta^2
$$

for some constant $C>0$, and thus condition (ii) in Assumption C.4 is met with $\delta_{1,n}\asymp\delta^*_{h,n}$.

I consider condition (iii) in Assumption C.4. Define $\tilde{\mathcal{L}}_{u,n}\equiv\{l_u(\mathbf{W},\theta,h)-l_u(\mathbf{W},\theta,h_0):\theta\in B_{2,K_2,n}(\theta_0),h\in B_{1,K_1,n}(h_0)\}$. Then, For any $\theta,\tilde{\theta}\in B_{2,K_2,n}(\theta_0)$ and $h,\tilde{h}\in B_{1,K_1,n}(h_0)$, one obtains that

$$
\begin{aligned}
&|\{l_u(\mathbf{W},\theta,h)-l_u(\mathbf{W},\theta,h_0)\}-\{l_u(\mathbf{W},\tilde{\theta},\tilde{h})-l_u(\mathbf{W},\tilde{\theta},h_0)\}|\\
&\leq |l_u(\mathbf{W},\theta,h)-l_u(\mathbf{W},\tilde{\theta},\tilde{h})|+|l_u(\mathbf{W},\theta,h_0)-l_u(\mathbf{W},\tilde{\theta},h_0)|\\
&\leq |\sum_t \{m(X_{it},r(\mathbf{X}_i,h(X_{it},Z_{it})))-\tilde{m}(X_{it},\tilde{r}(\mathbf{X}_i,\tilde{h}(X_{it},Z_{it})))\}|\\
&+|\sum_t \{m(X_{it},r(\mathbf{X}_i,h_0(X_{it},Z_{it})))-\tilde{m}(X_{it},\tilde{r}(\mathbf{X}_i,h_0(X_{it},Z_{it})))\}|\\
&\lesssim T\cdot\overline{\mathbb{A}(\mathbf{X}_i,Z_{it})}^2\cdot||\alpha-\tilde{\alpha}||^2_{\mathcal{A},E},
\end{aligned}
\tag{22}
$$

where the last inequality holds by Assumption 4-(ii) and the Cauchy-Schwarz inequality. It is straightforward to see that, by Assumptions 5 and 6,

$$
|l_u(\mathbf{W},\theta,h)-l_u(\mathbf{W},\theta,h_0)|\leq |m(X_{it},r(\mathbf{X}_i,h(X_{it},Z_{it})))-m(X_{it},r(\mathbf{X}_i,h_0(X_{it},Z_{it})))|\lesssim F_n(\mathbf{W}),
\tag{23}
$$

where $F_n(\mathbf{W})\equiv C\cdot|\partial_\gamma\phi^{k_{m,n}}(x,\gamma)|\cdot|\partial_p b^{k_{r,n}}(\mathbf{x},p)|\cdot\delta^*_{h,n}$ for some $C>0$ is an envelope function of $\tilde{\mathcal{L}}_{u,n}$, and hence $||F_n(\mathbf{W})||_2\lesssim\zeta_{1,m}(k_{m,n})\cdot\zeta_{1,r}(k_{r,n})\cdot\delta^*_{h,n}$. I make use of Theorem 2.14.2 in Van der Vaart and Wellner (1996) to obtain the convergence rate δ_n in condition (iii) of Assumption C.4. To this end, it remains to calculate $J_{[]}(1,\tilde{\mathcal{L}}_{u,n},||\cdot||_2)$. By equation (22) and Theorem 2.7.11 in Van der Vaart and Wellner (1996), $N_{[]}(\epsilon||F_n||_2,\tilde{\mathcal{L}}_{u,n},||\cdot||_2)\leq N(\frac{\epsilon}{C}||F_n||_2,\mathcal{M}_{n,K_2},||\cdot||_2)\times N(\frac{\epsilon}{C}||F_n||_2,\mathcal{R}_{n,K_2},||\cdot||_2)\times\frac{\epsilon}{C}||F_n||_2,B_{1,K_1n}(h_0),||\cdot||_2)$. By Lemma 2.5 in Van de Geer (2000), one obtains that

$$
\begin{aligned}
&J_{[]}(1,\tilde{\mathcal{L}}_{u,n},||\cdot||_2)\\
&=\int_0^1\sqrt{1+\log N_{[]}(\epsilon||F_n||_2,\tilde{\mathcal{L}}_{u,n},||\cdot||_2)}d\epsilon\\
&\lesssim\int_0^1\sqrt{\log N(\frac{\epsilon}{C}||F_n||_2,\mathcal{M}_{n,K_2},||\cdot||_2)+\log N(\frac{\epsilon}{C}||F_n||_2,\mathcal{R}_{n,K_2},||\cdot||_2)+\log N(\frac{\epsilon}{C}||F_n||_2,B_{1,K_1n}(h_0),||\cdot||_2)}d\epsilon\\
&=\int_0^{||F_n||_2/C}\sqrt{\log N(\epsilon,\mathcal{M}_{n,K_2},||\cdot||_2)+\log N(\epsilon,\mathcal{R}_{n,K_2},||\cdot||_2)+\log N(\epsilon,B_{1,K_1n}(h_0),||\cdot||_2)}d\epsilon\cdot\frac{C}{||F_n||_2}\\
&\lesssim\left\{\sqrt{k_{m,n}}+\sqrt{k_{r,n}}+\sqrt{\delta^*_{h,n}}\right\}.
\end{aligned}
$$

Applying Theorem 2.14.2 in Van der Vaart and Wellner (1996) yields that

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{\theta\in B_{2,K_2,n}(\theta_0),h\in B_{1,K_1,n}(h_0)}|\mu_n[l_u(\mathbf{W},\theta,h)-l_u(\mathbf{W},\theta,h_0)]|\right]\\
&\lesssim O\left(\left(\sqrt{k_{m,n}}+\sqrt{k_{r,n}}+1\right)\cdot\frac{1}{\sqrt{n}}||F_n||_2\right)\\
&=O\left(\left(\sqrt{k_{m,n}}+\sqrt{k_{r,n}}\right)\cdot\frac{1}{\sqrt{n}}\cdot\zeta_{1,m}(k_{m,n})\cdot\zeta_{1,r}(k_{r,n})\cdot\delta^*_{h,n})\right)\\
&\leq O\left(\left(\left(\sqrt{k_{m,n}}+\sqrt{k_{r,n}}\right)\cdot\frac{1}{\sqrt{n}}\right)^2+\left(\zeta_{1,m}(k_{m,n})\cdot\zeta_{1,r}(k_{r,n})\cdot\delta^*_{h,n}\right)^2\right),
\end{aligned}
$$

where the last inequality holds by the fact that $ab\leq(a+b)^2$ for any $a,b>0$. Therefore, one can set $\delta_n^2 \asymp \left((\sqrt{k_{m,n}} + \sqrt{k_{r,n}}) \cdot \frac{1}{\sqrt{n}} \right)^2 + \left(\zeta_{1,m}(k_{m,n}) \cdot \zeta_{1,r}(k_{r,n}) \cdot \delta_{h,n}^* \right)^2$.

Lastly, I verify condition (iv) in Assumption C.4. I make use of Lemma 19.36 in Van der Vaart (2000). Let $\tilde{B}_{2,K_2,n}^{\delta}(\theta_0) \equiv \{\theta \in B_{2,K_2,n}(\theta_0) : ||\theta - \theta_0||_{\Theta,2} \leq \delta\}$ and $\tilde{B}_{2,K_2,0}^{\delta}(\theta_0) \equiv \{\theta \in B_{2,K_2,0}(\theta_0) : ||\theta - \theta_0||_{\Theta,2} \leq \delta\}$. Define $\mathcal{F}_n \equiv \{l_u(\mathbf{W}, \theta, h_0) - l_u(\mathbf{W}, \theta_0, h_0) : \theta \in \tilde{B}_{2,K_2,0}^{\delta}(\theta_0)\}$. Then,

$$
\begin{aligned}
& \sup_{\theta \in \tilde{B}_{2,K_2,n}^{\delta}(\theta_0)} \mathbb{E}\left[|l_u(\mathbf{W}, \theta, h_0) - l_u(\mathbf{W}, \theta_0, h_0)|^2 \right] \\
\leq & \sup_{\theta \in \tilde{B}_{2,K_2,0}^{\delta}(\theta_0)} \mathbb{E}\left[|l_u(\mathbf{W}, \theta, h_0) - l_u(\mathbf{W}, \theta_0, h_0)|^2 \right] \\
\lesssim & \sup_{\theta \in \tilde{B}_{2,K_2,0}^{\delta}(\theta_0)} ||\theta - \theta_0||_{\Theta,2}^2 \leq \delta^2 \qquad (24)
\end{aligned}
$$

under Assumption 4-(ii). Furthermore, Assumption 4-(i) implies that $\sup_{\omega \in \mathcal{W}, \theta \in \Theta} |l_u(w, \theta, h_0) - l_u(w, \theta_0, h_0)| < \infty$. It remains to calculate $J_{[]}(C\delta, \mathcal{F}_n, || \cdot ||_2)$. Equation (24) enables to apply Theorem 2.7.11 in Van der Vaart and Wellner (1996), Let $B_{\mathcal{M},K_2,n}^{\delta}(m_0) \equiv \{m \in \mathcal{M}_n : ||m - m_0||_2 \leq \min(\delta, K_2)\}$ and $B_{\mathcal{R},K_2,n}^{\delta}(r_0) \equiv \{r \in \mathcal{R}_n : ||r - r_0||_2 \leq \min(\delta, K_2)\}$. Then, one can show that $N(C\delta, \tilde{B}_{2,K_2,n}^{\delta}(\theta_0), || \cdot ||_{\Theta,2}) \leq N(\frac{C}{2}\delta, B_{\mathcal{M},K_2,n}^{\delta}(m_0), || \cdot ||_2) \cdot N(\frac{C}{2}\delta, B_{\mathcal{R},K_2,n}^{\delta}(r_0), || \cdot ||_2)$. By, together with Assumption 5, Lemma 2.5 in Van de Geer (2000), one can show that $N(\frac{C}{2}\delta, B_{\mathcal{M},K_2,n}^{\delta}(m_0), || \cdot ||_2) \lesssim k_{m,n} \cdot \log(1 + \frac{Cc_m}{\delta})$ and $N(\frac{C}{2}\delta, B_{\mathcal{R},K_2,n}^{\delta}(r_0), || \cdot ||_2) \lesssim k_{r,n} \cdot \log(1 + \frac{Cc_r}{\delta})$ for some finite constant $C > 0$. Therefore,

$$
J_{[]}(C\delta, \mathcal{F}_n, || \cdot ||_2) = \int_0^{C\delta} \sqrt{1 + \log N_{[]}(||\bar{l}_u||_2 \epsilon, \mathcal{F}_n, || \cdot ||_2} d\epsilon \lesssim \sqrt{\max(k_{m,n}, k_{r,n})}\delta
$$

by the same logic as earlier. Applying Lemma 19.36 in Van der Vaart (2000) leads to that

$$
\mathbb{E}\left[\sup_{\theta \in \tilde{B}_{2,K_2,n}^{\delta}(\theta_0)} |\mu_n\left[l_u(\mathbf{W}, \theta, h_0) - l_u(\mathbf{W}, \theta_0, h_0) \right]| \right] \lesssim \frac{1}{\sqrt{n}} J_{[]}(C\delta, \mathcal{F}_n, ||\cdot||_2) \lesssim \frac{1}{\sqrt{n}} \sqrt{\max(k_{m,n}, k_{r,n})}\delta,
$$

and therefore one can set $\phi_n(\delta) = \sqrt{\max(k_{m,n}, k_{r,n})}\delta$. It is straightforward to see that the map $\delta \mapsto \delta^{-(1+\epsilon)}\phi_n(\delta)$ is decreasing for any $\epsilon \in (0,1)$. Then, the condition $(\delta_{\theta,n})^{-2}\sqrt{\max(k_{m,n}, k_{r,n})}\delta_{\theta,n} \lesssim \sqrt{n}$ holds if one chooses $\delta_{\theta,n} \equiv \sqrt{\frac{\max(k_{m,n}, k_{r,n})}{n}}$, and Assumption 5 implies that $\delta_{\theta,n} = o(1)$.

In all, by the fact that $\delta_{h,n}^* \leq \zeta_{1,m}(k_{m,n}) \cdot \zeta_{1,r}(k_{r,n}) \cdot \delta_{h,n}^*$ and Proposition B.2, one obtains that

$$
\delta_{\theta,n}^* = \sqrt{\frac{k_{m,n}}{n}} + k_{m,n}^{-(p_m/(d_x+1))} + \sqrt{\frac{k_{r,n}}{n}} + k_{r,n}^{-(p_r/(Td_x+1))} + \zeta_{1,m}(k_{m,n}) \cdot \zeta_{1,r}(k_{r,n}) \cdot \delta_{h,n}^*,
$$

and this ends the proof. ∎

A.3 Proof of Theorem 6.4

To prove Theorem 6.4, it is required to verify Assumption C.5 in Section B. To this end, I present several lemmas.

Lemma A.1. *Suppose that the conditionsn in Theorem 6.4 hold. Then, equations (37) and (38) in Assumption C.5 hold.*

Proof. By the definition of g^* and using Knight's identity, one obtains that

$$l_u(\mathbf{W}, g^*, h) - l_u(\mathbf{W}, g, h) - \Delta_2(\mathbf{W}, g, h)[g^* - g] = R_2(\mathbf{W}, g^*, g, h),$$

where $\Delta_2(\mathbf{W}, g, h)[g^* - g] = \Delta_2(\mathbf{W}, g, h)[\pm\kappa_n \cdot u^*_{g_n}]$ and $R_2(\mathbf{W}, g^*, g, h) \equiv -\sum_t \int_0^{g^*-g} \{\mathbf{1}(Y_t \leq g(\mathbf{X}, h(X_t, Z_t)) + s) - \mathbf{1}(Y_t \leq g(\mathbf{X}, h(X_t, Z_t)))\}ds$. By definition, $g^* - g = \pm\kappa_n u^*_{g_n}$. Without loss of generality, I assume that $g^* - g = \kappa_n u^*_{g_n}$. Then, by the change of variable, $s = \kappa_n q$,

$$R_2(\mathbf{W}, g^*, g, h) = -\kappa_n \sum_t \int_0^{u^*_{g_n}} \{\mathbf{1}(Y_t \leq g(\mathbf{X}, h(X_t, Z_t)) + \kappa_n q) - \mathbf{1}(Y_t \leq g(\mathbf{X}, h(X_t, Z_t)))\}dq$$

Therefore, (37) will hold if

$$\sup_{\alpha \in \mathcal{N}_{\alpha,n}} |\mu_n[\tilde{R}_2(\mathbf{W}, g, h)]| = o_p\left(n^{-1/2}\right)$$

where t $\tilde{R}_2(\mathbf{W}, g, h) = \sum_t \int_0^{u^*_{g_n}} \{\mathbf{1}(Y_t \leq g(\mathbf{X}, h(X_t, Z_t)) + \kappa_n q) - \mathbf{1}(Y_t \leq g(\mathbf{X}, h(X_t, Z_t)))\}dq$. To verify the condition, I apply Lemma 4.2 in Chen (2007). To this end, note that for any small $\delta > 0$ and for any $\alpha \in \mathcal{N}_\alpha$,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\tilde{\alpha} \in \mathcal{N}_{\alpha,n}, ||\alpha - \tilde{\alpha}||_{\mathcal{A},\infty} \leq \delta} |\tilde{R}_2(\mathbf{W}, g, h) - \tilde{R}_2(\mathbf{W}, \tilde{g}, \tilde{h})|^2\right] \\
\leq &\mathbb{E}\left[\sup_{\tilde{\alpha} \in \mathcal{N}_{\alpha,n}, ||\alpha - \tilde{\alpha}||_{\mathcal{A},\infty} \leq \delta} |R_2(\mathbf{W}, g^*, g, h) - R_2(\mathbf{W}, \tilde{g}^*, \tilde{g}, \tilde{h})|^2\right] \\
\lesssim &\mathbb{E}\left[\sup_{\tilde{\alpha} \in \mathcal{N}_{\alpha,n}, ||\alpha - \tilde{\alpha}||_{\mathcal{A},\infty} \leq \delta} \sum_t \int_0^{u^*_{g_n}} |\mathbf{1}(Y_t \leq g(\mathbf{X}, h(X_t, Z_t)) + \kappa_n s) - \mathbf{1}(Y_t \leq \tilde{g}(\mathbf{X}, \tilde{h}(X_t, Z_t)) + \kappa_n s)|ds\right]
\end{aligned}$$

Note that by Assumption 4, for any $\mathbf{x} \in Supp(\mathbf{X})$ and $z_t \in Supp(Z_t)$,

$$-\underline{\mathbb{A}(\mathbf{x}, z_t)} \cdot \delta \leq \tilde{g}(\mathbf{x}, \tilde{h}(x_t, z_t)) - g(\mathbf{x}, h(x_t, z_t)) \leq \overline{\mathbb{A}(\mathbf{x}, z_t)} \cdot \delta$$

and that

$$g(\mathbf{x}, h(x_t, z_t)) + \overline{\mathbb{A}(\mathbf{x}, z_t)} \cdot \delta \geq g(\mathbf{x}, h(x_t, z_t)) \geq g(\mathbf{x}, h(x_t, z_t)) - \underline{\mathbb{A}(\mathbf{x}, z_t)} \cdot \delta.$$

Therefore,

$$\begin{aligned}
&|\mathbf{1}(Y_t \leq g(\mathbf{X}, h(X_t, Z_t)) + \kappa_n s) - \mathbf{1}(Y_t \leq \tilde{g}(\mathbf{X}, \tilde{h}(X_t, Z_t)) + \kappa_n s)| \\
&\leq \mathbf{1}(Y_t \leq g(\mathbf{X}, h(X_t, Z_t)) + \kappa_n s + \overline{\mathbb{A}(\mathbf{X}, Z_t)} \cdot \delta) - \mathbf{1}(Y_t \leq g(\mathbf{X}, h(X_t, Z_t)) + \kappa_n s - \underline{\mathbb{A}(\mathbf{X}, Z_t)} \cdot \delta).
\end{aligned}$$

This leads to that

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{\tilde{\alpha}\in\mathcal{N}_{\alpha,n},||\alpha-\tilde{\alpha}||_{\mathcal{A},\infty}\leq\delta}\sum_t\int_0^{u^*_{g_n}}|\mathbf{1}(Y_t\leq g(\mathbf{X},h(X_t,Z_t))+\kappa_n s)-\mathbf{1}(Y_t\leq\tilde{g}(\mathbf{X},\tilde{h}(X_t,Z_t))+\kappa_n s)|ds\right]\\
\leq&\mathbb{E}\left[\sup_{\tilde{\alpha}\in\mathcal{N}_{\alpha,n},||\alpha-\tilde{\alpha}||_{\mathcal{A},\infty}\leq\delta}\sum_t\mathbf{1}(u^*_{g_n}\geq 0)\int_0^{u^*_{g_n}}|\mathbf{1}(Y_t\leq g(\mathbf{X},h(X_t,Z_t))+\kappa_n s)-\mathbf{1}(Y_t\leq\tilde{g}(\mathbf{X},\tilde{h}(X_t,Z_t))+\kappa_n s)|ds\right]\\
&+\mathbb{E}\left[\sup_{\tilde{\alpha}\in\mathcal{N}_{\alpha,n},||\alpha-\tilde{\alpha}||_{\mathcal{A},\infty}\leq\delta}\sum_t\mathbf{1}(u^*_{g_n}<0)\int_{u^*_{g_n}}^{0}|\mathbf{1}(Y_t\leq g(\mathbf{X},h(X_t,Z_t))+\kappa_n s)-\mathbf{1}(Y_t\leq\tilde{g}(\mathbf{X},\tilde{h}(X_t,Z_t))+\kappa_n s)|ds\right]\\
\lesssim&\mathbb{E}\left[\sum_t\int_0^{u^*_{g_n}}F_{Y_t|\mathbf{X},Z_t}(g(\mathbf{X},h(X_t,Z_t))+\kappa_n s+\overline{\mathbb{A}(\mathbf{X},Z_t)}\cdot\delta))-F_{Y_t|\mathbf{X},Z_t}(g(\mathbf{X},h(X_t,Z_t))+\kappa_n s-\underline{\mathbb{A}(\mathbf{X},Z_t)}\cdot\delta)ds\right]\\
\lesssim&\delta.
\end{aligned}
$$

Therefore, condition (4.2.1) in Lemma 4.2 in Chen (2007) holds with $s=1/2$ in its notation. By Assumptions 4 and 6 and Theorem 2.7.1 in Van der Vaart and Wellner (1996), $\log N(C\epsilon^2,\mathcal{G},||\cdot||_\infty)\cdot\log N(C\epsilon^2,\mathcal{H},||\cdot||_\infty)<\infty$. In all, applying Lemma 4.2 in Chen (2007) results in that $\sup_{\alpha\in\mathcal{N}_{\alpha,n}}|\mu_n[\tilde{R}_2(\mathbf{W},g,h)]|=o_p\left(n^{-1/2}\right)$. This implies that (37) holds.

Recall that $\Delta_2(\mathbf{W},g,h)[u^*_{g_n}]=\sum_t\{u-\mathbf{1}(Y_t\leq g(\mathbf{X},h(X_t,Z_t))\}u^*_{g_n}$, and therefore

$$
\Delta_2(\mathbf{W},g,h)[u^*_{g_n}]-\Delta_2(\mathbf{W},g_0,h_0)[u^*_{g_n}]=\sum_t\left\{\mathbf{1}(Y_t\leq g_0(\mathbf{X},h_0(X_t,Z_t)))-\mathbf{1}(Y_t\leq g(\mathbf{X},h(X_t,Z_t)))\right\}u^*_{g_n}.
$$

By the same logic above, it follows that

$$
\sup_{\alpha\in\mathcal{N}_{\alpha,n}}|\mu_n\left[\Delta_2(\mathbf{W},g,h)[u^*_{g_n}]-\Delta_2(\mathbf{W},g_0,h_0)[u^*_{g_n}]\right]|=o_p\left(n^{-1/2}\right),
$$

which implies that it is $O_p(\kappa_n)$, and thus (38) also holds. ■

Lemma A.2. *Suppose that all conditions in Theorem 6.4 hold. Then, condition (ii) of Assumption C.5 is satisfied.*

Proof. By (17) and Assumption 10, one can show that

$$
\begin{aligned}
\mathbb{K}(g,h)&=-\mathbb{E}\left[\int_0^{g(\mathbf{X},Z_t;h)-g_0(\mathbf{X},Z_t;h_0)}F_{Y^*_t|\mathbf{X},Z_t}(g_0(\mathbf{X},Z_t;h_0)+s)-F_{Y^*_t|\mathbf{X},Z_t}(g_0(\mathbf{X},Z_t;h_0))ds\right]\\
&=-\mathbb{E}\left[\int_0^{g(\mathbf{X},Z_t;h)-g_0(\mathbf{X},Z_t;h_0)}f_{Y^*_t|\mathbf{X},Z_t}(g_0(\mathbf{X},Z_t;h_0))s+f'_{Y^*_t|\mathbf{X},Z_t}(\tilde{g}(\mathbf{X},Z_t;\tilde{h}))s^2ds\right]\\
&=-\mathbb{E}\left[\frac{f_{Y^*_t|\mathbf{X},Z_t}(g_0(\mathbf{X},Z_t;h_0))}{2}\left(g(\mathbf{X},Z_t;h)-g_0(\mathbf{X},Z_t;h_0)\right)^2\right]-\mathbb{E}[R_{\mathbb{K},1}(\mathbf{X},Z_t)],
\end{aligned}
$$

where $R_{\mathbb{K},1}(\mathbf{X},Z_t)\equiv\int_0^{g(\mathbf{X},Z_t;h)-g_0(\mathbf{X},Z_t;h_0)}f'_{Y^*_t|\mathbf{X},Z_t}(\tilde{g}(\mathbf{X},Z_t;h_0))s^2ds$ and $\tilde{g}(\mathbf{X},Z_t;\tilde{h})$ is between $g(\mathbf{X},Z_t;h)$ and $g_0(\mathbf{X},Z_t;h_0)$. Similarly,

$$
\mathbb{K}(g^*,h)=-\mathbb{E}\left[\frac{f_{Y^*_t|\mathbf{X},Z_t}(g_0(\mathbf{X},Z_t;h_0))}{2}\left(g^*(\mathbf{X},Z_t;h)-g_0(\mathbf{X},Z_t;h_0)\right)^2\right]-\mathbb{E}[R_{\mathbb{K},2}(\mathbf{X},Z_t)]
$$

with $R_{\mathbb{K},2}(\mathbf{X},Z_t)\equiv\int_0^{g^*(\mathbf{X},Z_t;h)-g_0(\mathbf{X},Z_t;h_0)}f'_{Y^*_t|\mathbf{X},Z_t}(\tilde{g}^*(\mathbf{X},Z_t;\tilde{h}^*))s^2ds$ for some $\tilde{g}^*$ and $\tilde{h}^*$. By the definition of g^*, one obtains that

$$\begin{aligned}\mathbb{K}(g,h)-\mathbb{K}(g^*,h) &= \frac{1}{2}\left\{||g^*-g_0||^2_{\Delta,2}-||g-g_0||^2_{\Delta,2}\right\}\\&+\mathbb{E}\left[\pm\kappa_n u^*_{g_n} f_{Y_t^*|\mathbf{X},Z_t}(g_0(\mathbf{X},Z_t;h_0))\left(g(\mathbf{X},Z_t;h)-g(\mathbf{X},Z_t;h_0)\right)\right]\\&+\mathbb{E}[R_{\mathbb{K},2}(\mathbf{X},Z_t)]-\mathbb{E}[R_{\mathbb{K},1}(\mathbf{X},Z_t)].\end{aligned}$$

Observe that

$$\begin{aligned}g(\mathbf{X},Z_t;h)-g(\mathbf{X},Z_t;h_0) &= \partial_\gamma m(X_t,r(\mathbf{X},h_0(X_t,Z_t)))\cdot\partial_p r(\mathbf{X},h_0(X_t,Z_t))\cdot(h-h_0)\\&+\Bigg\{\partial^2_\gamma m(X_t,r(\mathbf{X},\bar{h}(X_t,Z_t)))\cdot\left(\partial_p r(\mathbf{X},\bar{h}(X_t,Z_t))\right)^2\\&+\partial_\gamma m(X_t,r(\mathbf{X},\bar{h}(X_t,Z_t)))\cdot\partial^2_p r(\mathbf{X},\bar{h}(X_t,Z_t))\Bigg\}\cdot(h-h_0)^2,\end{aligned}$$

where $\bar{h}$ is between h and h_0, and thus

$$\begin{aligned}\mathbb{K}(g,h)-\mathbb{K}(g^*,h) &= \frac{1}{2}\left\{||g^*-g_0||^2_{\Delta,2}-||g-g_0||^2_{\Delta,2}\right\}\\&+\mathbb{E}\left[\pm\kappa_n u^*_{g_n} f_{Y_t^*|\mathbf{X},Z_t}(g_0(\mathbf{X},Z_t;h_0))\partial_\gamma m(X_t,r(\mathbf{X},h_0(X_t,Z_t)))\cdot\partial_p r(\mathbf{X},h_0(X_t,Z_t))\cdot(h-h_0)\right]\\&+\mathbb{E}[R_{\mathbb{K},2}(\mathbf{X},Z_t)-R_{\mathbb{K},1}(\mathbf{X},Z_t)\pm\kappa_n R_{\mathbb{K},3}(\mathbf{X},Z_t)]\end{aligned}$$

with

$$\begin{aligned}R_{\mathbb{K},3}(\mathbf{X},Z_t) &\equiv \left\{\partial^2_\gamma m(X_t,r(\mathbf{X},\tilde{h}(X_t,Z_t)))\cdot\Big(\partial_p r(\mathbf{X},\tilde{h}(X_t,Z_t))\Big)^2+\partial_\gamma m(X_t,r(\mathbf{X},\tilde{h}(X_t,Z_t)))\cdot\partial^2_p r(\mathbf{X},\tilde{h}(X_t,Z_t))\right\}\\&\times(h-h_0)^2\cdot u^*_{g_n} f_{Y_t^*|\mathbf{X},Z_t}(g_0(\mathbf{X},Z_t;h_0)|\mathbf{X},Z_t).\end{aligned}$$

By definition of $\Gamma(\alpha_0)[v_h,v_g]$ in (20), it follows that

$$\begin{aligned}\mathbb{K}(g,h)-\mathbb{K}(g^*,h) &= \mp\kappa_n\Gamma(\alpha_0)[(h-h_0),u^*_{g_n}]+\frac{1}{2}\left\{||g^*-g_0||^2_{\Delta,2}-||g-g_0||^2_{\Delta,2}\right\}\\&+\mathbb{E}[R_{\mathbb{K},2}(\mathbf{X},Z_t)-R_{\mathbb{K},1}(\mathbf{X},Z_t)\pm\kappa_n R_{\mathbb{K},3}(\mathbf{X},Z_t)]+\mathbb{E}[\pm\kappa_n\cdot R_{\mathbb{K},4}(\mathbf{X},Z_t)]\end{aligned}$$

with

$$\begin{aligned}R_{\mathbb{K},4}(\mathbf{X},Z_t) &= \{\partial_\gamma m(X_t,r(\mathbf{X},h_0(X_t,Z_t)))\cdot\partial_p r(\mathbf{X},h_0(X_t,Z_t))-\partial_\gamma m_0(X_t,r_0(\mathbf{X},h_0(X_t,Z_t)))\cdot\partial_p r_0(\mathbf{X},h_0(X_t,Z_t))\}\\&\times f_{Y_t^*|\mathbf{X},Z_t}(g_0(\mathbf{X},Z_t;h_0))\cdot(h-h_0).\end{aligned}$$

I first consider $R_{\mathbb{K},1}(\mathbf{X}, Z_t)$ and $R_{\mathbb{K},2}(\mathbf{X}, Z_t)$.

$$\begin{aligned}
\mathbb{E}[|R_{\mathbb{K},1}(\mathbf{X}, Z_t)|] &\leq \mathbb{E}\left[\int_0^{g(\mathbf{X},Z_t;h)-g_0(\mathbf{X},Z_t;h_0)} |f'_{Y_t^*|\mathbf{X},Z_t}(\tilde{g}(\mathbf{X}, Z_t; h_0))s^2|ds\right] \\
&\lesssim \mathbb{E}\left[|g(\mathbf{X}, Z_t; h) - g_0(\mathbf{X}, Z_t; h_0)|^3\right] \\
&\lesssim \mathbb{E}\left[|g(\mathbf{X}, Z_t; h) - g_0(\mathbf{X}, Z_t; h_0)|^2\right] \cdot \sup_{\mathbf{x},z_t} |g(\mathbf{x}, z_t; h) - g_0(\mathbf{x}, z_t; h_0)|,
\end{aligned}$$

where the second inequality holds by Assumption 10-(i). Then,

$$\begin{aligned}
\mathbb{E}\left[|g(\mathbf{X}, Z_t; h) - g_0(\mathbf{X}, Z_t; h_0)|^2\right] &\lesssim \mathbb{E}\left[|g(\mathbf{X}, Z_t; h) - g(\mathbf{X}, Z_t; h_0)|^2\right] + \mathbb{E}\left[|g(\mathbf{X}, Z_t; h_0) - g_0(\mathbf{X}, Z_t; h_0)|^2\right] \\
&\lesssim \zeta_{1,m}(k_{m,n})^2 \cdot \zeta_{1,r}(k_{r,n})^2 \cdot \delta_{h,n}^2 + \delta_{\theta,n}^2.
\end{aligned}$$

It follows that

$$\sup |g(\mathbf{x}, z_t; h) - g(\mathbf{x}, z_t; h_0)| \lesssim \zeta_{1,m}(k_{m,n}) \cdot \zeta_{1,r}(k_{r,n}) \cdot \delta_{h,n}^{\sup},$$

and that

$$\sup |g(\mathbf{x}, z_t; h_0) - g_0(\mathbf{x}, z_t; h_0)| \lesssim (\zeta_{0,m}(k_{m,n}) + \zeta_{0,r}(k_{r,n})) \cdot \delta_{\theta,n}$$

by Assumptions 4 and 5. Therefore, one obtains that $\mathbb{E}[|R_{\mathbb{K},1}(\mathbf{X}, Z_t)|] = o(n^{-1})$ by Assumption 11. Similarly,

$$\begin{aligned}
\mathbb{E}[|R_{\mathbb{K},2}(\mathbf{X}, \mathbf{Z})|] &\leq \sum_t \int_0^{g^*(\mathbf{X},Z_t;h)-g_0(\mathbf{X},Z_t;h_0)} |f'_{Y_t^*|\mathbf{X},Z_t}(\tilde{g}^*(\mathbf{X}, Z_t; \tilde{h}^*))s^2|ds \\
&\lesssim \sum_t \mathbb{E}\left[|g^*(\mathbf{X}, Z_t; h) - g_0(\mathbf{X}, Z_t; h_0)|^3\right] = o(n^{-1}).
\end{aligned}$$

Now I consider $\mathbb{E}[|R_{\mathbb{K},3}(\mathbf{X}, Z_t)|]$. Note that

$$\mathbb{E}[|R_{\mathbb{K},3}(\mathbf{X}, Z_t)|] \lesssim \left\{\zeta_{2,m}(k_{m,n}) \cdot \zeta_{1,r}(k_{r,n})^2 + \zeta_{1,m}(k_{m,n}) \cdot \zeta_{2,r}(k_{r,n})\right\} \delta_{h,n}^2 = o(n^{-1/2})$$

by Assumption 11. Lastly, consider $\mathbb{E}[|R_{\mathbb{K},4}(\mathbf{X}, Z_t)|]$. By the Cauchy-Schwarz inequality,

$$\mathbb{E}[|R_{\mathbb{K},4}(\mathbf{X}, Z_t)|] \lesssim \sqrt{\mathbb{E}\left[G(\mathbf{X}, Z_t)^2\right] \cdot \mathbb{E}\left[(h - h_0)^2\right]}$$

where $G(\mathbf{X}, Z_t) \equiv \partial_\gamma m(X_t, r(\mathbf{X}, h_0(X_t, Z_t))) \cdot \partial_p r(\mathbf{X}, h_0(X_t, Z_t)) - \partial_\gamma m_0(X_t, r_0(\mathbf{X}, h_0(X_t, Z_t))) \cdot \partial_p r_0(\mathbf{X}, h_0(X_t, Z_t))$. Then, it follows that

$$\begin{aligned}
\mathbb{E}\left[G(\mathbf{X}, Z_t)^2\right] &\lesssim \mathbb{E}\left[\partial_p r(\mathbf{X}, h_0(X_t, Z_t))^2 \cdot \{\partial_\gamma m(X_t, r(\mathbf{X}, h_0(X_t, Z_t))) - \partial m_0(X_t, r(\mathbf{X}, h_0(X_t, Z_t)))\}^2\right] \\
&\quad + \mathbb{E}\left[\{\partial_p r(\mathbf{X}, h_0(X_t, Z_t)) - \partial_p r_0(\mathbf{X}, h_0(X_t, Z_t))\}^2\right] + \mathbb{E}\left[\{r(\mathbf{X}, h_0(X_t, Z_t)) - r_0(\mathbf{X}, h_0(X_t, Z_t))\}^2\right] \\
&\lesssim \zeta_{1,r}(k_{r,n})^2 \delta_{\theta,n}^2 + \delta_{\theta,n}^2 + O\left(k_{r,n}^{-2(p_r-1)/(Td_x+1)}\right) + \delta_{\theta,n}^2
\end{aligned}$$

where the first inequality holds by Assumption 10 and the second inequality holds by Corollary 3.1

in Chen and Christensen (2018). Therefore,

$$\mathbb{E}[|R_{\mathbb{K},4}(\mathbf{X}, Z_t)|] \lesssim \zeta_{1,r}(k_{r,n})\delta_{\theta,n}\delta_{h,n} = o(n^{-1/2})$$

by Assumption 11. In all, it follows that

$$\mathbb{K}(g,h) - \mathbb{K}(g^*,h) = \mp\kappa_n\Gamma(\beta_0)[(h-h_0), u^*_{g_n}] + \frac{1}{2}\left\{||g^* - g_0||^2_{\Delta,2} - ||g-g_0||^2_{\Delta,2}\right\} + o(n^{-1}),$$

and thus Assumption C.5-(ii) holds. ■

Proof of Thoerem 6.4

Proof. Assumption 7 restricts attention to the class of regular functionals. Note that for any $v_h \in \mathbb{V}_1$ and $v_g \in \mathbb{V}_2$, $||v_h||_{1,\Delta} \lesssim ||v_h||_2$ and $||v_g||_{2,\Delta} \lesssim ||v_g||_2$ under Assumption 3-(ii). Assumption 8, together with Assumption 3-(ii), corresponds to Assumption 3.1 in Hahn et al. (2018a). Assumption 9 imply Assumption 3.3 in Hahn et al. (2018a). It remains to show that

$$\mp\kappa_n\Gamma(\alpha_0)[(h-h_0), u^*_{g_n}] = \mp\kappa_n\Gamma(\alpha_0)[(h-h_{0,n}), u^*_{g_n}] + o(n^{-1}).$$

Observe that

$$\Gamma(\alpha_0)[(h-h_0), u^*_{g_n}] = \Gamma(\alpha_0)[(h-h_{0,n}), u^*_{g_n}] + \Gamma(\alpha_0)[(h_{0,n}-h_0), u^*_{g_n}]$$

and that

$$\begin{aligned}\left|\Gamma(\alpha_0)[(h_{0,n}-h_0), u^*_{g_n}]\right| &= \Big|\sum_t \mathbb{E}[f_{Y_t^*|\mathbf{X},Z_t}(m_0(X_{it}, r_0(\mathbf{X}_i, h_0(\mathbf{X}_i Z_{it})))) \cdot \partial_\gamma m_0(X_{it}, r_0(\mathbf{X}_i, h_0(\mathbf{X}_i Z_{it}))) \\ &\quad \times \partial_p r_0(\mathbf{X}_i, h_0(X_{it}Z_{it})) \cdot u^*_{g_n} \cdot (h_{0,n}-h_0)]\Big| \\ &\lesssim \sup|h_{0,n}-h_0| = o(n^{-1/2}).\end{aligned}$$

Therefore, $\mp\kappa_n\Gamma(\alpha_0)[(h_{0,n}-h_0), u^*_{g_n}] = o(n^{-1})$. This, together with Lemmas A.1 and A.2, implies Assumption C.5, and the result follows by Proposition B.3. ■

B Asymptotic Results in Hahn et al. (2018a)

In this section, I restate the asymptotic results in Hahn et al. (2018a) and Hahn et al. (2018b) to make the proofs of the main results in this paper more concrete.

B.1 Consistency

I first present consistency and rate results in Hahn et al. (2018b). I slightly modify some parts of assumptions if necessary, but such modifications do not affect the proof strategies in Hahn et al. (2018b).

Assumption C.1. *(i)* $\mathbb{E}[l_u(\mathbf{W}_i, \alpha_0)] > -\infty$; *(ii) for all* $\epsilon > 0$, *there exists some non-increasing positive sequence* $c_n(\epsilon)$ *such that for all* $n \geq 1$,

$$\mathbb{E}[l_u(\mathbf{W}_i, \alpha_0)] - \sup_{\{\theta \in \Theta_n : d(\theta, \theta_0) \geq \epsilon\}} \mathbb{E}[l_u(\mathbf{W}_i, \theta, h_0)] \geq c_n(\epsilon)$$

with $\liminf_n c_n(\epsilon) > 0$.

Assumption C.2. *(i)* $\theta_0 \in \Theta$ *and* $d(\cdot, \cdot) : \Theta \times \Theta \to \mathbb{R}_+$ *is a (pseudo-) metric; (ii) for all* $n \geq 1$, $\Theta_n \subseteq \Theta_{n+1} \subseteq \Theta$ *and (iii) there exists* $\pi_n \theta_0 \in \Theta_n$ *such that*

$$|\mathbb{E}[l_u(\mathbf{W}_i, \pi_n \theta_0, h_0) - l_u(\mathbf{W}_i, \theta_0, h_0)]| = O(\eta_{2,n})$$

for some finite positive non-increasing sequence $\eta_{2,n}$ *with* $\eta_{2,n} \downarrow 0$.

Assumption C.3. *(i) There exists some finite positive non-increasing sequence* $\eta_{0,n}$ *such that* $\eta_{0,n} \downarrow 0$ *and* $\sup_{\theta \in \Theta_n, h \in B_n(h_0)} |\mathbb{E}_n[l_u(\mathbf{W}, \theta, h)]| = O_p(\eta_{0,n})$; *(ii) there is a finite positive sequence* $\eta_{1,n}$ *such that* $\eta_{1,n} \downarrow 0$ *and*

$$\sup_{\theta \in \Theta_n, h \in B_n(h_0)} |\mathbb{E}[l_u(\mathbf{W}_i, \theta, h) - l_u(\mathbf{W}_i, \theta, h_0)]| = O(\eta_{1,n}).$$

Proposition B.1 (Consistency). *Let* $u \in \mathcal{U}$ *be given. Suppose that Assumptions C.1, C.2 and C.3 hold. Then,* $d(\hat{\theta}_n, \theta_0) = o_p(1)$.

Proof. See the proof of Theorem 5.1 in Hahn et al. (2018b). ■

B.2 Rate of Convergence

I present a rate result which is similar to that of Hahn et al. (2018b), and the result is specialized in the L_2-convergence rate. Let $B_{1,K_1,n}(h_0) \equiv \{h \in \mathcal{H} : ||h - h_0||_2 \leq K_1 \delta^*_{h,n}\} \cap \mathcal{H}_n$, $B_{2,K_2,n}(\theta_0) \equiv \{\theta \in \Theta : ||\theta - \theta_0||_2 \leq K_2\} \cap \Theta_n$, and $B_{2,K_2,0}(\theta_0) \equiv \{\theta \in \Theta : ||\theta - \theta_0||_2 \leq K_2\}$. To establish the convergence rate of the two-step estimator, the following assumption is additionally required.

Assumption C.4. *The following conditions hold:*

(i) There exists some positive non-increasing sequence $\delta^2_{2,n}$ *such that*

$$\sup_{h \in B_{K_1}(h_0)} |\mathbb{E}[l_u(\mathbf{W}, \pi_n \theta_0, h) - l_u(\mathbf{W}, \theta_0, h)]| = O(\delta^2_{2,n})$$

for some constant $K_1 > 0$;

(ii) for any small constant $\delta, \tilde{\delta} > 0$ *and for any* $\theta \in B_{K_2}(\theta_0)$ *with* $\tilde{\delta} < d_\Theta(\theta, \theta_0) < \delta$, *there exist some positive non-increasing sequence* $\delta_{1,n}$ *and finite constants* $c_{K_1,1}, c_{K_1,2} > 0$ *such that*

$$\sup_{h \in B_{1,K_1,n}(h_0)} \mathbb{E}[l_u(\mathbf{W}, \theta, h) - l_u(\mathbf{W}, \theta_0, h)] \leq c_{K_1,1} \delta^2_{1,n} - c_{K_1,2} \delta^2;$$

(iii) there exists a non-increasing sequence δ_n *such that*

$$\sup_{\theta \in B_{2,K_2,n}(\theta_0), h \in B_{1,K_1,n}(h_0)} |\mu_n[l_u(\mathbf{W}, \theta, h) - l_u(\mathbf{W}, \theta, h_0)]| = O_p(\delta^2_n);$$

(iv) for all n large enough and for any sufficiently small δ,

$$\mathbb{E}[\sup_{\{\theta \in B_{2,K_2,n}(\theta_0):||\theta-\theta_0||_2 \leq \delta\}} |\mu_n[l_u(\mathbf{W}, \theta, h_0) - l_u(\mathbf{W}, \theta_0, h_0)]| \leq \frac{c_1 \phi_n(\delta)}{\sqrt{n}},$$

where $c_1 > 0$ is some constant and $\phi_n(\cdot)$ is some function such that $\delta^{-\gamma}\phi_n(\delta)$ is a decreasing function for some $\gamma \in (0, 2)$.

Assumption C.4 corresponds to Assumption 5.4 in Hahn et al. (2018b) with a minor modification for condition (ii).

Proposition B.2 (Convergence Rate)**.** *Let $u \in \mathcal{U}$ be given. Suppose that Assumptions C.1 – C.4 hold and that there exists a finite non-increasing sequence $\delta_{\theta,n}$ such that*

$$(\delta_{\theta,n})^{-2}\phi_n(\delta_{\theta,n}) \leq c_2\sqrt{n}.$$

If $||\pi_n\theta_0 - \theta_0||_{\Theta,2} = O(\tilde{\delta}_{\theta,n})$, then,

$$||\hat{\theta}_n - \theta_0||_{\Theta,2} = O_p\left(\delta^*_{\theta,n}\right),$$

where $\delta^*_{\theta,n} \equiv \delta_{1,n} + \delta_{2,n} + \delta_{\theta,n} + \tilde{\delta}_n$ (or $\delta^*_{\theta,n} \equiv \max\left(\delta_{1,n}, \delta_{2,n}, \delta_{\theta,n}, \tilde{\delta}_{\theta,n}\right)$).

Proof. I follow the proof of Theorem 5.2 in Hahn et al. (2018b). Let $\omega > 0$ be a small constant. Since $\hat{\theta}_n$ is consistent, one can choose a large $K_M > 0$ such that

$$\Pr(||\hat{\theta}_n - \theta_0||_{\Theta,\infty} > K_M) \leq \omega. \quad (25)$$

Since $||\pi_n\theta_0 - \theta_0||_{\Theta,\infty} = o(1)$, there exists a large constant $K_{\theta_0} > 0$ such that $||\pi_n\theta_0 - \theta_0||_{\Theta,\infty} < K_{\theta_0}$. Define $K^*_M \equiv \max(K_M, K_{\theta_0})$ and

$$\Theta_n(M) \equiv \{\theta \in \Theta_n : 2^M\delta^*_{\theta,n} < ||\theta - \theta_0||_{\Theta,2} < K^*_M\}.$$

Let $I_{M,n}(\omega) \equiv \Pr(||\hat{\theta}_n - \theta_0||_{\Theta,2} > 2^M\delta^*_{\theta,n})$. Then, by (25),

$$I_{M,n}(\omega) \leq \Pr(\hat{\theta}_n \in \Theta_n(M)) + \omega. \quad (26)$$

Claim B.1. *Under the conditions,*

$$I_{M,n}(\omega) \leq \Pr(\sup_{\theta \in \Theta_n, h \in B_{1,K_1,n}(h_0)} [I_{1,n}(\theta, h_0) + I_{2,n}(\theta, h)] + K(\delta_n^2 + \delta_{2n}^2) \geq 0) + 5\omega$$

for some constant K.

Proof of the Claim From the definition of $\hat{\theta}_n$, one can choose some sufficiently large constant $K_1 > 0$ such that

$$\Pr(L_{u,n}(\hat{\theta}_n, \hat{h}_n) - L_{u,n}(\pi_n\theta_0, \hat{h}_n) < 0) \leq \omega. \quad (27)$$

Combining (26) and (27), it follows that

$$I_{M,n}(\omega) \leq \Pr(\sup_{\theta\in\Theta_n(M)} L_{u,n}(\theta,\hat{h}_n) - L_{u,n}(\pi_n\theta_0,\hat{h}_n) \geq 0) + 2\omega. \tag{28}$$

Considering the even in (28), it is straightforward to see that

$$\begin{aligned}
& L_{u,n}(\theta,\hat{h}_n) - L_{u,n}(\pi_n\theta_0,\hat{h}_n) \\
=& \mu_n[l_u(\mathbf{W},\theta,\hat{h}_n) - l_u(\mathbf{W},\pi_n\theta_0,\hat{h}_n)] + L_{u,0}(\theta,\hat{h}_n) - L_{u,0}(\pi_n\theta_0,\hat{h}_n) \\
=& \mu_n[l_u(\mathbf{W},\theta,\hat{h}_n) - l_u(\mathbf{W},\theta,h_0)] + \mu_n[l_u(\mathbf{W},\pi_n\theta_0,h_0) - l_u(\mathbf{W},\pi_n\theta_0,\hat{h}_n)] \\
& + \underbrace{\mu_n[l_u(\mathbf{W},\theta,h_0) - l_u(\mathbf{W},\pi_n\theta_0,h_0)]}_{=I_{1,n}(\theta,h_0)} + \underbrace{L_{u,0}(\theta,\hat{h}_n) - L_{u,0}(\theta_0,\hat{h}_n)}_{=I_{2,n}(\theta,\hat{h}_n)} \\
& + L_{u,0}(\theta_0,\hat{h}_n) - L_{u,0}(\pi_n\theta_0,\hat{h}_n).
\end{aligned} \tag{29}$$

By Assumption C.4-(iii), one can choose a large constant K_2 such that

$$\begin{aligned}
& \Pr(\sup_{\theta\in\Theta_n(M)} \mu_n[l_u(\mathbf{W},\theta,\hat{h}_n) - l_u(\mathbf{W},\theta,h_0)] \geq K_2\delta_n^2, \hat{h}_n \in \mathcal{N}_{1,K_1}) \\
& \leq \Pr(\sup_{\theta\in B_{2,K_M^*,n}(\theta_0), h\in B_{1,K_1,n}(h_0)} |\mu_n[l_u(\mathbf{W},\theta,h) - l_u(\mathbf{W},\theta,h_0)]| \geq K_2\delta_n^2) \leq \omega.
\end{aligned} \tag{30}$$

Combining (28), (29), and (30), one obtains that

$$I_{M,n}(\omega) \leq \Pr\left[\begin{pmatrix} \mu_n[l_u(\mathbf{W},\pi_n\theta_0,h_0) - l_u(\mathbf{W},\pi_n\theta_0,\hat{h}_n)] \\ +L_{u,0}(\theta_0,\hat{h}_n) - L_{u,0}(\pi_n\theta_0,\hat{h}_n) \\ + \sup_{\theta\in\Theta_n(M)} [I_{1,n}(\theta,h_0) + I_{2,n}(\theta,\hat{h}_n)] + K_2\delta_n^2 \end{pmatrix} \geq 0, \hat{h}_n \in \mathcal{N}_{1,K_1}\right] + 4\omega. \tag{31}$$

By the definition of $B_{2,K_M^*,n}(\theta_0)$, it is clear that $\pi_n\theta_0 \in B_{2,K_M^*,n}(\theta_0)$, and this, together with Assumption C.4-(iii), implies that

$$\begin{aligned}
& \Pr(\mu_n[l_u(\mathbf{W},\pi_n\theta_0,h_0) - l_u(\mathbf{W},\pi_n\theta_0,\hat{h}_n)] \geq K_2\delta_n^2, \hat{h}_n \in \mathcal{N}_{1,K_1}) \\
& \leq \Pr(\sup_{\theta\in B_{2,K_M^*,n}(\theta_0), h\in B_{1,K_1,n}(h_0)} |\mu_n[l_u(\mathbf{W},\theta,h_0) - l_u(\mathbf{W},\theta,h)]| \geq K_2\delta_n^2) \leq \omega.
\end{aligned} \tag{32}$$

By the same argument for (31), one can show that

$$I_{M,n}(\omega) \leq \Pr\left[\begin{pmatrix} L_{u,0}(\theta_0,\hat{h}_n) - L_{u,0}(\pi_n\theta_0,\hat{h}_n) \\ + \sup_{\theta\in\Theta_n(M)} [I_{1,n}(\theta,h_0) + I_{2,n}(\theta,\hat{h}_n)] \\ +2K_2\delta_n^2 \end{pmatrix} \geq 0, \hat{h}_n \in B_{1,K_1,n}(h_0)\right] + 5\omega. \tag{33}$$

From Assumption C.4-(i), one can choose a large constant K_3 such that

$$\sup_{h\in B_{1,K_1,n}(h_0)} |\mathbb{E}[l_u(\mathbf{W},\theta_0,h) - l_u(\mathbf{W},\pi_n\theta_0,h)]| \leq K_3\delta_{2n}^2,$$

and this implies that

$$\Pr(L_{u,0}(\theta_0,\hat{h}_n) - L_{u,0}(\pi_n\theta_0,\hat{h}_n) \geq K_3\delta_{2n}^2, \hat{h}_n \in B_{1,K_1,n}(h_0))$$
$$\leq \Pr(\sup_{h\in B_{1,K_1,n}(h_0)} |\mathbb{E}[l_u(\mathbf{W},\theta_0,h) - l_u(\mathbf{W},\pi_n\theta_0,h)]| \geq K_3\delta_{2n}^2) = 0.$$

Then, it follows that

$$I_{M,n}(\omega) \leq \Pr\left[\left(\begin{array}{r}\sup_{\theta\in\Theta_n(M)}[I_{1,n}(\theta,h_0) + I_{2,n}(\theta,\hat{h}_n)] \\ +2K_2\delta_n^2 + K_3\delta_{2n}^2\end{array}\right) \geq 0\right] + 6\omega \tag{34}$$

by the same way as before. Therefore,

$$I_{M,n}(\omega) \leq \Pr\left[\left(\begin{array}{r}\sup_{\theta\in\Theta_n(M)}[I_{1,n}(\theta,h_0) + I_{2,n}(\theta,\hat{h}_n)] \\ +K(\delta_n^2 + \delta_{2n}^2)\end{array}\right) \geq 0, \hat{h}_n \in B_{1,K_1,n}(h_0)\right] + 6\omega,$$

where $K \equiv \max\{2K_2, K_3\}$. This ends the proof of the claim.

Claim B.2. *Under the conditions,*

$$\Pr(\sup_{\theta\in\Theta_n(M), h\in B_{1,K_1,n}(h_0)} [I_{1,n}(\theta,h_0) + I_{2,n}(\theta,h)] + K(\delta_n^2 + \delta_{2n}^2) \geq 0)$$
$$\leq \sum_{j\geq M, 2^{j-1}\cdot\delta_{\theta,n}^* \leq K_M^*} \Pr(\sup_{\theta\in\Theta_{n,j}} I_{1,n}(\theta,h_0) \geq c_{K_1,2}2^{2j} - K - c_{K_1,1})\delta_{\theta,n}^{*^2},$$

where $\Theta_{n,j} \equiv \{\theta : 2^j\delta_{\theta,n}^* < ||\theta - \theta_0||_2 \leq 2^{j+1}\delta_{\theta,n}^*\}$.

Proof of Claim Partition $\Theta_n(M)$ into countably infinitely many disjoint pieces $\{\Theta_{n,j}\}_{j=M}^{\infty}$ (i.e. $\Theta_n(M) = \cup_{j=M}^{\infty}\Theta_{n,j}$ and $\Theta_{n,j} \cap \Theta_{n,j'} = \emptyset$ for any $j \neq j'$). Then,

$$\Pr(\sup_{\theta\in\Theta_n(M), h\in B_{1,K_1,n}(h_0)} [I_{1,n}(\theta,h_0) + I_{2,n}(\theta,h)] + K(\delta_n^2 + \delta_{2n}^2) \geq 0)$$
$$\leq \sum_{j\geq M, 2^{j-1}\delta_{\theta,n}^* \leq K_M^*} \Pr(\sup_{\theta\in\Theta_{n,j}, h\in\mathcal{N}_{1,K_1}} [I_{1,n}(\theta,h_0) + I_{2,n}(\theta,h)] + K(\delta_n^2 + \delta_{2n}^2) \geq 0). \tag{35}$$

Then, by Assumption C.4-(ii), one obtains that

$$\begin{aligned}\sup_{\theta\in\Theta_{n,j}, h\in B_{1,K_1,n}(h_0)} I_{2,n}(\theta,h) &= \sup_{\theta\in\Theta_{n,j}, h\in B_{1,K_1,n}(h_0)} L_{u,0}(\theta,h) - L_{u,0}(\theta_0,h) \\ &\leq c_{K_1,1}\delta_{1n}^2 - c_{K_1,2}(2^j\delta_{\theta,n}^*)^2 \\ &\leq (c_{K_1,1} - c_{K_1,2}2^{2j})\delta_{\theta,n}^{*^2}. \end{aligned} \tag{36}$$

Combining (35) and (36) ends the proof of claim.

Claim B.3. *Under the conditions,*

$$\Pr(\sup_{\theta\in\Theta_{n,j}} I_{1,n}(\theta,h_0)\geq (c_{K_1,2}2^{2j}-K-c_{K_1,1})\delta^{*2}_{\theta,n})\leq \frac{c_1c_2[2^{(j+1)\gamma}+K_\epsilon^\gamma]}{|c_{K_1,2}2^{2j}-K-c_{K_1,1}|}$$

where c denotes the generic constant and K_ϵ is some constant.

Proof of Claim Using Markov inequity and triangular inequality results in

$$\begin{aligned}\Pr(\sup_{\theta\in\Theta_{n,j}} I_{1,n}(\theta,h_0)\geq (c_{K_1,2}2^{2j}-K-c_{K_1,1})\delta^{*2}_{\theta,n}) &\leq \frac{\mathbb{E}[\sup_{\theta\in\Theta_{n,j}}|\mu_n[l_u(\mathbf{W},\theta,h_0)-l_u(\mathbf{W},\pi_n\theta_0,h_0)]|]}{|(c_{K_1,2}2^{2j}-K-c_{K_1,1})\delta^{*2}_{\theta,n}|}\\ &\leq \frac{\mathbb{E}[\sup_{\theta\in\Theta_{n,j}}|\mu_n[l_u(\mathbf{W},\theta,h_0)-l_u(\mathbf{W},\theta_0,h_0)]|]}{|(c_{K_1,2}2^{2j}-K-c_{K_1,1})\delta^{*2}_{\theta,n}|}\\ &+\frac{\mathbb{E}[|\mu_n[l_u(\mathbf{W},\theta_0,h_0)-l_u(\mathbf{W},\pi_n\theta_0,h_0)]|]}{|(c_{K_1,2}2^{2j}-K-c_{K_1,1})\delta^{*2}_{\theta,n}|}.\end{aligned}$$

By Assumption C.4-(iv), it follows that

$$\begin{aligned}\frac{\mathbb{E}[\sup_{\theta\in\Theta_{n,j}}|\mu_n[l_u(\mathbf{W},\theta,h_0)-l_u(\mathbf{W},\theta_0,h_0)]|]}{|(c_{K_1,2}2^{2j}-K-c_{K_1,1})\delta^{*2}_{\theta,n}|} &\leq \frac{c_1\phi_n(2^{j+1}\delta^*_{\theta,n})}{\sqrt{n}|(c_{K_1,2}2^{2j}-K-c_{K_1,1})\delta^{*2}_{\theta,n}|}\\ &=\frac{c_1(2^{j+1}\delta^*_{2n})^\gamma}{\sqrt{n}|(c_{K_1,2}2^{2j}-K-c_{K_1,1})\delta^{*2}_{\theta,n}|}\cdot\frac{\phi_n(2^{j+1}\delta^*_{\theta,n})}{(2^{j+1}\delta^*_{\theta,n})^\gamma}\\ &\leq \frac{c_1(2^{j+1})^\gamma}{|(c_{K_1,2}2^{2j}-K-c_{K_1,1})|}\frac{\phi_n(\delta^*_{\theta,n})}{\sqrt{n}\delta^*_{\theta,n}}\\ &\leq \frac{c_1c_2(2^{j+1})^\gamma}{|(c_{K_1,2}2^{2j}-K-c_{K_1,1})|},\end{aligned}$$

where the last two inequalities hold by the fact that $\delta\mapsto\phi_n(\delta)/\delta^\gamma$ is a decreasing function and the definition of $\delta_{\theta,n}$.

Since $||\pi_n\theta_0-\theta_0||_{\Theta,2}=O(\tilde{\delta}_{\theta,n})$, choose $K_\epsilon>1$ large enough so that $||\pi_n\theta_0-\theta_0||_2\leq K_\epsilon\tilde{\delta}_{\theta,n}$. By C.4-(iv) and the same argument above, one obtains that

$$\begin{aligned}\frac{\mathbb{E}[|\mu_n[l_u(\mathbf{W},\theta_0,h_0)-l_u(\mathbf{W},\pi_n\theta_0,h_0)]|]}{|(c_{K_1,2}2^{2j}-K-c_{K_1,1})\delta^{*2}_{\theta,n}|} &\leq \frac{\mathbb{E}[|\sup_{\theta\in\Theta_n:||\theta-\theta_0||_{\Theta,2}\leq K_\epsilon\tilde{\delta}_{\theta,n}}\mu_n[l_u(\mathbf{W},\theta_0,h_0)-l_u(\mathbf{W},\theta,h_0)]|]}{|(c_{K_1,2}2^{2j}-K-c_{K_1,1})|}\\ &\leq \frac{c_1c_2K_\epsilon^\gamma}{|(c_{K_1,2}2^{2j}-K-c_{K_1,1})|},\end{aligned}$$

and therefore the claim holds. Then, it follows that

$$I_{M,n}(\omega)\leq \sum_{j\geq M,2^{j-1}\delta^*_{\theta,n}\leq K^*_M}\frac{c_1c_2[(2^{j+1})^\gamma+K_\epsilon^\gamma]}{|(c_{K_1,2}2^{2j}-K-c_{K_1,1})|}+5\omega.$$

Since $\gamma<2$, one can choose M sufficiently large so that $\sum_{j\geq M,2^{j-1}\delta^*_{\theta,n}\leq K^*_M}\frac{c_1c_2[(2^{j+1})^\gamma+K_\epsilon^\gamma]}{|(c_{K_1,2}2^{2j}-K-c_{K_1,1})|}\leq\omega$, which leads to that $I_{M,n}(\omega)=\Pr(||\hat{\theta}_n-\theta_0||_{\Theta,2}>2^M\delta^*_{\theta,n})\leq 6\omega$ for sufficiently large n and $M>0$. This ends the proof. ■

B.3 Asymptotic Normality

The following assumption is a version of Assumption 3.2 in Hahn et al. (2018a), which is a high-level assumption, that is needed to establish the asymptotic normality of functionals. For any $g \in \mathcal{N}_{g,n}$, let $g^* \equiv g \pm \kappa_n u^*_{g_n}$ with $\kappa_n = o(n^{-1/2})$, and recall that $(u^*_{h_n}, u^*_{g_n}, u^*_{\Gamma_n}) \equiv ||v_n||^{-1}_{sd}(v^*_{h_n}, v^*_{g_n}, v^*_{\Gamma_n})$.

Assumption C.5. *(i) The following conditions hold:*

$$\sup_{\alpha \in \mathcal{N}_{\alpha,n}} \left| \mu_n \left[l_u(\mathbf{W}, g^*, h) - l_u(\mathbf{W}, g, h) - \Delta_2(\mathbf{W}, g, h)[\pm \kappa_n u^*_{g_n}] \right] \right| = O_p(\kappa_n^2), \tag{37}$$

$$\sup_{\alpha \in \mathcal{N}_{\alpha,n}} \left| \mu_n \left[\Delta_2(\mathbf{W}, g, h)[u^*_{g_n}] - \Delta_2(\mathbf{W}, g_0, h_0)[u^*_{g_n}] \right] \right| = O_p(\kappa_n); \tag{38}$$

(ii) define $\mathbb{K}(g,h) \equiv \mathbb{E}[l_u(\mathbf{W}, g, h) - l_u(\mathbf{W}, g_0, h_0)]$*. Then, uniformly over* $(h,g) \in \mathcal{N}_{\alpha,n}$*,*

$$\mathbb{K}(g,h) - \mathbb{K}(g^*,h) = \mp \kappa_n \Gamma(\alpha_0)[h - h_{o,n}] + \frac{||g^* - g_0||^2_{2,\Delta} - ||g - g_0||^2_{2,\Delta}}{2} + O(\kappa_n^2).$$

Proposition B.3. *Suppose that Assumptions 7, 8, 9, and C.5 hold. Then,*

$$\sqrt{n} \frac{f(\hat{\alpha}_n) - f(\alpha_0)}{||v^*_n||_{sd}} \xrightarrow{d} N(0,1).$$

Proof. This result is a direct consequence of Theorem 3.1 in Hahn et al. (2018a). ■