The Dissertation Committee for Sungwon Lee
certifies that this is the approved version of the following dissertation:

Essays on Semi-/Non-parametric Methods in Econometrics

Committee:

Stephen Donald, Supervisor

Jason Abrevaya

Sukjin Han

Thomas Shively

Essays on Semi-/Non-parametric Methods in Econometrics

by

Sungwon Lee

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2018

Dedicated to my parents.

Acknowledgments

This work could not have been done without the support and encouragement from many people around me. I am deeply indebted to my advisor Stephen Donald for his guidance and support throughout my graduate life. He has provided me with many valuable lessons and the time to pursue my academic goals. I am also very grateful to Jason Abrevaya and Sukjin Han for their thoughtful discussions and suggestions. They have always helped me whenever I was in trouble and inspired me to become a good researcher. I also thank Brendan Kline, Thomas Shively, and Haiqing Xu for their sharp comments on my dissertation.

My gratitude also goes to my friends. I have benefited from discussions on my research with my research colleagues, Jessie, Peter, and Xinchen. I am sincerely grateful to other friends, Yeonjoon, Choongryul, Jaemin, Narae, David, Gabe, Joon, Eunjoo, Changseung, Haejung, Byungjae, and Jiwon.

I am grateful to my family for their support and encouragement as well. Lastly but not least, I would thank God for his guidance throughout my life.

Essays on Semi-/Non-parametric Methods in Econometrics

Publication No. _____

Sungwon Lee, Ph.D.
The University of Texas at Austin, 2018

Supervisor: Stephen Donald

My dissertation contains three chapters focusing on semi-/non-parametric models in econometrics. The first chapter, which is a joint work with Sukjin Han, considers parametric/semiparametric estimation and inference in a class of bivariate threshold crossing models with dummy endogenous variables. We investigate the consequences of common practices employed by empirical researchers using this class of models, such as the specification of the joint distribution of the unobservables to be a bivariate normal distribution, resulting in a bivariate probit model. To address the problem of misspecification, we propose a semiparametric estimation framework with parametric copula and nonparametric marginal distributions. This specification is an attempt to ensure robustness while achieving point identification and efficient estimation. We establish asymptotic theory for the sieve maximum likelihood estimators that can be used to conduct inference on the individual structural parameters and the average treatment effects. Numerical studies suggest the sensitivity of parametric specification and the robustness of semiparametric estimation. This paper also shows that the absence of excluded instruments may result in the

failure of identification, unlike what some practitioners believe.

The second chapter develops nonparametric significance tests for quantile regression models with duration outcomes. It is common for empirical studies to specify models with many covariates to eliminate the omitted variable bias, even if some of them are potentially irrelevant. In the case where models are nonparametrically specified, such a practice results in the curse of dimensionality. I adopt the integrated conditional moment (ICM) approach, which was developed by Bierens (1982); Bierens (1990), to construct test statistics. The proposed test statistics are functionals of a stochastic process which converges weakly to a centered Gaussian process. The test has non-trivial power against local alternatives at the parametric rate. A subsampling procedure is proposed to obtain critical values.

The third chapter considers identification of treatment effect and its distribution under some distributional assumptions. I assume that a binary treatment is endogenously determined. The main identification objects are the quantile treatment effect and the distribution of the treatment effect. I construct a counterfactual model and apply Manski's approach (Manski (1990)) to find the quantile treatment effects. For the distribution of the treatment effect, I adapt the approach proposed by Fan and Park (2010). Some distributional assumptions called stochastic dominance are imposed on the model to tighten the bounds on the parameters of interest. It also provides confidence regions for identified sets that are pointwise consistent in level. An empirical study on the return to college confirms that the stochastic dominance assumptions improve the bounds on the distribution of the treatment effect.

Table of Contents

List of Tables

List of Figures

Chapter 1

Sensitivity Analysis in Triangular Systems of Equations with Binary Endogenous Variables

This is a joint work with Sukjin Han.

1.1 Introduction

This paper considers parametric/semiparametric estimation and inference in a class of bivariate threshold crossing models with dummy endogenous variables. Let Y denote the binary outcome variable and D the observed binary endogenous treatment variable. We consider a bivariate triangular system for (Y, D):

$$\begin{aligned} Y &= \mathbf{1}[X'\beta + \delta_1 D - \varepsilon \geq 0], \\ D &= \mathbf{1}[X'\alpha + Z'\gamma - \nu \geq 0], \end{aligned} \tag{1.1.1}$$

where X denotes the vector of exogenous regressors that determine both Y and D, and Z denotes a vector of exogenous regressors that directly affect D but not Y (i.e., instruments for D). In this paper, we investigate the consequences of common practices employed by empirical researchers using this class of models. As important part of this investigation, we conduct a sensitivity analysis regarding the specification of the unobservables (ε, ν)'s joint distribution, which is the component of the model that practitioners are mostly agnostic about and for which a parametric assumption

is typically imposed. To address the problem of misspecification, we propose a semiparametric estimation framework with parametric copula and nonparametric marginal distributions. This specification is an attempt to ensure robustness while achieving point identification and efficient estimation.

A parametric class of models (1.1.1) includes the *bivariate probit model* in which the joint distribution of (ε, ν) is assumed to be a bivariate normal distribution. This model has been widely used in empirical research such as Evans and Schwab (1995), Neal (1997), Goldman et al. (2001), Altonji et al. (2005), Bhattacharya et al. (2006), Rhine et al. (2006) and Marra and Radice (2011) to name a just few. The distributional assumption in this model, however, is made out of convenience or convention and hardly justified by underlying economic theory, thereby susceptible to misspecification. With binary endogenous regressors, the objects of interest in model (1.1.1) are mean treatment parameters besides the individual structural parameters. As the outcome variable is also binary, mean treatment parameters such as the average treatment effect (ATE) are expressed as the differential between the marginal distributions of ε. The problem of misspecification in estimating these treatment parameters can therefore be even more severe than that in estimating individual parameters.

To one extreme, a nonparametric joint distribution of (ε, ν) can be used in bivariate threshold crossing models as in Shaikh and Vytlacil (2011). As their results suggests, however, the ATE is only partially identified in this fully flexible setting. Instead of sacrificing point identification, we impose a parametric assumption on the dependence structure between the unobservables using copula functions that are

known up to a scalar parameter. At the same time, in order to ensure robustness, we allow to be unspecified the marginal distribution of ε (as well as of ν), which is involved in the calculation of the ATE. In this way, our class of models encompasses both parametric and semiparametric classes of models with parametric copula and either parametric or nonparametric marginal distributions. This broad range of models allows us to conduct a sensitivity analysis in terms of the specification of the joint distribution of (ε, ν).

The identification of the individual parameters as well as the ATE in this class of models is established in (Han and Vytlacil, 2017, hereafter HV17). They show that when the copula function for (ε, ν) satisfies a certain stochastic ordering, identification is achieved in both parametric and semiparametric models under an exclusion restriction and mild support conditions. The present paper, building on these results, considers estimation and inference in the same setting. For the semiparametric class of models (1.1.1) with parametric copula and nonparametric marginal distributions, the likelihood contains infinite dimensional parameters, i.e., the unknown marginal distributions. For the estimation of this model, we consider sieve maximum likelihood (ML) estimators for the finite and infinite dimensional parameters of the model as well as the functionals of them. The estimation of the parametric model is within the standard ML framework.

The contributions of this paper can be summarized as follows. This paper is intended to provide a guideline to empirical researchers through these contributions. First, we establish the asymptotic theory for the sieve ML estimators in a class of semiparametric copula-based models. This result can be used to conduct inference

on the functionals of the finite and infinite dimensional parameters, such as inference on the individual structural parameters and the ATE. We show that the sieve ML estimators are consistent and their smooth functionals are root-n asymptotically normal.

Second, based on these theoretical results, we conduct a sensitivity analysis via Monte Carlo simulation studies. We find that the parametric ML estimates can be very sensitive to the misspecification of the marginal distributions of the unobservables. We show that, on the other hand, sieve ML estimates perform well in terms of the mean squared error (MSE) as they are robust to this misspecification, while their performance is comparable to the parametric estimates under correct specification. We also show that copula misspecification does not have substantial effects in estimation as long as the true copula is within the stochastic ordering class for identification. Since copula misspecification is a problem common to both parametric and semiparametric models, our sensitivity analysis suggests to practitioners that semiparametric consideration can be desirable in estimation and inference.

Third, we formally show that identification may fail without the exclusion restriction, unlike what is argued in Wilde (2000). The bivariate probit model is sometimes used in applied work without instruments (e.g., White and Wolaver (2003) and Rhine et al. (2006)). We show, however, that this restriction is not only sufficient but also necessary for identification in parametric and semiparametric models when there is a single binary exogenous variable common to both equations. We also show that, under joint normality of the unobservables, the parameters are at best weakly

identified when there are common (possibly continuous) exogenous variables[1]. We also note that another source of identification failure is the absence of restrictions on the dependence structure of the unobservables as mentioned above.

The sieve estimation method is a useful nonparametric estimation framework that allows flexible specification while guarantees tractability of the estimation problem; see Chen (2007) for a survey of sieve estimation in semi-nonparametric models. The estimation method is also easy to implement in practice. The sieve ML estimation is used in various contexts: (Chen et al., 2006, hereafter CFT06) consider the sieve estimation of semiparametric multivariate distributions that are modeled using parametric copulas; Bierens (2008) applies it to the mixed proportional hazard model; Hu and Schennach (2008) and Chen et al. (2009) use the method to estimate nonparametric models with non-classical measurement errors. The asymptotic theory developed in this paper is based on the results established in the sieve extremum estimation literature, e.g., Chen et al. (2006); Chen (2007); Bierens (2014). A semiparametric version of bivariate threshold crossing models is also considered in Marra and Radice (2011) and Ieva et al. (2014), but unlike in the present paper, flexibility is introduced for the index of the threshold and not for the distribution of the unobservables.

The paper is organized as follows. We start the next section by reviewing the identification results of HV17, and then discuss the lack of identification in the absence of exclusion restrictions and in the absence of restrictions on the depen-

[1] HV17 only show sufficiency of this restriction for identification. Mourifié and Méango (2014) show necessity of the restriction but their argument does not exploit all the information available in the model; see Section 2.2 of the present paper for details.

dence structure of the unobservables. Section 1.3 introduces the sieve ML estimation framework for the semiparametric class of model (1.1.1), and Section 1.4 establishes the large sample theory for the sieve ML estimators. The sensitivity analysis is conducted in Section 1.5 by investigating finite sample performances of parametric ML and sieve ML estimates in various different specifications. Section 1.6 concludes.

1.2 Identification and Failure of Identification

1.2.1 Identification

In model (1.1.1), let $\underset{(k+1)\times 1}{X} \equiv (1, X_1, ..., X_k)'$ and $\underset{l\times 1}{Z} \equiv (Z_1, ..., Z_l)'$, and conformably, let $\alpha \equiv (\alpha_0, \alpha_1, ..., \alpha_k)'$, $\beta \equiv (\beta_0, \beta_1, ..., \beta_k)'$, and $\gamma \equiv (\gamma_1, \gamma_2, ..., \gamma_l)'$.

Assumption 1.2.1. *X and Z satisfy that $(X, Z) \perp (\varepsilon, \nu)$, where "$\perp$" denotes statistical independence..*

Assumption 1.2.2. *(X', Z') does not lie in a proper linear subspace of $\mathbb{R}^{k+l}$ a.s.*[2]

Assumption 1.2.3. *There exists a copula function $C : (0,1)^2 \to (0,1)$ such that the joint distribution $F_{\varepsilon\nu}$ of (ε, ν) satisfies $F_{\varepsilon\nu}(\varepsilon, \nu) = C(F_\varepsilon(\varepsilon), F_\nu(v))$, where F_ε and F_ν are marginal distributions of ε and ν, respectively, that are strictly increasing and absolutely continuous with respect to Lebesgue measure.*[3]

Assumption 1.2.4. *As scale and location normalizations, $\alpha_1 = \beta_1 = 1$ and $\alpha_0 = \beta_0 = 0$.*

[2] A proper linear subspace of $\mathbb{R}^{k+l}$ is a linear subspace with a dimension strictly less than $k + l$. The assumption is that, if M is a proper linear subspace of $\mathbb{R}^{k+l}$, then $\Pr[(X', Z') \in M] < 1$.

[3] The Sklar's theorem (e.g., Nelsen (1999)) guarantees the existence of such a copula, which is in fact unique as F_ε and F_ν are continuous.

A model with alternative scale and location normalizations, $Var(\varepsilon) = Var(\nu) = 1$ and $E[\varepsilon] = E[\nu] = 0$, can be seen as a reparametrized version of the model with the normalizations in Assumption 1.2.4; see e.g., the reparametrization (1.2.1) below. For $x \in \text{supp}(X)$ and $z \in \text{supp}(Z)$, write a one-to-one map (by Assumption 1.2.3) as

$$\begin{aligned} s_{xz} &\equiv F_\nu(x'\alpha + z'\gamma), \\ r_{0,x} &\equiv F_\varepsilon(x'\beta), \\ r_{1,x} &\equiv F_\varepsilon(x'\beta + \delta_1). \end{aligned} \tag{1.2.1}$$

Take (x, z) and $(x, \tilde{z})$ for some $x \in \text{supp}(X|Z = z) \cap \text{supp}(X|Z = \tilde{z})$ where $\text{supp}(X|Z)$ is the conditional support of X given Z. Then by Assumption 1.2.1, model (1.1.1) implies that the fitted probabilities are written as

$$\begin{aligned} p_{11,xz} &= C(r_{1,x}, s_{xz}), \\ p_{11,x\tilde{z}} &= C(r_{1,x}, s_{x\tilde{z}}), \\ p_{10,xz} &= r_{0,x} - C(r_{0,x}, s_{xz}), \\ p_{10,x\tilde{z}} &= r_{0,x} - C(r_{0,x}, s_{x\tilde{z}}), \\ p_{01,xz} &= s_{xz} - C(r_{1,x}, s_{xz}), \\ p_{01,x\tilde{z}} &= s_{x\tilde{z}} - C(r_{1,x}, s_{x\tilde{z}}), \end{aligned} \tag{1.2.2}$$

where $p_{yd,xz} \equiv \Pr[Y = y, D = d|X = x, Z = z]$ for $(y, d) \in \{0, 1\}^2$. The equation (1.2.2) serves as the basis for identification and estimation of the model. Depending upon whether one is willing to impose an additional assumption on the dependence

structure of the unobservables (ε, ν) via $C(\cdot,\cdot)$, the underlying parameters of the model is either point identified or partially identified.

We first consider point identification. The results for point identification can be found in HV17, which we adapt here given Assumption 1.2.4. The additional dependence structure can be characterized in terms of the stochastic ordering of the copula parametrized with a scalar parameter.

Definition 1.2.5 (Strictly More SI or Less SD)**.** *Let $C(u_2|u_1)$ and $\tilde{C}(u_2|u_1)$ be conditional copulas, for which $1 - C(u_2|u_1)$ and $1 - \tilde{C}(u_2|u_1)$ are either increasing or decreasing in u_1 for all u_2. Such copulas are called to be stochastically increasing (SI) or stochastically decreasing (SD), respectively. Then $\tilde{C}$ is strictly more SI (or less SD) than C if $\psi(u_1, u_2) \equiv \tilde{C}^{-1}(C(u_2|u_1)|u_1)$ is strictly increasing in u_1,*[4] *which is denoted as $C \prec_S \tilde{C}$.*

This ordering is equivalent to having a ranking in terms of the first order stochastic dominance. Let $(U_1, U_2) \sim C$ and $(\tilde{U}_1, \tilde{U}_2) \sim \tilde{C}$. When $\tilde{C}$ is strictly more SI (less SD) than C, then $\Pr[\tilde{U}_2 > u_2|\tilde{U}_1 = u_1]$ increases even more than $\Pr[U_2 > u_2|U_1 = u_1]$ as u_1 increases.[5]

Assumption 1.2.6. *The copula in Assumption 1.2.3 satisfies $C(\cdot,\cdot) = C(\cdot,\cdot;\rho)$ with a scalar dependence parameter $\rho \in \Omega$, is twice differentiable in u_1, u_2 and ρ, and satisfies*

$$C(u_1|u_2;\rho_1) \prec_S C(u_1|u_2;\rho_2) \text{ for any } \rho_1 < \rho_2. \tag{1.2.3}$$

[4] Note that $\psi(u_1, u_2)$ is increasing in u_2 by definition.

[5] The SI dependence ordering is also called the (strictly) "more regression dependent" or "more monotone regression dependent" ordering in the statistics literature; see Joe (1997) for details.

The meaning of the last part of this assumption is that the copula is ordered in ρ in the sense of the stochastic ordering defined above. This requirement defines a class of copulas that we allow for identification. Many well-known copulas satisfy (1.2.3): the normal copula, Plackett copula, Frank copula, Clayton copula and many more; see HV17 for the full list of copulas and their expressions. Under these assumptions, we first discuss the identification in a fully parametric model:

Assumption 1.2.7. *F_ε and F_ν are known with means $\mu \equiv (\mu_\varepsilon, \mu_\nu)$ and variances $\sigma^2 \equiv (\sigma^2_\varepsilon, \sigma^2_\nu)$.*

Given this assumption, $F_\nu(\nu) = F_{\tilde{\nu}}(\tilde{\nu})$ and $F_\varepsilon(\varepsilon) = F_{\tilde{\varepsilon}}(\tilde{\varepsilon})$ where $F_{\tilde{\nu}}$ and $F_{\tilde{\varepsilon}}$ are the distributions of $\tilde{\nu} \equiv (\nu - \mu_\nu)/\sigma_\nu$ and $\tilde{\varepsilon} \equiv (\varepsilon - \mu_\varepsilon)/\sigma_\varepsilon$, respectively. Define

$$\mathcal{X} \equiv \bigcup_{\substack{z'\gamma \neq \tilde{z}'\gamma \\ z,\tilde{z} \in \text{supp}(Z)}} \text{supp}(X|Z=z) \cap \text{supp}(X|Z=\tilde{z}).$$

Theorem 1.2.8. *In model (1.1.1), suppose Assumptions 1.2.1–1.2.7 hold. Then $(\alpha', \beta', \delta_1, \gamma, \rho, \mu, \sigma)$ are point identified in an open and convex parameter space if (i) γ is a nonzero vector; and (ii) $\mathcal{X}$ does not lie in a proper linear subspace of $\mathbb{R}^k$ a.s.*

The proofs of this theorem is minor modification of the proof of Theorem 5.1 in HV17.

Although the parametric structure on the copula is necessary for point identification of the parameters, HV17 show that the parametric assumption for F_ε and F_ν are not necessary. Additionally, if we make a large support assumption, we can also identify the nonparametric marginal distributions F_ε and F_ν.

Assumption 1.2.9. *(i) The distributions of X_j (for $1 \leq j \leq k$) and Z_j (for $1 \leq j \leq l$) are absolutely continuous with respect to Lebesgue measure; (ii) There exists at least one element X_j in X such that its support conditional on $(X_1, ..., X_{j-1}, X_{j+1}, ..., X_k)$ is $\mathbb{R}$ and $\alpha_j \neq 0$ and $\beta_j \neq 0$, where, without loss of generality, we let $j = 1$.*

Theorem 1.2.10. *In model (1.1.1), suppose Assumptions 1.2.1–1.2.6, and 1.2.9(i) hold. Then $(\alpha', \beta', \delta_1, \gamma, \rho)$ are point identified in an open and convex parameter space if (i) γ is a nonzero vector; and (ii) $\mathcal{X}$ does not lie in a proper linear subspace of $\mathbb{R}^k$ a.s. Additionally, if Assumption 1.2.9(ii) holds, $F_\varepsilon(\cdot)$ and $F_\nu(\cdot)$ are identified up to additive constants.*

An interesting function of the underlying parameters that are point identified in under the parametric and semiparametric distributional assumptions is the conditional ATE:

$$ATE(x) = E[Y_1 - Y_0 | X = x] = F_\varepsilon(x'\beta + \delta_1) - F_\varepsilon(x'\beta). \tag{1.2.4}$$

1.2.2 The Failure of Identification

In this section, we discuss two sources of identification failure, namely, the absence of exclusion restrictions and the absence of restrictions on the dependence structure of the unobservables (ε, ν).

1.2.2.1 No Exclusion Restrictions

There is applied work where (1.1.1) is used without excluded instruments; see e.g., White and Wolaver (2003) and Rhine et al. (2006). Identification in these

papers relies on Wilde (2000), which provides an identification argument of counting the number of equations and unknowns in the system. Here we show that this argument is insufficient for identification. We show that without excluded instruments, i.e., when $\gamma = 0$, the structural parameters are not identified even with full parametric specification of the joint distribution (Assumptions 1.2.6 and 1.2.7). The existence of common exogenous covariates X in both equations is not very helpful for identification in a sense that becomes clear below.

Before considering the lack of identification in a general case with possibly continuous X_1 in $X = (1, X_1)$, we start the analysis with binary X_1. Mourifié and Méango (2014) show the lack of identification when there is no excluded instrument in the bivariate probit model with binary X_1. They, however, only provide a numerical counter-example. Moreover, their analysis does not consider the full set of observed fitted probabilities, and hence possibly neglects information that could have contributed for identification. Here we provide an analytical counter-example in a more general parametric class of model (1.1.1) that nests the bivariate probit model. We shows that there exists two distinct values of $(\delta_1, \rho, \mu_\varepsilon, \sigma_\varepsilon)$ that generate the same observed fitted probabilities, even if the full set of probabilities are used. Note that the reduced-form parameters (μ_ν, σ_ν) are always identified from the equation for D, and $\alpha = \beta = (0, 1)'$ as normalization with scalar X_1.

Theorem 1.2.11. *In model (1.1.1) with $X = (1, X_1)$ where $X_1 \in supp(X_1) = \{0, 1\}$, suppose that the assumptions in Theorem 1.2.8 hold, except that $\gamma = 0$. Then there exist two distinct sets of $(\delta_1, \rho, \mu_\varepsilon, \sigma_\varepsilon)$ that generate the same observed data.*

In showing this result, we find a counter-example where the copula density

induced by $C(u_1, u_2)$ is symmetric around $u_2 = u_1$ and $u_2 = 1 - u_1$, and the density induced by F_ε is symmetric. Note that the bivariate normal distribution, namely, the normal copula with normal marginals, satisfies these symmetry properties. That is, *in the bivariate probit model with a common binary exogenous covariate and no excluded instruments, the structural parameters are not identified.*

Under Assumption 1.2.4, let

$$
\begin{aligned}
q_0 &\equiv F_{\tilde{\nu}}(-\mu_\nu/\sigma_\nu), \\
q_1 &\equiv F_{\tilde{\nu}}((1-\mu_\nu)/\sigma_\nu), \\
t_0 &\equiv F_{\tilde{\varepsilon}}(-\mu_\varepsilon/\sigma_\varepsilon), \\
t_1 &\equiv F_{\tilde{\varepsilon}}((1-\mu_\varepsilon)/\sigma_\varepsilon),
\end{aligned}
$$

we have

$$
\begin{aligned}
\tilde{p}_{11,0} &= C(F_{\tilde{\varepsilon}}(F_{\tilde{\varepsilon}}^{-1}(t_0) + \delta_1), q_0; \rho), \\
\tilde{p}_{11,1} &= C(F_{\tilde{\varepsilon}}(F_{\tilde{\varepsilon}}^{-1}(t_1) + \delta_1), q_1; \rho), \\
\tilde{p}_{10,0} &= t_0 - C(t_0, q_0; \rho), \\
\tilde{p}_{10,1} &= t_1 - C(t_1, q_1; \rho), \\
\tilde{p}_{00,0} &= 1 - t_0 - q_0 + C(t_0, q_0; \rho), \\
\tilde{p}_{00,1} &= 1 - t_1 - q_1 + C(t_1, q_1; \rho),
\end{aligned}
$$

where $\tilde{p}_{yd,x} \equiv \Pr[Y = y, D = d | X_1 = x]$. We want to show that, given (q_0, q_1) which are identified from the reduced-form equation, there are two distinct sets of parameter values $(t_0, t_1, \delta_1, \rho)$ and $(t_0^*, t_1^*, \delta_1^*, \rho^*)$ (with $(t_0, t_1, \delta_1, \rho) \neq (t_0^*, t_1^*, \delta_1^*, \rho^*)$)

that generate the same observed fitted probabilities $\tilde{p}_{yd,0}$ and $\tilde{p}_{yd,1}$ for all $(y,d) \in \{0,1\}^2$. In showing this, the following lemma is useful:

Lemma 1.2.1. *Assumption 1.2.6 implies that, for any* $(u_1,u_2) \in (0,1)^2$ *and* $\rho \in \Omega$,

$$C_\rho(u_1,u_2;\rho) > 0. \tag{1.2.5}$$

The proofs of this lemma and other results below are collected in the Appendix.

Now fix $(q_0,q_1) \in (0,1)^2$. First, consider the fitted probability $\tilde{p}_{10,0}$. Given $t_0 \in (0,1)$ and $\rho \in \Omega$, note that, for $\rho^* > \rho$[6], there exists a solution $t_0^* = t_0^*(t_0,q_0,\rho,\rho^*)$ such that

$$\begin{aligned} t_0 - C(t_0,q_0;\rho) &= \Pr[u_1 \le t_0, u_2 \ge q_0;\rho] && (1.2.6)\\ &= \Pr[u_1 \le t_0^*, u_2 \ge q_0;\rho^*] && (1.2.7)\\ &= t_0^* - C(t_0^*,q_0;\rho^*), \end{aligned}$$

and note that by Assumption 1.2.6 and a variant of Lemma 1.2.1, we have that $t_0^* > t_0$. Here, (t_0,q_0,ρ) and (t_0^*,q_0,ρ^*) result in the same observed probability $\tilde{p}_{10,0} = t_0 - C(t_0,q_0;\rho) = t_0^* - C(t_0^*,q_0;\rho^*)$. Now consider the fitted probability $\tilde{p}_{11,0}$. Choose $\delta_1 = 0$. Also let $F_{\tilde{\varepsilon}} \sim Unif(0,1)$ only for simplicity, which is relaxed in the

[6] The inequality here and other inequalities implied from this (e.g., $t_0^* > t_0$, and etc.) are assumed only for concreteness.

Appendix. Then there exists a solution $t_0^\dagger = t_0^\dagger(t_0, q_0, \rho, \rho^*)$ such that

$$\begin{aligned} C(t_0, q_0; \rho) &= \Pr[u_1 \leq t_0, u_2 \leq q_0; \rho] && (1.2.8) \\ &= \Pr[u_1 \leq t_0^\dagger, u_2 \leq q_0; \rho^*] && (1.2.9) \\ &= C(t_0^\dagger, q_0; \rho^*), \end{aligned}$$

and note that $t_0^\dagger < t_0$ by Assumption 1.2.6 and Lemma 1.2.1. Then, by letting $\delta_1^* = t_0^\dagger - t_0^*$, $(t_0, q_0, \delta_1, \rho)$ and $(t_0^*, q_0, \delta_1^*, \rho^*)$ satisfy $\tilde{p}_{11,0} = C(t_0 + 0, q_0; \rho) = C(t_0^* + \delta_1^*, q_0; \rho^*)$. Lastly, note that $\tilde{p}_{00,0} = 1 - q_0 - \tilde{p}_{10,0}$ and $\tilde{p}_{01,0} = q_0 - \tilde{p}_{11,0}$, and so (t_0, δ_1, ρ) and $(t_0^*, \delta_1^*, \rho^*)$ above will also result in the same values of $\tilde{p}_{00,0}$ and $\tilde{p}_{01,0}$.

It is tempting to have a parallel argument for $\tilde{p}_{10,1}$, $\tilde{p}_{11,1}$, $\tilde{p}_{00,1}$, and $\tilde{p}_{01,1}$, but there is a complication. Although other parameters are not, δ_1 and ρ are common in both sets of probabilities. Therefore, we proceed as follows. First, consider $\tilde{p}_{10,1}$. Given $t_1 \in (0, 1)$ and the above choice of $\rho^* \in \Omega$, note that there exists a solution $t_1^* = t_1^*(t_1, q_1, \rho, \rho^*)$ such that

$$\begin{aligned} t_1 - C(t_1, q_1; \rho) &= \Pr[u_1 \leq t_1, u_2 \geq q_1; \rho] && (1.2.10) \\ &= \Pr[u_1 \leq t_1^*, u_2 \geq q_1; \rho^*] && (1.2.11) \\ &= t_1^* - C(t_1^*, q_1; \rho^*), \end{aligned}$$

and similarly as before, we have $t_1^* > t_1$. Here, (t_1, q_1, ρ) and (t_1^*, q_1, ρ^*) result in the same observed probability $\tilde{p}_{10,1} = t_1 - C(t_1, q_1; \rho) = t_1^* - C(t_1^*, q_1; \rho^*)$. Now consider $\tilde{p}_{11,1}$. Recall $\delta_1 = 0$ and $F_\varepsilon \sim Unif(0, 1)$. Then there exists a solution

$t_1^\dagger = t_1^\dagger(t_1, q_1, \rho, \rho^*)$ such that

$$C(t_1, q_1; \rho) = \Pr[u_1 \leq t_1, u_2 \leq q_1; \rho] \tag{1.2.12}$$

$$= \Pr[u_1 \leq t_1^\dagger, u_2 \leq q_1; \rho^*] \tag{1.2.13}$$

$$= C(t_1^\dagger, q_1; \rho^*),$$

and thus $t_1^\dagger < t_1$. Then, if we can show that

$$t_1^\dagger = t_1^* + \delta_1^*, \tag{1.2.14}$$

where t_1^* and δ_1^* are the values already determined above, then $(t_1, q_1, \delta_1, \rho)$ and $(t_1^*, q_1, \delta_1^*, \rho^*)$ result in $\tilde{p}_{11,1} = C(t_1 + 0, q_1; \rho) = C(t_1^* + \delta_1^*, q_1; \rho^*)$. Then similar as before, the two sets of parameters will generate the same values of $\tilde{p}_{00,1} = 1 - q_1 - \tilde{p}_{10,1}$ and $\tilde{p}_{01,1} = q_1 - \tilde{p}_{11,1}$. Consequently, $(t_0, t_1, q_0, q_1, \delta_1, \rho)$ and $(t_0^*, t_1^*, q_0, q_1, \delta_1^*, \rho^*)$ generate the same entire observed fitted probabilities. The remaining question is whether we can find $(t_0, t_1, \delta_1, \rho)$ and $(t_0^*, t_1^*, \delta_1^*, \rho^*)$ such that (1.2.14) holds; this is shown in the Appendix where $F_{\tilde{\varepsilon}} \sim Unif(0, 1)$ is also relaxed.

One might argue that the lack of identification in Theorem 1.2.11 is due to the limited variation of X. Although it is a plausible conjecture, this does not seem to be the case with the model considered in this paper.[7] We now consider a general case with possibly continuous X_1 and discuss what can be said about the existence of two distinct sets of $(\beta, \delta_1, \rho, \mu_\varepsilon, \sigma_\varepsilon)$ that generate the same observed data. To this

[7] In fact, in Heckman (1979)'s sample selection model under normality, although identification fails with binary exogenous covariates in the absence of exclusion restriction, it is well-known that identification is achieved with continuous covariates by exploiting the nonlinearity of the model (Vella (1998)).

end, define

$$q(x) \equiv F_{\tilde{\nu}}((x'\alpha - \mu_\nu)/\sigma_\nu),$$
$$t(x) \equiv F_{\tilde{\varepsilon}}((x'\beta - \mu_\varepsilon)/\sigma_\varepsilon),$$

and then

$$p_{11,x} = C(F_{\tilde{\varepsilon}}(F_{\tilde{\varepsilon}}^{-1}(t(x)) + \delta_1), q(x); \rho),$$
$$p_{10,x} = t(x) - C(t(x), q(x); \rho),$$
$$p_{00,x} = 1 - t(x) - q(x) + C(t(x), q(x); \rho).$$

Similar to the proof strategy for the binary X_1 case, we want to show that, given $(\alpha, \mu_\nu, \sigma_\nu)$, there are two distinct sets of parameter values $(\beta, \delta_1, \rho, \mu_\varepsilon, \sigma_\varepsilon)$ and $(\beta^*, \delta_1^*, \rho^*, \mu_\varepsilon^*, \sigma_\varepsilon^*)$ that generate the same observed fitted probabilities $p_{yd,x}$ for all $(y, d) \in \{0, 1\}^2$ and $x \in \text{supp}(X)$.

Let $t(x) \equiv F_{\tilde{\varepsilon}}(x'\beta) \in (0, 1)$ for all x and for some β. Also, choose $\delta_1 = 0$ and some $\rho \in \Omega$. For $\rho^* > \rho$, we will show that there exists (β^*, δ_1^*) such that, for $t^*(x) \equiv F_{\tilde{\varepsilon}}(x'\beta^*)$,

$$p_{10,x} = t(x) - C(t(x), q(x); \rho) = t^*(x) - C(t^*(x), q(x); \rho^*) \tag{1.2.15}$$
$$p_{11,x} = C(F_{\tilde{\varepsilon}}(F_{\tilde{\varepsilon}}^{-1}(t(x)) + 0), q(x); \rho) = C(s^\dagger(x), q(x); \rho^*) \tag{1.2.16}$$

for all x, where

$$s^\dagger(x) = F_{\tilde{\varepsilon}}(F_{\tilde{\varepsilon}}^{-1}(t^*(x)) + \delta_1^*). \tag{1.2.17}$$

The question is whether we find (β, δ_1, ρ) and $(\beta^*, \delta_1^*, \rho^*)$ such that (1.2.15)–(1.2.17)

simultaneously hold. First note that, since $\rho^* > \rho$, $t^* > t$ and hence $\beta^* \neq \beta$ by the assumption that there is no linear subspace in the space of X. Now as before, take $C(\cdot, \cdot; \rho)$ to be a normal copula and choose $\rho = 0$ and $\rho^* = 1$. Then by arguments similar to the binary case, we obtain

$$t^*(x) = q(x) + (1 - q(x))t(x), \tag{1.2.18}$$

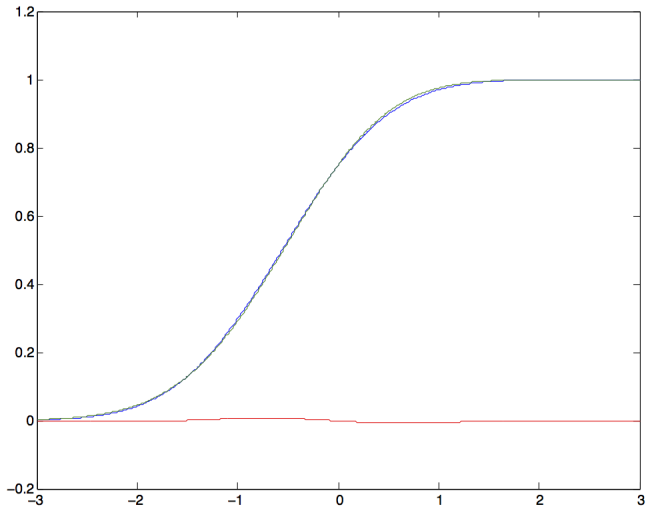
and $s^\dagger(x) = q(x)t(x)$. Then (1.2.17) can be rewritten as

$$\begin{aligned} \delta_1^* &= F_{\tilde{\varepsilon}}^{-1}(s^\dagger(x)) - F_{\tilde{\varepsilon}}^{-1}(t^*(x)) \\ &= F_{\tilde{\varepsilon}}^{-1}(q(x)t(x)) - F_{\tilde{\varepsilon}}^{-1}(q(x) + (1 - q(x))t(x)). \end{aligned} \tag{1.2.19}$$

The complication here is to make this equation satisfied for all x. Note that (1.2.18) and (1.2.19) are consistent with the definition of a distribution function of a continuous r.v.: $F_{\tilde{\varepsilon}}(+\infty) = 1$, $F_{\tilde{\varepsilon}}(-\infty) = 0$, and $F_{\tilde{\varepsilon}}(\varepsilon)$ is strictly increasing. We can then numerically show that a distribution function that is close to a normal distribution satisfies the conditions with a particular choice of (β^*, δ_1^*); see Figure 1.1. This figure compares that distribution function (blue line) to a normal distribution function (green line).

Although, no formal derivation of counterexample is given, this result suggests the following: (i) In the bivariate probit model with continuous common exogenous covariates and no excluded instruments, the parameters will be *at best* weakly identified; (ii) This also implies that the structural parameters and the marginal distributions of the semiparametric model considered in Theorem 1.2.10 are not

Figure 1.1: A Numerical Calculation of a Distribution Function under which Identification Fails



identified without an exclusion restriction even if X_1 has large support.

1.2.2.2 No Restrictions on Dependence Structures

When the restriction imposed on $C(\cdot,\cdot)$ (i.e., Assumption 1.2.6) is completely relaxed, the underlying parameters of model (1.1.1) may fail to be identified whether or not the exclusion restriction holds. That is, a structure on how the unobservables (ε,ν) are dependent to each other is necessary for identification. This is closely related to the results in the literature that the treatment parameters (which is a lower dimensional function of the individual parameters) in triangular models similar to (1.1.1) is only partially identified without distributional assumptions; see Bhattacharya et al. (2008), Chiburis (2010), Shaikh and Vytlacil (2011), and Mourifié (2015).

Suppose Assumptions 1.2.1–1.2.4 hold. Then the model becomes a semiparametric threshold crossing model in that the joint distribution is completely unspecified. Then as a special case of Shaikh and Vytlacil (2011), one can easily derive bounds for the ATE $F_\varepsilon(x'\beta+\delta_1)-F_\varepsilon(x'\beta)$. The sharpness of these bounds is shown in their paper under a rectangular support assumption for (X,Z), which in turn is relaxed in Mourifié (2015). Additionally with Assumption 1.2.7, one can also derive bounds for the individual parameters $x'\beta$ and δ_1, as it is shown in Chiburis (2010). When there is no excluded instruments in the model, Chiburis (2010) shows that the bounds on the ATE do not improve over Manski (1990)'s bounds, which argument applies for the individual parameters.

1.3 Estimation

Let $W_i \equiv (Y_i, D_i, X_i', Z_i')'$ be an observation of individual i and let w be a realization of W_i. We denote the supports of W, ϵ, and ν by $\mathcal{S}_W$, $\mathcal{S}_\epsilon$, and $\mathcal{S}_\nu$, respectively. We assume that the distribution functions F_ϵ and F_ν admit the density functions f_ϵ and f_ν, respectively. Then we can define $\theta \equiv (\psi', f_\epsilon, f_\nu)'$ as the parameter of the model. The parameter space needs to be defined carefully. Since we want the density functions f_ϵ and f_ν to be nonnegative, we define the parameter spaces of f_ϵ and f_ν by using square root density functions. That is, we consider

$$\mathcal{F}_j = \{f = g^2 : g \in \mathcal{F}, \int \{g(x)\}^2 dx = 1\}, \tag{1.3.1}$$

where $j \in \{\epsilon, \nu\}$ and $\mathcal{F}$ is a space of functions, which will be specified later, as the parameter space of f_j . Then we can define $\tilde{\Theta} \equiv \tilde{\Psi} \times \mathcal{F}_\epsilon \times \mathcal{F}_\nu$ as the parameter

space of θ. Note that, by defining θ as the parameter of the model, we can consider $F_\epsilon(\epsilon) = \int_{\mathcal{S}_\epsilon} \mathbf{1}[t \leq \epsilon] f_\epsilon(t)dt$ and $F_\nu(\nu) = \int_{\mathcal{S}_\nu} \mathbf{1}[t \leq \nu] f_\nu(t)dt$ as functionals of θ. To distinguish an element $\theta \in \tilde{\Theta}$ from the true parameter, let $\theta_0 = (\psi_0', f_{\epsilon 0}, f_{\nu 0})' \in \tilde{\Theta}$ be the true parameter.

We adopt the maximum likelihood (ML) method to estimate the parameters in the model. Assuming that the data are i.i.d, we define the conditional density function of (Y_i, D_i) on $(X_i', Z_i')'$ as

$$f(Y_i, D_i | X_i, Z_i; \theta) = \prod_{y,d=0,1} [p_{yd}(X_i, Z_i; \xi)]^{\mathbf{1}\{Y_i = y, D_i = d\}},$$

where $p_{yd}(x, z; \xi)$ abbreviates the right hand side expression that equates $p_{yd,xz}$ in (1.2.2) and $f(y, d|x, z; \theta)$ is the conditional density of (Y_i, D_i) on $(X_i', Z_i') = (x', z')$. Then the log of density $l(\theta, w) \equiv \log f(y, d|x, z; \theta)$ becomes

$$l(\theta, W_i) \equiv \sum_{y,d=0,1} \mathbf{1}_{yd}(Y_i, D_i) \cdot \log p_{yd}(X_i, Z_i; \theta), \tag{1.3.2}$$

where $\mathbf{1}_{yd}(Y_i, D_i) \equiv \mathbf{1}\{Y_i = y, D_i = d\}$. Then the ML estimator of θ_0, $\tilde{\theta}_n$, is define as

$$\tilde{\theta}_n \equiv \arg\max_{\theta \in \tilde{\Theta}} Q_n(\theta), \tag{1.3.3}$$

where $Q_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} l(\theta, W_i)$ is the log-likelihood function.

Since function $l(\theta, W_i)$ contains both finite-dimensional and infinite-dimensional parameters, it is not easy to solve the optimization problem in Equation (1.3.3) without additional information on f_ϵ and f_ν. If the infinite-dimensional parameters f_ϵ and f_ν are fully characterized by finite-dimensional parameters, say $\eta \equiv (\eta_\epsilon', \eta_\nu')' \in$

$H \subset \mathbb{R}^{d_\eta}$ for some integer $d_\eta > 0$, then the estimator $\tilde{\theta}_n$ becomes a standard ML estimator. For example, if one imposes Assumption 1.2.7, then $\eta_\epsilon = (\mu_\epsilon, \sigma_\epsilon)'$ and $\eta_\nu = (\mu_\nu, \sigma_\nu)'$. This parametrization leads us to redefine the parameter θ and the parameter space $\tilde{\Theta}$ as $\theta = (\psi', \eta')' \in \tilde{\Psi} \times H \subset \mathbb{R}^{d_\psi + d_\eta}$ and $\tilde{\Theta} \equiv \tilde{\Psi} \times H$, and the ML estimator $\tilde{\theta}_n$ is obtained by maximizing $Q_n(\theta)$ over the parameter space $\tilde{\Theta} = \tilde{\Psi} \times H$. One can show that the parametric ML estimator $\tilde{\theta}_n$ is consistent, asymptotically normal, and efficient under some regularity conditions, and those conditions are provided by, for example, Newey and McFadden (1994).

Although the parametric ML estimator possesses many desirable properties, the model needs to be correctly specified to guarantee that those properties of the ML estimator hold. Since most of economic theories do not suggest choice of distributions, people have tried to seek for more robust estimation methods to misspecification. In this paper, we adopt the sieve method to estimate the unknown density functions to obtain robustness and flexibility of the model.

Let $\mathcal{F}_{\varepsilon n}$ and $\mathcal{F}_{\nu n}$ be appropriate sieve spaces for $\mathcal{F}_\varepsilon$ and $\mathcal{F}_\nu$, respectively, and let $f_{\epsilon n}(\cdot; a_{\epsilon n})$ and $f_{\nu n}(\cdot; a_{\nu n})$ be the sieve approximations of f_ϵ and f_ν on their sieve spaces $\mathcal{F}_{\epsilon n}$ and $\mathcal{F}_{\nu n}$, respectively. Then we define the sieve ML estimator $\hat{\theta}_n$ as following :

$$\hat{\theta}_n \equiv \arg\max_{\theta \in \tilde{\Theta}_n} Q_n(\theta), \tag{1.3.4}$$

where $\tilde{\Theta}_n \equiv \tilde{\Psi} \times \mathcal{F}_{\epsilon n} \times \mathcal{F}_{\nu n}$.

We also point out that the sieve ML estimator can be equivalently obtained

from the following unconstrained optimization problem:

$$\max_{\psi,a_n} \sum_{i=1}^{n} l(\psi, a_n, W_i) - \lambda_n \text{Pen}(a_n) + \tau_\varepsilon \left\{1 - \int_{\mathcal{S}_\varepsilon} f_\varepsilon(t, a_{\epsilon n}) dt\right\} + \tau_\nu \left\{1 - \int_{\mathcal{S}_\nu} f_\nu(t, a_{\nu n}) dt\right\},$$

where $a_n = (a'_{\epsilon n}, a'_{\nu n})'$, $l(W_i, \psi, a_n)$ is the log likelihood with sieve approximations $f_\varepsilon(\cdot, a_{\epsilon n})$ and $f_\nu(\cdot, a_{\nu n})$, $\text{Pen}(a_n)$ is the penalization term that imposes, for example, the properties of Holder space, and the remaining penalization terms are to impose the properties of a density function. Note that $\tau_\varepsilon > 0$ and $\tau_\nu > 0$.

We are interested in a class of "smooth" univariate densities and focus on approximation of a square root density. Specifically, we assume that $\sqrt{f_\epsilon}$ and $\sqrt{f_\nu}$ belong to the class of *p-smooth* functions [8]and we restrict our attention to the linear sieve spaces for $\mathcal{F}_\epsilon$ and $\mathcal{F}_\nu$. In this case, the choice of sieve spaces for $\mathcal{F}_\epsilon$ and $\mathcal{F}_\nu$ depends on $\mathcal{S}_\epsilon$ and $\mathcal{S}_\nu$, respectively. If the supports are bounded, then one can use the polynomial sieve, the trigonometric sieve, or the cosine sieve. When the supports are unbounded, then we can use the Hermite polynomial sieve or the spline wavelet sieve to approximate a square root density.

We confine our attention to cases where the copula function is correctly specified for establishing the asymptotic theory. Since the copula is specified by some finite-dimensional parameter, the model is vulnerable to misspecification of the copula function. It is well-known that if the density function is misspecified in a ML problem, the ML estimator converges to a pseudo-true value which minimizes the Kullback-Leibler Information Criterion (KLIC) (e.g. White (1982), Chen and Fan

[8]The definition of p-smooth functions can be found on in (Chen, 2007, p.5570) or CFT06 (p.1230). We give the formal definition of p-smooth functions in Section 4.

(2006a) and Chen and Fan (2006b)). We do not pursue investigating the asymptotic properties of the sieve estimators under copula misspecification, but there are several tests that can be useful to check misspecification of the copula in some classes of models. Chen and Fan (2006a) propose a test procedure for model selection, which is based on the test of Vuong (1989). Liao and Shi (2017) extend Vuong's test to the one for models containing infinite dimensional parameters and propose a uniformly asymptotically valid Vuong test for semi/non-parametric models. Their setting encompasses the models that can be estimated by the sieve ML as a special case, so one may refer to the paper for model selection in our context. Even if we assume that the copula function is correctly specified to develop the asymptotic theory, we address the issue on misspecification of the copula in part by conducting some simulations to see how misspecification of copula affects the performance of estimators.

1.4 Asymptotic Theory for Semiparametric Models

In this section, we provide the asymptotic theory for the sieve ML estimator. We slightly modify the model to investigate the asymptotic properties of the sieve M-estimator. Specifically, we consider the following specification:

$$\begin{aligned} F_{\epsilon 0}(x) &= H_{\epsilon 0}(G_\epsilon(x)) \\ F_{\nu 0}(x) &= H_{\nu 0}(G_\nu(x)), \end{aligned} \tag{1.4.1}$$

where $H_{\epsilon 0}(\cdot)$ and $H_{\nu 0}(\cdot)$ are unknown distribution functions on $[0,1]$ and $G_\epsilon(\cdot)$ and $G_\nu(\cdot)$ are known and strictly increasing functions mapping from $\mathbb{R}$ into $[0,1]$. It is

possible that $G_\epsilon(\cdot)$ and $G_\nu(\cdot)$ are different from each other, but this is not crucial when it comes to estimating the parameters in the model as the main difficulty with estimation relies on the unknown functions $H_{\epsilon 0}$ and $H_{\nu 0}$. We assume that $G_\epsilon(\cdot) = G_\nu(\cdot) \equiv G(\cdot)$ to avoid the complexity of notations. The transformation in Equation (1.4.1) can be found in the literature (e.g. Bierens (2014)) and we do not have any loss of generality. Furthermore, the transformation may make it easier to derive the asymptotic properties of the estimator because the unknown infinite-dimensional parameters are defined on a bounded set. For the known distribution function G, we can choose $G(x) \equiv \Phi(x)$ for $x \in \mathbb{R}$, where $\Phi(\cdot)$ is the standard normal distribution function, and assume that $H_{\epsilon 0}(\cdot)$ and $H_{\nu 0}(\cdot)$ have their density functions $h_{\epsilon 0}(\cdot)$ and $h_{\nu 0}(\cdot)$, respectively, on $[0, 1]$. With this modification, we redefine the parameter as $\theta = (\psi', h_{\epsilon 0}, h_{\nu 0})' \in \tilde{\Theta}^\dagger$, where $\tilde{\Theta}^\dagger = \tilde{\Psi} \times \mathcal{H}_\epsilon \times \mathcal{H}_\nu$, and the sieve space becomes $\tilde{\Theta}_n^\dagger = \tilde{\Psi} \times \mathcal{H}_{\epsilon n} \times \mathcal{H}_{\nu n}$.

Let G be a mapping from $\mathbb{R}$ to $[0, 1]$, which is strictly increasing on $\mathbb{R}$. Then one may wonder if there exist $H_{\epsilon 0}$ and $H_{\nu 0}$ satisfying (1.4.1). Since G is assumed to be strictly increasing, there exists its inverse function G^{-1}. Letting $H_{\epsilon 0}(\cdot) = F_{\epsilon 0}(G^{-1}(\cdot))$ and $H_{\nu 0}(\cdot) = F_{\nu 0}(G^{-1}(\cdot))$, it is straightforward to see that $H_{\epsilon 0}$ and $H_{\nu 0}$ are mapping from $[0, 1]$ to $[0, 1]$ and satisfying the relations in (1.4.1). We also note that such a transformation does not change identification results. Since $G(\cdot)$ is strictly increasing on $\mathbb{R}$, it has the inverse function $G^{-1}(\cdot)$. Then it is straightforward to show that, with the transformation given by Equation (1.4.1), $H_0(\cdot) = F_0(G^{-1}(\cdot))$ and thus F_0 is identified on $\mathbb{R}$ if and only if H_0 is identified on $[0, 1]$. Assuming that G is differentiable and that its derivative is bounded away from zero on $\mathbb{R}$ and bounded

above, the unknown density function h_0 can be written as $h_0(x) = \frac{f_0(G^{-1}(x))}{g(G^{-1}(x))}$, where $g(x) = \frac{dG(x)}{dx}$. This expression draws the conclusion that f_0 is identified if any only if $h_0(x)$ is identified. Hence, we can conclude that $h_{\epsilon 0}$ and $h_{\nu 0}$ are identified if and only if the unknown marginal density functions $f_{\epsilon 0}$ and $f_{\nu 0}$ are identified and G admits the density g on $\mathbb{R}$. Therefore, we choose G such that G is differentiable and that the derivative, denoted by g, is bounded away from zero on $\mathbb{R}$. It is clear that using Φ as G satisfies those requirements.

1.4.1 Consistency of the Sieve MLE

The consistency of the sieve ML estimator has been established in several papers (e.g. Geman and Hwang (1982); Gallant and Nychka (1987); White and Wooldridge (1991); Bierens (2014)). Chen (2007) provides sufficient conditions under which the sieve M-estimator is consistent, and we establish the consistency by verifying the conditions in Theorem 3.1 in Chen (2007).

We redefine the parameter space to facilitate developing the asymptotic theory. The identification requires the space of the finite-dimensional parameter $\tilde{\Psi}$ to be open and convex (see Theorems 1.2.8 and 1.2.10), and thus $\tilde{\Psi}$ cannot be compact. We introduce an "optimization space" which contains the true parameter ψ_0 and consider it as the parameter space of ψ. Formally, we restrict the parameter space for estimation in the following way.

Assumption 1.4.1. *There exists a compact and convex subset $\Psi \subseteq \tilde{\Psi}$ such that $\psi_0 \in int(\Psi)$, where $int(A)$ is the interior of a set A.*

With the optimization space, we define the parameter space as $\Theta \equiv \Psi \times \mathcal{H}_\epsilon \times$

$\mathcal{H}_\nu$ and the corresponding sieve space is denoted by $\Theta_n \equiv \Psi \times \mathcal{H}_{\epsilon n} \times \mathcal{H}_{\nu n}$. Then the sieve ML estimator in Equation (1.3.4) is also redefined as following :

$$\hat{\theta}_n \equiv \arg\max_{\theta \in \Theta_n} Q_n(\theta) \tag{1.4.2}$$

Define $Q_0(\theta) \equiv \mathbb{E}[l(\theta, W_i)]$ and let $||\cdot||_c$ be a norm on Θ, whose the form is of $||\theta||_c \equiv ||\psi||_E + ||h_\epsilon||_{\mathcal{H}_\epsilon} + ||h_\nu||_{\mathcal{H}_\nu}$, where $||\cdot||_E$ is the Euclidean norm and $||\cdot||_{\mathcal{H}_\epsilon}$ and $||\cdot||_{\mathcal{H}_\nu}$ are norms on $\mathcal{H}_\epsilon$ and $\mathcal{H}_\nu$, respectively. Let $d_c(\cdot,\cdot) : \Theta \times \Theta \to [0, \infty)$ be a pseudo metric induced by the norm $||\cdot||_c$.

We introduce some classes of functions to define the parameter space. Let $\mathcal{C}^m(\mathcal{X})$ be the space of m-times continuously differentiable real-valued functions on $\mathcal{X}$. Let $\zeta \in (0, 1]$ and, given a d-tuple ω, let $[\omega] = \omega_1 + ... + \omega_d$. Denote the differential operator by $\mathcal{D}$ and let $\mathcal{D}^\omega = \frac{\partial^{[\omega]}}{\partial x_1^{\omega_1} ... \partial x_d^{\omega_d}}$. Letting $p = m + \zeta$, we define the Hölder norm for $h \in \mathcal{C}^m(\mathcal{X})$ as following :

$$||h||_{\Lambda^p} \equiv \sup_{[\omega] \leq m, x} |\mathcal{D}^\omega h(x)| + \sup_{[\omega]=m} \sup_{x,y \in \mathcal{X}, ||x-y||_E \neq 0} \frac{|\mathcal{D}^\omega h(x) - \mathcal{D}^\omega h(y)|}{||x - y||_E^\zeta} < \infty,$$

where ζ is the Hölder exponent. We define a Hölder class as $\Lambda^p(\mathcal{X}) \equiv \{h \in \mathcal{C}^m(\mathcal{X}) : ||h||_{\Lambda^p} < \infty\}$. A Hölder ball with radius R, $\Lambda_R^p(\mathcal{X})$, is defined as $\Lambda_R^p(\mathcal{X}) \equiv \{h \in \Lambda^p(\mathcal{X}) : ||h||_{\Lambda^p} \leq R < \infty\}$.

We first need to choose the norms $||\cdot||_{\mathcal{H}_\epsilon}$ and $||\cdot||_{\mathcal{H}_\nu}$ on $\mathcal{H}_\epsilon$ and $\mathcal{H}_\nu$, respectively, to prove the consistency. It is important to choose appropriate norms to ensure compactness of the original parameter space as compactness plays an important role in establishing the asymptotic theory. Since the parameter space contains infinite

dimensional spaces, the parameter space may be compact under certain norms and may not be compact under other norms. Since closedness and boundedness of an infinite dimensional space are no longer equivalent to compactness, it is much harder to show that the parameter space is compact under certain norms. To overcome this difficulty, we take the approach introduced by Gallant and Nychka (1987), which uses two norms to obtain the consistency. Their idea is to use the strong norm to define the parameter space as a ball and then obtain compactness of the parameter space by equipping another norm, the consistency norm. Freyberger and Masten (2015) recently extend the idea to more cases and present compactness results for several parameter spaces. Note that, using the transformation of the distribution functions in Equation (1.4.1), the unknown infinite dimensional parameters are defined on bounded domains.

We present assumptions under which the sieve ML estimator in Equation (1.4.2) is consistent with respect to some pseudo-metric $d_c(\cdot,\cdot)$.

Assumption 1.4.2. *There exists a measurable function* $\underline{p}(X,Z)$ *such that for all* $\theta \in \Theta$ *and for all* $y,d = 0,1$, $p_{yd,XZ}(\theta) \geq \underline{p}(X,Z)$ *with* $\mathbb{E}|\log(\underline{p}(X,Z))| < \infty$ *and* $\mathbb{E}[\frac{1}{\underline{p}(X,Z)^2}] < \infty$.

Assumption 1.4.3. $(W_i)_{i=1}^n$ *are i.i.d. and* $\mathbb{E}[||(X_i', Z_i')'||_E^2] < \infty$.

Assumption 1.4.4. *(i)* $\sqrt{h_{\epsilon 0}}, \sqrt{h_{\nu 0}} \in \Lambda_R^p([0,1])$ *with* $p > \frac{1}{2}$ *and some* $R > 0$*; (ii)* $\mathcal{H} = \{h = b^2 : b \in \Lambda_R^p([0,1]), \int_0^1 h = 1\}$*, where* R *is the same to the one in (i), and* $\mathcal{H}_\epsilon = \mathcal{H}_\nu = \mathcal{H}$*; (iii) the density functions* $h_{\epsilon 0}$ *and* $h_{\nu 0}$ *are bounded away from zero on* $[0,1]$*; (iv) For* $h \in \mathcal{H}$,$||h||_{\mathcal{H}} \equiv \sup_{x\in[0,1]} |h(x)|$*, denoted by* $||h||_\infty$.

Assumption 1.4.5. *(i)* $\mathcal{H}_{\epsilon n} = \mathcal{H}_{\nu n} = \{h \in \mathcal{H} : h(x) = p^{k_n}(x)' a_{k_n}, a_{k_n} \in \mathbb{R}^{k_n}, ||h||_\infty < 2R^2\}$, *where* $k_n \to \infty$ *and* $k_n/n \to 0$ *as* $n \to \infty$; *(ii) For all* $j \geq 1$, *we have* $\Theta_j \subseteq \Theta_{j+1}$ *and there exists sequence* $\{\pi_j \theta_0\}_j$ *such that* $d_c(\pi_j \theta_0, \theta_0) \to 0$ *as* $j \to \infty$.

Assumption 1.4.6. *For* $j = 1, 2$, *denote* $C_j(u_1, u_2; \rho) \equiv \frac{\partial C(u_1, u_2; \rho)}{\partial u_j}$ *and* $C_\rho(u_1, u_2; \rho) \equiv \frac{\partial C(u_1, u_2; \rho)}{\partial \rho}$. *The derivatives* $C_j(\cdot, \cdot; \cdot)$ *and* $C_\rho(\cdot, \cdot; \cdot)$ *are uniformly bounded for all* $j = 1, 2$.

Assumption 1.4.2 guarantees that the log-likelihood function $l(\theta, W_i)$ is well-defined for all $\theta \in \Theta$ and that $Q_0(\theta_0) > -\infty$. Assumption 1.4.3 restricts the data generating process and assumes existence of moments of the data. Assumption 1.4.4 defines the parameter space and implies that the infinite dimensional parameters are in some smooth class. Note that the conditions (i) and (ii) in Assumption 1.4.4 together imply that $h_{\epsilon 0}$ and $h_{\nu 0}$ belong to $\Lambda^p_{\tilde{R}}([0,1])$ where $\tilde{R} \equiv 2^{m+1} R^2 < \infty$[9]. Thus, we may assume that $h_{\epsilon 0}$ and $h_{\nu 0}$ belong to a Hölder ball with smoothness p under Assumption 1.4.4. While the condition (i) implicitly defines the strong norm (Hölder norm), Assumption 1.4.4-(iv) defines the sup-norm as the weak norm (consistency norm). Note that since the parameter space for the finite-dimensional parameter ψ, Ψ, is assumed to be compact in Assumption 1.4.1, the whole parameter space Θ is compact under the $||\cdot||_c$ by Theorems 1 and 2 in Freyberger and Masten (2015). The first part of Assumption 1.4.5 restricts our choice of sieve spaces for $\mathcal{H}_\epsilon$ and $\mathcal{H}_\nu$ to be among linear sieve spaces with order k_n, and this can be relaxed so that the choice of k_n is different for h_ϵ and h_ν. The latter part of Assumption 1.4.5 requires that

[9] See Appendix A for details.

the sieve space should be appropriately chosen so that the unknown parameters can be well-approximated. Since the unknown infinite-dimensional parameters belong to a Hölder ball and they are defined on bounded supports, one may choose the polynomial sieve, the trigonometric sieve, the cosine sieve, or the spline sieve [10]. For example, if we choose the polynomial sieve or the spline sieve, then one can show that $d_c(\pi_{k_n}\theta_0, \theta_0) = O(k_n^{-p})$ (e.g. Lorentz (1966)). Assumption 1.4.6 imposes boundedness of the derivatives of the copula function.

The following theorem demonstrates that under the assumptions above, the sieve estimator $\hat{\theta}_n$ is consistent with respect to the pseudo metric d_c.

Theorem 1.4.7. *Suppose that Assumptions 1.2.1-1.2.6 and 1.2.9 hold. If Assumptions 1.4.1-1.4.6 are satisfied, then* $d_c(\hat{\theta}_n, \theta_0) \xrightarrow{p} 0$.

1.4.2 Convergence Rates

The convergence rate is one of objects of interest in the semiparametric or nonparametric estimation by itself. More importantly, the convergence rate plays an important role in deriving the asymptotic normality. To be more specific, the convergence rate needs to be fast enough to establish the asymptotic normality. Therefore, in this section we derive the convergence rate of the sieve ML estimator with respect to a certain norm. The convergence rate of sieve M-estimators has been studied by, for example, Shen and Wong (1994); Chen and Shen (1998), and Chen (2007). Unlike that we use a sup-norm type pseudo-metric to show consistency, we

[10] Refer to Chen (2007) or CFT06 for more details on choice of sieve spaces.

establish the convergence rate with respect to a L^2-type norm given below:

$$||\theta - \theta_0||_2 \equiv ||\psi - \psi_0||_E + ||h_\epsilon - h_{\epsilon 0}||_2 + ||h_\nu - h_{\nu 0}||_2, \tag{1.4.3}$$

where $||h - \tilde{h}||_2^2 \equiv \int_0^1 (h(t) - \tilde{h}(t))^2 dt$ for any $h, \tilde{h} \in \mathcal{H}$. It is straightforward to show that $||\theta - \theta_0||_2 \leq d_c(\theta, \theta_0)$. To establish the convergence rate with respect to the norm $||\cdot||_2$, we consider the assumption which imposes the equivalence between $K(\cdot, \cdot)$ and $||\cdot||_2^2$, where $K(\theta_0, \theta)$ is the Kullback-Leibler information.

Assumption 1.4.8. *Let* $K(\theta_0, \theta) \equiv \mathbb{E}[l(\theta_0, W_i) - l(\theta, W_i)]$. *Then there exist* $B_1, B_2 > 0$ *such that*

$$B_1 K(\theta_0, \theta) \leq ||\theta - \theta_0||_2^2 \leq B_2 K(\theta_0, \theta)$$

for all $\theta \in \Theta_n$ *with* $d_c(\theta, \theta_0) = o(1)$.

Assumption 1.4.8 implies that the norm $||\cdot||_2$ and the square-root of the Kullback-Leibler information are equivalent. The next theorem demonstrates the convergence rate of the sieve ML estimator with respect to the norm $||\cdot||_2$.

Theorem 1.4.9. *Suppose that Assumptions 1.2.1-1.2.6 and 1.2.9-1.4.8 hold. Then we have*

$$||\hat{\theta}_n - \theta_0||_2 = O_p(\max\{\sqrt{\frac{k_n}{n}}, k_n^{-p}\}). \tag{1.4.4}$$

Furthermore, if we choose $k_n \propto n^{\frac{1}{2p+1}}$, *then we have*

$$||\hat{\theta}_n - \theta_0||_2 = O_p(n^{-\frac{p}{2p+1}}).$$

The convergence rate given in (1.4.4) depends on two components. The first

component is related to the convergence rate of the "variance term", and this rate increases as the complexity of the sieve space, k_n, becomes higher. In contrast, the latter component, which reflects the convergence rate of the deterministic approximation error $||\theta_0 - \pi_k \theta_0||_2$, decreases as k_n becomes larger. The choice of $k_n \propto n^{\frac{1}{2p+1}}$ yields the best convergence rate, and we can see that with this choice, the convergence rate of the sieve estimator $\hat{\theta}_n$ becomes faster as the degree of the smoothness, p, increases.

1.4.3 Asymptotic Normality a Smooth Functional

Once the parameters of the model are estimated, it is important to find out the asymptotic distribution of the parameters to conduct statistical inference. Since the parameters in our model consist of both finite and infinite dimensional parameters and many objects of interest in inference are considered as a functional of the parameters, we focus on establishing the asymptotic distribution of functionals rather than the parameters themselves.

In the literature, $\sqrt{n}$-estimable functionals are called *regular* functionals and functional slower than $\sqrt{n}$-estimable are referred to as *irregular* functionals. While the asymptotic distribution of a class of regular functionals has been established in the sieve M-estimation literature (e.g. Chen and Shen (1998), CFT06, Bierens (2014)), there are few studies on the asymptotic theory on irregular functionals[11]. Since the class of smooth functionals encompasses a large class of objects of interest,

[11] See Chen et al. (2014) or Chen and Pouzo (2015) for inference for the irregular functionals based on the sieve methods.

we restrict our attention to a class of smooth functionals.

Let $T : \Theta \to \mathbb{R}$ be a functional and define $\mathbb{V}$ as the linear span of $\Theta - \{\theta_0\}$. We also let $r_{10} = F_{\epsilon 0}(x^{'}\beta_0 + \delta_{10})$, $r_{00} = F_{\epsilon 0}(x^{'}\beta_0)$, and $s_0 = F_{\nu 0}(x^{'}\alpha_0 + z^{'}\gamma_0)$. For $t \in [0,1]$, define the directional derivative of $l(\theta, W)$ at the direction $v \in \mathbb{V}$ as

$$\begin{aligned}
\frac{dl(\theta_0 + tv, W)}{dt}|_{t=0} &= \lim_{t \to 0} \frac{l(\theta_0 + tv, W) - l(\theta_0)}{t} \\
&= \frac{\partial l(\theta_0, W)}{\partial \psi'}[v_\psi] + \sum_{j \in \{\epsilon, \nu\}} \frac{\partial l(\theta_0, W)}{\partial h_j}[v_j] \\
&= \frac{\partial l(\theta_0, W)}{\partial \psi'} v_\psi + \sum_{j \in \{\epsilon, \nu\}} \frac{\partial l(\theta_0, W)}{\partial h_j}[v_j], \qquad (1.4.5)
\end{aligned}$$

where for $v = (v_\psi^{'}, v_\epsilon, v_\nu)^{'}$,

$$\frac{\partial l(\theta_0, w)}{\partial \psi'} v_\psi = \sum_{\tilde{y}, \tilde{d} \in \{0,1\}} (\mathbf{1}_{\tilde{y}, \tilde{d}} \cdot \frac{1}{p_{\tilde{y}\tilde{d}, xz}(\theta_0)} \cdot \frac{\partial p_{\tilde{y}\tilde{d}, xz}(\theta_0)}{\partial \psi'}) v_\psi,$$

$$\begin{aligned}
\frac{\partial l(\theta_0, w)}{\partial h_\epsilon}[v_\epsilon] = {} & \mathbf{1}_{11}(y,d) \times [\frac{1}{p_{11,xz}(\theta_0)} C_1(r_{10}, s_0; \rho_0) \int_0^{G(x^{'}\beta_0 + \delta_0)} v_\epsilon(t) dt]] \\
& + \mathbf{1}_{10}(y,d) \times [\frac{1}{p_{10,xz}(\theta_0)} [(1 - C_1(r_{00}, s_0; \rho_0)) \int_0^{G(x^{'}\beta_0)} v_\epsilon(t) dt]] \\
& + \mathbf{1}_{01}(y,d) \times [\frac{1}{p_{01,xz}(\theta_0)} [-C_1(r_{10}, s_0; \rho_0) \int_0^{G(x^{'}\beta_0 + \delta_0)} v_\epsilon(t) dt]] \\
& + \mathbf{1}_{00}(y,d) \times [\frac{1}{p_{00,xz}(\theta_0)} [(1 - C_1(r_{00}, s_0; \rho_0)) \int_0^{G(x^{'}\beta_0)} v_\epsilon(t) dt]],
\end{aligned}$$

and

$$\frac{\partial l(\theta_0, w)}{\partial h_\nu}[v_\nu] = \{\frac{\mathbf{1}_{11}(y,d)}{p_{11,xz}(\theta_0)} C_2(r_{10}, s_0; \rho_0) + \frac{\mathbf{1}_{10}(y,d)}{p_{10,xz}(\theta_0)}(-C_2(r_{00}, s_0; \rho_0))$$
$$+ \frac{\mathbf{1}_{01}(y,d)}{p_{01,xz}(\theta_0)}(1 - C_2(r_{10}, s_0; \rho_0)) + \frac{\mathbf{1}_{00}(y,d)}{p_{00,xz}(\theta_0)}(1 - C_2(r_{00}, s_0; \rho_0))\}$$
$$\times \int_0^{G(x'\alpha_0 + z'\gamma_0)} v_\nu(t)dt.$$

Before presenting results on the asymptotic normality of smooth functionals, we strengthen the smoothness condition in Assumptions 1.2.6 and 1.4.6. We let $C_{ij}(u_1, u_2; \rho)$ denote the second-order partial derivative of a copula function $C(u_1, u_2; \rho)$ w.r.t. i and j for $i, j \in \{u_1, u_2, \rho\}$.

Assumption 1.4.10. *The copula function $C(u_1, u_2; \rho)$ is twice continuously differentiable with respect to u_1, u_2, and ρ and its first- and second- order partial derivatives are well-defined in a neighborhood of θ_0.*

Define the Fisher inner product on the space $\mathbb{V}$ as

$$< v, \tilde{v} > \equiv E[(\frac{\partial l(\theta_0, W)}{\partial \theta}[v])(\frac{\partial l(\theta_0, W)}{\partial \theta}[\tilde{v}])] \tag{1.4.6}$$

and the Fisher norm for $v \in \mathbb{V}$ as $||v||^2 =< v, v >$. If we let $\bar{\mathbb{V}}$ be the closed linear span of $\mathbb{V}$ under the Fisher norm, then $(\bar{\mathbb{V}}, ||\cdot||)$ is a Hilbert space as CFT06 demonstrated.

For the functional T and for any $v \in \mathbb{V}$, we denote

$$\frac{\partial T(\theta_0)}{\partial \theta'}[v] \equiv \lim_{t \to 0} \frac{T(\theta_0 + tv) - T(\theta_0)}{t}.$$

Note that for any $\theta_1, \theta_2 \in \Theta$, we have

$$\begin{aligned}||\theta_1 - \theta_2||^2 &= \mathbb{E}(\frac{\partial l(\theta_0, W_i)}{\partial \theta}[\theta_1 - \theta_2])^2 \\ &\leq B\{\mathbb{E}[\frac{\partial l(\theta_0, W_i)}{\partial \psi'}(\psi_1 - \psi_2)]^2 + \mathbb{E}[\frac{\partial l(\theta_0, W_i)}{\partial h_\epsilon}[h_{\epsilon 1} - h_{\epsilon 2}]]^2 \\ &+ \mathbb{E}[\frac{\partial l(\theta_0, W_i)}{\partial h_\nu}[h_{\nu 1} - h_{\nu 2}]]^2\} \\ &\leq B||\theta_1 - \theta_2||_2^2 \end{aligned} \tag{1.4.7}$$

for some $B > 0$ under Assumptions 1.4.3, 1.4.4, and 1.4.6. This implies that we can use the convergence rate of the sieve estimator $\hat{\theta}_n$ w.r.t. the norm $||\cdot||_2$ for the one w.r.t. the norm $||\cdot||$.

Assumption 1.4.11. *The following conditions hold:*

(i) there exist a constants $w > 1 + \frac{1}{2p}$ and a small $\epsilon_0 > 0$ such that for any $v \in \mathbb{V}$ with $||v|| \leq \epsilon_0$,

$$|T(\theta_0 + v) - T(\theta_0) - \frac{\partial T(\theta_0)}{\partial \theta'}[v]| = O(||v||^w);$$

(ii) For any $v \in \mathbb{V}$, $T(\theta_0 + tv)$ is continuously differentiable in $t \in [0, 1]$ around $t = 0$, and

$$||\frac{\partial T(\theta_0)}{\partial \theta'}|| \equiv \sup_{v \in \mathbb{V}, ||v|| > 0} \frac{|\frac{\partial T(\theta_0)}{\partial \theta'}[v]|}{||v||} < \infty.$$

Assumption 1.4.11 defines a smooth functional T and guarantees the existence of $v^* \in \bar{\mathbb{V}}$ such that $< v^*, v >= \frac{\partial T(\theta_0)}{\partial \theta'}[v]$ for all $v \in \mathbb{V}$ and $||v^*||^2 = ||\frac{\partial T(\theta_0)}{\partial \theta'}||^2$, and we call v^* the Riesz representer for the functional T. The next assumption requires the Riesz representer be well-approximated over the sieve space and converge

at a rate with respect to the Fisher norm.

Assumption 1.4.12. *There exists* $\pi_n v^* \in \Theta_n - \{\theta_0\}$ *such that* $||\pi_n v^* - v^*|| = o(n^{-1/4})$.

We derive the asymptotic normality of a smooth functional T by modifying the conditions in CFT06. Let $\mu_n(g) = \frac{1}{n}\sum_{i=1}^n \{g(W_i) - \mathbb{E}[g(W_i)]\}$ be the empirical process indexed by g. We denote the convergence rate of the sieve estimator by δ_n (i.e. $||\hat{\theta}_n - \theta_0|| = O_p(\delta_n)$).

Assumption 1.4.13. *There exist* $\xi_1 > 0$ *and* $\xi_2 > 0$ *with* $2\xi_1 + \xi_2 < 1$ *and a constant* K *such that* $(\delta_n)^{3-(2\xi_1+\xi_2)} = o(n^{-1})$, *and the followings hold for all* $\tilde{\theta} \in \Theta_n$ *with* $||\tilde{\theta} - \theta_0|| \leq \delta_n$ *and all* $v \in \mathbb{V}$ *with* $||v|| \leq \delta_n$ *:*

(i) $|\mathbb{E}[\frac{\partial^2 l(\tilde{\theta},W)}{\partial\psi\partial\psi'} - \frac{\partial^2 l(\theta_0,W)}{\partial\psi\partial\psi'}]| < K||\tilde{\theta} - \theta_0||^{1-\xi_2}$;

(ii) $|\mathbb{E}[\sum_{j\in\{\epsilon,\nu\}}\{\frac{\partial^2 l(\tilde{\theta},W)}{\partial\psi\partial h_j}[v_j] - \frac{\partial^2 l(\theta_0,W)}{\partial\psi\partial h_j}[v_j]\}| \leq K||v||^{1-\xi_1}||\tilde{\theta} - \theta_0||^{1-\xi_2}$;

(iii) $|\mathbb{E}[\sum_{i,j\in\{\epsilon,\nu\}}\{\frac{\partial^2 l(\tilde{\theta},W)}{\partial h_i\partial h_j}[v,v] - \frac{\partial^2 l(\theta_0,W)}{\partial h_i\partial h_j}[v,v]\}] \leq K||v||^{2(1-\xi_1)}||\tilde{\theta} - \theta_0||^{1-\xi_2}$.

Assumption 1.4.14. *The followings hold:*

(i) $\sup_{\theta\in\Theta_n:||\theta-\theta_0||=O(\delta_n)} \mu_n(\frac{\partial l(\theta,W)}{\partial\psi'} - \frac{\partial l(\theta_0,W)}{\partial\psi'}) = o_p(n^{-\frac{1}{2}})$;

(ii) For all $j \in \{\epsilon, \nu\}$,

$$\sup_{\theta\in\Theta_n:||\theta-\theta_0||=O(\delta_n)} \mu_n(\frac{\partial l(\theta,W)}{\partial h_j}[\pi_n v_j^*] - \frac{\partial l(\theta_0,W)}{\partial h_j}[\pi_n v_j^*]) = o_p(n^{-\frac{1}{2}}).$$

Assumptions 1.4.13 and 1.4.14 are modifications of Assumptions 5 and 6 in CFT06, which are needed to control for the second-order expansion of the log-likelihood function $l(\theta, W)$. Under Assumption 1.4.10, these conditions require the

unknown marginal density functions to be smooth enough. For example, the sieve estimator needs to converge at a faster rate than $1/(3-(2\xi_1+\xi_2))$ to satisfy $(\delta_n)^{3-(2\xi_1+\xi_2)} = o(n^{-1})$. Usually, the convergence rate positively depends on the smoothness parameter p in Assumption 1.4.4 and thus the class of models should be restricted to one whose density functions are smooth enough.

Proposition 1.4.1. *Suppose that Assumptions 1.2.1-1.2.6 and 1.2.9-1.4.14 are satisfied. If $k_n \propto n^{\frac{1}{2p+1}}$, then we have*

$$\sqrt{n}(T(\hat{\theta}_n) - T(\theta_0)) \xrightarrow{d} N(0, ||\frac{\partial T(\theta_0)}{\partial \theta'}||^2).$$

1.4.3.1 Asymptotic normality of $\hat{\psi}_n$

In many cases, the finite-dimensional parameter ψ_0 is the parameter of interest and we demonstrate the asymptotic normality of the sieve estimators of the finite-dimensional parameter ψ_0. For any arbitrary $\lambda \in \mathbb{R}^{d_\psi}$ with $|\lambda| \in (0, \infty)$, let $T : \Theta \to \mathbb{R}$ be a functional of the form $T(\theta) = \lambda' \psi$. Then we have for any $v \in \mathbb{V}$,

$$\frac{\partial T(\theta_0)}{\partial \theta}[v] = \lambda' v_\psi \tag{1.4.8}$$

and that there exist a small $\eta > 0$ such that $||v|| \leq \eta$ and a constant $\tilde{c} > 0$ such that

$$|T(\theta_0 + v) - T(\theta_0) - \frac{\partial T(\theta_0)}{\partial \theta}| \leq \tilde{c}||v||^w \tag{1.4.9}$$

with $w = \infty$. In addition, we have

$$\begin{aligned}\sup_{v\in\mathbb{V}:||v||>0}\frac{|\lambda^{'}v_{\psi}|^2}{||v||^2} &= \sup_{v\in\mathbb{V}:||v||>0}\frac{|\lambda^{'}v_{\psi}|^2}{\mathbb{E}[(\frac{\partial l(\theta_0,W)}{\partial\psi^{'}}v_{\psi}+\sum_{j\in\{\epsilon,\nu\}}\frac{\partial l(\theta_0,W)}{\partial h_j}[v_j])^2]}\\ &= \lambda^{'}\mathcal{I}_*(\theta_0)^{-1}\lambda\\ &= \lambda^{'}\mathbb{E}[\mathcal{S}_{\psi_0}\mathcal{S}_{\psi_0}^{'}]^{-1}\lambda,\end{aligned}$$

where

$$\mathcal{S}_{\psi_0}^{'} = \frac{\partial l(\theta_0,W)}{\partial\psi^{'}} - (\frac{\partial l(\theta_0,W)}{\partial h_\epsilon}[b_\epsilon^*] + \frac{\partial l(\theta_0,W)}{\partial h_\nu}[b_\nu^*]), \tag{1.4.10}$$

$b_\epsilon^* = (b_{\epsilon 1}^*,...,b_{\epsilon d_\psi}^*) \in \Pi_{k=1}^{d_\psi}(\mathcal{H}_\epsilon - \{h_{\epsilon 0}\})$, and $b_\nu^* = (b_{\nu 1}^*,...,b_{\nu d_\psi}^*) \in \Pi_{k=1}^{d_\psi}(\mathcal{H}_\nu - \{h_{\nu 0}\})$ are the solutions to the following optimization problems for $k = 1, 2, ..., d_\psi$,

$$\inf_{(b_{\epsilon k},b_{\nu k})\in\bar{\mathbb{V}}_\epsilon\times\bar{\mathbb{V}}_\nu}\mathbb{E}[(\frac{\partial l(\xi_0,W)}{\partial\theta_k} - (\frac{\partial l(\xi_0,W)}{\partial h_\epsilon}[b_{\epsilon k}] + \frac{\partial l(\xi_0,W)}{\partial h_\nu}[b_{\nu k}]))^2].$$

Since the Riesz representer v^* exists if and only if $\mathbb{E}[\mathcal{S}_{\psi_0}\mathcal{S}_{\psi_0}^{'}] = \mathcal{I}_*(\psi_0)$ is non-singular, we impose the following assumption.

Assumption 1.4.15. $\mathbb{E}[\mathcal{S}_{\psi_0}\mathcal{S}_{\psi_0}^{'}]$ *is non-singular.*

Theorem 1.4.16. *Suppose that Assumptions 1.2.1-1.2.6, 1.2.9-1.4.10, 1.4.11-(iii), and 1.4.13-1.4.15 hold. Then we have*

$$\sqrt{n}(\hat{\psi}_n - \psi_0) \xrightarrow{d} N(0, \mathcal{I}_*(\psi_0)^{-1}). \tag{1.4.11}$$

The covariance matrix in Equation (1.4.11) needs to be estimated, and CFT06 adopt the covariance estimation established in Ai and Chen (2003). Since an infinite-dimensional optimization is involved in calculating $\mathcal{S}_{\psi_0}$, we provide a sieve

estimator of $\mathcal{I}_*(\theta_0)^{-1}$ and the sieve spaces for b_ϵ and b_ν are the same to the ones for h_ϵ and h_ν, respectively. By the same way in Ai and Chen (2003), we first estimate b_j^*'s in Equation (1.4.10) by solving the following minimization problem : for all $k = 1, 2, ..., d_\psi$,

$$(\hat{b}_{\epsilon k}, \hat{b}_{\nu k}) \equiv \arg \min_{(b_{\epsilon k}, b_{\nu k}) \in \mathcal{H}_{\epsilon n} \times \mathcal{H}_{\nu n}} \frac{1}{n} \sum_{i=1}^{n} [(\frac{\partial l(\hat{\theta}_n, W_i)}{\partial \psi_k} - (\frac{\partial l(\hat{\theta}_n, W_i)}{\partial h_\epsilon}[b_{\epsilon k}] + \frac{\partial l(\hat{\theta}_n, W_i)}{\partial h_\nu}[b_{\nu k}]))^2].$$

Let $\hat{b}_j = (\hat{b}_{j1}, \hat{b}_{j2}, ..., \hat{b}_{jd_\psi})'$ for given $j \in \{\epsilon, \nu\}$, then we compute

$$\begin{aligned}\hat{\mathcal{I}}_*(\hat{\psi}_n) =& \frac{1}{n} \sum_{i=1}^{n} \{[\frac{\partial l(\hat{\theta}_n, W_i)}{\partial \psi} - (\frac{\partial l(\hat{\theta}_n, W_i)}{\partial h_\epsilon}[\hat{b}_\epsilon] + \frac{\partial l(\hat{\theta}_n, W_i)}{\partial h_\nu}[\hat{b}_\nu])] \\ & \times [\frac{\partial l(\hat{\theta}_n, W_i)}{\partial \psi} - (\frac{\partial l(\hat{\theta}_n, W_i)}{\partial h_\epsilon}[\hat{b}_\epsilon] + \frac{\partial l(\hat{\theta}_n, W_i)}{\partial h_\nu}[\hat{b}_\nu])]' \}\end{aligned}$$

to obtain a consistent estimator of $\mathcal{I}_*(\psi_0)$. We illustrate the following result and the proof can be found in Theorem 5.1 in Ai and Chen (2003).

Theorem 1.4.17. *Suppose that Assumptions in Theorem 1.4.16 hold. Then* $\hat{\mathcal{I}}_*(\hat{\psi}_n) = \mathcal{I}_*(\psi_0) + o_p(1)$.

1.4.3.2 Asymptotic normality of $\hat{\psi}_n$ when the unknown marginals are equal

CFT06 consider the case where the unknown marginal distributions are the same. Let $h_{\epsilon 0} = h_{\nu 0} = h_0 \in \mathcal{H}$ and H_0 is the distribution function which has the density h_0. With the Fisher norm defined by Equation (1.4.6), we can show that

$$\frac{\partial l(\theta_0, W)}{\partial \theta}[v] = \frac{\partial l(\theta_0, W)}{\partial \psi'} v_\psi + \frac{\partial l(\theta_0, W)}{\partial h}[v_h],$$

where $\frac{\partial l(\theta_0,W)}{\partial h}[v_h] = (\frac{\partial l(\theta_0,W)}{\partial h_\epsilon}[v_h] + \frac{\partial l(\theta_0,W)}{\partial h_\nu}[v_h])|_{h_{\epsilon 0}=h_{\nu 0}=h_0}$. We can obtain the asymptotic distribution of $\hat{\psi}_n$ with the following one :

Assumption 1.4.18. *$E[\tilde{\mathbb{S}}_\psi \tilde{\mathbb{S}}'_{\psi_0}]$ is non-singular, where*

$$\tilde{\mathbb{S}}_{\psi_0} = \inf_{b_h \in \Pi_{k=1}^{d_\psi} \bar{\mathbb{V}}_h} \{(\frac{\partial l(\theta_0,W)}{\partial \psi'} - \frac{\partial l(\theta_0,W)}{\partial h}[b_h])'(\frac{\partial l(\theta_0,W)}{\partial \psi'} - \frac{\partial l(\theta_0,W)}{\partial h}[b_h])\} \quad (1.4.12)$$

and $\bar{\mathbb{V}}_h = \bar{\mathbb{V}}_\epsilon = \bar{\mathbb{V}}_\nu$.

We present the asymptotic normality of $\hat{\psi}_n$ under the assumption of the same marginal distributions in the following theorem.

Theorem 1.4.19. *Suppose that the conditions in Theorem 1.4.16 are satisfied. If Assumption 1.4.18 hold and the unknown marginal distributions $H_{\epsilon 0}$ and $H_{\nu 0}$ are equal, then*

$$\sqrt{n}(\hat{\psi}_n - \psi_0) \Rightarrow N(0, \tilde{\mathbb{J}}_*(\psi_0)^{-1}),$$

where $\tilde{\mathbb{J}}_*(\psi_0)^{-1} \equiv E[\tilde{\mathbb{S}}_{\psi_0}\tilde{\mathbb{S}}'_{\psi_0}]^{-1}$. Furthermore, $\mathbb{J}_*(\psi_0)^{-1} \geq \tilde{\mathbb{J}}_*(\psi_0)^{-1}$ and the inequality holds in the sense that $\mathbb{J}_*(\psi_0)^{-1} - \tilde{\mathbb{J}}_*(\psi_0)^{-1}$ is positive semi-definite.

Remark 1.4.20. *We can also estimate the covariance matrix $\tilde{\mathbb{J}}_*(\psi_0)^{-1}$ in the same way of Theorem 1.4.17. Since we assume that both marginal distributions are the same, the infinite-dimensional parameter in $\tilde{\mathbb{S}}_{\psi_0}$ is estimated by $\hat{b} \equiv (\hat{b}_1, \hat{b}_2, ..., \hat{b}_{d_\psi})'$, where*

$$\hat{b}_k \equiv \arg\min_{b_k \in \mathcal{H}_n} \frac{1}{n}\sum_{i=1}^{n}[(\frac{\partial l(\hat{\theta}_n, W_i)}{\partial \psi_k} - (\frac{\partial l(\hat{\theta}_n, W_i)}{\partial h_0}[b_k])^2]$$

for $k = 1, 2, ..., d_\psi$ and $\mathcal{H}_n$ is the sieve space for $\mathcal{H}$. Then we can construct an estimator of $\tilde{\mathbb{J}}_*(\psi_0)^{-1}$ by using $\hat{b}$ and this estimator is consistent.

1.4.3.3 Asymptotic Normality of the CATEs

As mentioned above, the CATE is one of parameters of interest. Under the model in this paper, we define the CATE on $X = x$ as $\mathbb{E}[Y_1 - Y_0|X = x]$, where $\mathbb{E}[Y_d|X = x] = F_{\epsilon 0}(x^{'}\beta_0 + d)$ with $d \in \{0, 1\}$, denoted by $CATE(x)$. To derive the asymptotic normality of $CATE(x)$, we consider the case of $T(\theta_0) = CATE(\theta_0)$. For all $v \in \mathbb{V}$, we have

$$\frac{\partial CATE(\theta_0)}{\partial \theta'}[v] = \{f_{\epsilon 0}(x^{'}\beta_0 + \delta_0)(x^{'}v_\beta + v_\delta) - f_{\epsilon 0}(x^{'}\beta_0)x^{'}v_\beta\} + \int_{G(x^{'}\beta_0)}^{G(x^{'}\beta_0+\delta_0)} v_\epsilon(t)dt, \tag{1.4.13}$$

where $f_{\epsilon 0}(x) = h_{\epsilon 0}(G(x))g(x)$.

From Proposition 1.4.1, we present the following result without proof :

Theorem 1.4.21. *Let $x \in supp(X)$ be given. Suppose that the conditions in Proposition 1.4.1 hold with $T(\theta_0) = CATE(\theta_0, x)$. Then we have*

$$\sqrt{n}(CATE(\hat{\theta}_n; x) - CATE(\theta_0; x)) \xrightarrow{d} N(0, ||\frac{\partial CATE(\theta_0; x)}{\partial \theta'}[v]||^2), \tag{1.4.14}$$

where $||\frac{\partial CATE(\theta_0;x)}{\partial \theta'}[v]||^2 = \sup_{v\in\mathbb{V},||v||>0} \frac{|\frac{\partial CATE(\theta_0;x)}{\partial \theta'}[v]|}{||v||}$.

The asymptotic variance in Equation (1.4.14) can be estimated by the same way described above and an estimator is given as following :

$$\sigma^2_{CATE} = \max_{v\in\Theta_n} ||\frac{\partial CATE(\hat{\theta}_n; x)}{\partial \theta'}[v]||^2.$$

1.5 Monte Carlo Simulation and Sensitivity Analysis

1.5.1 Simulation Design

We carry out a simulation study to investigate the finite sample performance of the sieve M-estimator $\hat{\theta}_n$ defined in equation (1.4.2). We consider a various data generating processes (DGPs) for both copulas and marginal distributions. The identification of the model requires the copula function to satisfy the stochastic increasing property. HV17 provide several examples of copulas, including the Gaussian, Frank, Clayton, and Gumbel copulas. We consider these copulas to generate the sample. We are also interested in a comparison of performances of the parametric estimators and the semiparametric ones when the marginal distributions are misspecified. To do so, we consider two marginal distributions for ϵ and ν: the standard normal distribution and a mixture of normal distributions. To estimate parametric models, we specify normal distributions with unknown mean and variance parameters for the marginal distributions due to their popularity. We refer to the parameters characterizing the marginal distributions as the nuisance parameters. Specifically, the nuisance parameters are the marginal distribution functions (or density functions) of ϵ and ν themselves in semiparametric models. On the other hand, the nuisance parameters in parametric estimation are the mean and variance parameters of the marginal distribution functions of ϵ and ν. We consider two sample sizes 500 and 1000, and all results are obtained from 2000 Monte Carlo replications.

In simulations, we consider the following data generating process:

$$Y_i = \mathbf{1}\{X_i\beta + D_i\delta_1 \geq \epsilon\}$$

$$D_i = \mathbf{1}\{X_i\alpha + Z_i\gamma \geq \nu\},$$

where $(\alpha, \gamma, \beta, \delta_1) = (-1, 0.8, -1, 1.1)$ and $(X, Z) \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0.1 \\ -0.1 & 1 \end{pmatrix}\right)$.

There are several ways to normalize location and scale for the model, and we impose a restriction that X has no constant for location normalization and set the coefficient on X_1 to one for scale normalization. As we mentioned in the previous section, this normalization allows us to easily compare the performances of parametric and semiparametric estimators and can be used in more generalized models. To apply this normalization to our simulation design, α and β are assigned -1 in our simulation design.

The dependence structure between ϵ and ν is characterized by one-dimensional parameters in all copulas considered in this paper, but the interpretation of the dependence parameter differs across the copulas. To resolve the difficulty in comparison of the degree of dependence between ϵ and ν, we report the Spearman's ρ corresponding to the estimated dependence parameter in each copula specification. We also estimate the models with several values of Spearman's ρ to examine whether the performances of estimators vary across the degree of dependence. Since the Clayton and the Gumbel copulas do not allow negative dependence, we only examine the results from the Gaussian and the Frank copulas in the case where negative dependence is imposed.

We consider two marginal distributions for the DGPs: (i) the standard normal distribution and (ii) a mixture of normal distributions, and we set parameters so that $E[\epsilon] = E[\nu] = 0$ and $Var(\epsilon) = Var(\nu) = 1$ in both cases. Specifically, ϵ and ν are generated from $0.6N(-1, \sigma^2) + 0.4N(1.5, \sigma^2)$ for some $\sigma > 0$ when the true marginal distributions are the mixture of normals. We denote the mixture distribution of normals by TN.

The finite dimensional parameter ψ and the marginal effect of the treatment at some value of the covariate (i.e. CATE) are objects of interest in this class of models. In particular, we focus on CATE at the mean of covariate X as well as ψ. As a performance measure of estimators, we consider the root mean squared errors (RMSEs) in our simulation.

1.5.2 Parametric Models

The parametric models can be estimated by the standard maximum likelihood method. Since it is common to use bivariate probit models for parametric estimation, we specify normal distributions for the marginal distributions. With such a choice of marginal distributions, the model becomes the bivariate probit model if we choose the Gaussian copula. Even if it is commonly assumed that $E[\epsilon] = E[\nu] = 0$ and $Var(\epsilon) = Var(\nu) = 1$ as location and scale normalizations in parametric binary choice models, we adopt the same normalization to the one for the semiparametric model and thus we can easily compare performance of estimators between parametric and semiparametric models.

1.5.3 Semiparametric Models

Since we assume that $\sqrt{h_j} \in \Lambda^p([0,1])$, we approximate h_j to

$$h_j(x) = \frac{(\sum_{k=0}^{k_{nj}} a_{jk}\psi_{jk}(x))^2}{\int_0^1 (\sum_{k=0}^{k_{nj}} a_{jk}\psi_{jk}(x))^2 dx},$$

where $j \in \{\epsilon, \nu\}$, $\{\psi_{jk}(\cdot)\}_{k=0}^{k_{nj}}$ is the set of approximating functions for $h_j(\cdot)$, and k_{nj} is the number of approximating functions. Since h_j's are density functions on the unit interval, we need to impose a restriction that $\int_0^1 h_j(x)dx = 1$ for all $j \in \{\epsilon, \nu\}$. However, the approximation above implies $\int_0^1 h_j(x)dx = 1$ by construction, so we can omit this restriction on the unknown density functions when estimating the model. We take the space of polynomials as the sieve space for h_ϵ and h_ν. The orders of polynomials ($k_{n\epsilon}$ and $k_{n\nu}$) are set to be proportional to $n^{1/7}$. To incorporate the specification given in(1.4.1), we choose the standard normal distribution function for $G(\cdot)$ (i.e. $G(\cdot) = \Phi(\cdot)$).

1.5.4 Copula Misspecification

Although we assume that the copula is correctly specified, the economic theory does not provide a justification for the choice of the copula. In this simulation study, we examine the effect of copula misspecification on the performance of estimators. Misspecification problems in copula-based models have been addressed in the statistic literature (e.g. Kim et al. (2007a,b); Lawless and Yilmaz (2011)). As a related work, Lawless and Yilmaz (2011) compare the performances of the parametric and the semiparametric ML estimators in a copula-based model and show that the semiparametric two-step method performs better than the parametric estimation

method when the copula function is misspecified. To examine the effect of copula misspecification, we only consider a family of copulas that satisfy the stochastic ordering property (Assumption 1.2.6) to ensure the identification of the model in our simulations.

1.5.5 Simulation Results

To compare the performance of the sieve ML estimators with the one of parametric ML estimators, we examine the results from the cases where both the marginal distributions and the copula function are correctly specified (i.e. the true marginal distributions are the standard normal distribution). Table 1.1 shows the estimation results and we find that the estimators of ψ and CATE perform well in both the parametric and the semiparametric models. The biases of estimators are negligible in both models and the variances are small. In addition, the performances of estimators in the semiparametric models are as good as those in the parametric models. For example, the RMSE of the estimator of δ_1 in the semiparametric model with the Gaussian copula is 0.4181 and the one in the parametric model is 0.3982 when the sample size is 500. We also find that the marginal distribution functions are estimated well in the parametric models, and thus both the parametric and the semiparametric models estimate CATE well. The estimator of the dependence parameter ρ also performs well in both models. These results remain the same when the sample size increases. Table 1.7 contains the simulation results with 1000 observations. The results in Table 1.7 demonstrate that the RMSEs decrease and that the parameters are more precisely estimated in both models as the sample size

increases.

Now we consider the cases where the marginal distributions are misspecified in parametric models. Table 1.2 shows simulation results from the cases where the true marginal distributions are TN but a researcher specifies normal distributions for them. Table 1.8 is obtained from simulations under the same situation but with 1000 observations. From these tables, we can find that the MSEs of estimators in parametric models are larger than those in semiparametric models uniformly in the parameters and thus all parameters are estimated more precisely in semiparametric models under the misspecification of the marginal distributions. Moreover, the parametric estimators of the CATE are hugely distorted when the marginal distributions are misspecified and the poor performance of parametric estimators is attributed not only to bias, but also to variance. To be more specific, the bias of the CATE estimator from the parametric model with the Gaussian copula is 0.1377 which is about 8 times larger than the one from the corresponding semiparametric model when the sample size is 500. These biases of CATE estimators are substantial regarding that they do not disappear even when we increase the sample size. Comparing Tables 1.2 and 1.8, we can find that the decreases in RMSEs of CATE estimators with a larger sample size are due to smaller variances and that biases of estimators are generally not reduced even with a larger sample size. Therefore, the simulation results demonstrate that when the marginal distributions are misspecified, the semiparametric models outperform the parametric models since the CATE is one of the most important quantities in the sense that many empirical studies are interested in the CATE in this class of models rather than individual structural parameters

themselves.

Finally, we examine the simulation results when both the copula and the marginal distributions are misspecified. Tables 1.3-1.6 and 1.9-1.12 show the simulation results under misspecification of both copula and marginal distributions. If both copula and marginal distributions are misspecified, the performance of parametric ML estimators are comparable to or slightly worse than the one under marginal misspecification. For example, when the true copula function is a Frank copula and the sample size is 500, we find out that the RMSEs of parametric estimators under marginal misspecification (Table 1.2) are similar to those under both copula and marginal misspecification (Table 1.4). The estimators of ψ under both copula and marginal misspecification (Table 1.4) have slightly larger RMSEs than corresponding ones under marginal misspecification (Table 1.2), but the performance of CATE estimators varies across copula specifications. The degree of distortion is more severe when the true copula function is either the Clayton or the Gumbel copula and the copula function is misspecified (Tables 1.5 and 1.6). In particular, when the true DGP is based on the Gumbel copula, copula misspecification has a significant effect on the performance of estimators of ψ and CATE in parametric models. The RMSEs of estimators of ψ and CATE under copula and marginal misspecification are larger than those under marginal misspecification. Considering Tables 1.2 and 1.6 with a focus on the CATE, the RMSE under marginal misspecification is 0.1637, whereas the RMSEs under both copula and marginal misspecification are 0.1835, 0.2178, and 0.2732 for the Gaussian, Frank, and Clayton copulas, respectively). Even if the sample size increases, these observations remain the same. On the other hand, there is

no clear evidence that the performance of the semiparametric estimators under misspecification of marginal distributions is better than the one under both copula and marginal misspecification. For example, when the true copula is the Frank family, we can see that the finite dimensional parameter except for γ and the CATE in the semiparametric model are estimated better under both misspecification than under misspecification of marginal distributions if the copula is specified by the Gaussian or the Gumbel copula. In contrast, the Clayton copula specification provides the exactly reverse conclusion. Regardless of which copula is used for estimation, however, we can conclude that the semiparametric models dominate the parametric models in terms of RMSE in the presence of marginal misspecification and that the CATE estimators in parametric models are very misleading.

The simulation results suggest that researchers use semiparametric models proposed in this paper when they are concerned about the model misspecification in respect of the marginal distributions. We summarize the main findings from our simulation study:

1. The performance of the sieve ML estimators is comparable to the one of the parametric ML estimators when the model is correctly specified.

2. When the marginal distributions are misspecified, the sieve ML estimator is recommended in regard to the performance of the CATE estimator.

3. If both the copula and the marginal distributions are misspecified, the performance of the parametric ML estimators becomes worse and the semiparametric models are preferred over the parametric models.

Table 1.1: Correctly Specified Models ($n = 500$)

	Parametric Estimation				Semiparametric Estimation			
				Gaussian Copula				
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.3643	0.8000	1.1000	0.5000	0.3643
Estimate	0.8074	1.1469	0.4956	0.3657	0.8070	1.1577	0.5037	0.3584
S.D	0.0934	0.3954	0.1537	0.0897	0.0940	0.4141	0.1528	0.0935
Bias	0.0074	0.0469	-0.0044	0.0014	0.0070	0.0577	0.0038	-0.0060
RMSE	0.0936	0.3982	0.1537	0.0897	0.0943	0.4181	0.1528	0.0937
				Frank Copula				
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.3643	0.8000	1.1000	0.5000	0.3643
Estimate	0.8027	1.1450	0.4909	0.3681	0.8028	1.1556	0.4981	0.3598
S.D	0.0936	0.3379	0.1310	0.0781	0.0943	0.3588	0.1314	0.0829
Bias	0.0027	0.0450	-0.0091	0.0037	0.0028	0.0556	-0.0019	-0.0045
RMSE	0.0936	0.3409	0.1313	0.0781	0.0944	0.3631	0.1314	0.0830
				Clayton Copula				
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.3643	0.8000	1.1000	0.5000	0.3643
Estimate	0.8024	1.1083	0.5075	0.3598	0.8027	1.1275	0.5140	0.3504
S.D	0.0942	0.3371	0.1368	0.0791	0.0935	0.3719	0.1354	0.0816
Bias	0.0024	0.0083	0.0075	-0.0045	0.0027	0.0275	0.0139	-0.0139
RMSE	0.0942	0.3372	0.1370	0.0792	0.0936	0.3729	0.1361	0.0828
				Gumbel Copula				
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.3643	0.8000	1.1000	0.5000	0.3643
Estimate	0.8026	1.1339	0.5060	0.3605	0.8035	1.1564	0.5102	0.3562
S.D	0.0974	0.4002	0.1488	0.0894	0.0994	0.4300	0.1535	0.0978
Bias	0.0026	0.0339	0.0060	-0.0038	0.0035	0.0564	0.0102	-0.0081
RMSE	0.0974	0.4016	0.1489	0.0895	0.0995	0.4337	0.1539	0.0981

* The true DGP marginal distributions are the standard normal distribution.

Table 1.2: Marginal Misspecfication ($n = 500$)

	Parametric Estimation				Semiparametric Estimation			
	Gaussian Copula							
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.7994	1.0925	0.4496	0.2443	0.8562	1.2696	0.4895	0.1241
S.D	0.1281	0.6285	0.1651	0.1129	0.1113	0.3728	0.1059	0.0653
Bias	-0.0006	-0.0075	-0.0504	0.1377	0.0562	0.1696	-0.0105	0.0174
RMSE	0.1281	0.6285	0.1726	0.1780	0.1247	0.4096	0.1064	0.0675
	Frank Copula							
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.8056	1.3088	0.3976	0.2894	0.8377	1.2541	0.4829	0.1276
S.D	0.1272	0.5093	0.1221	0.0883	0.1141	0.3564	0.0963	0.0689
Bias	0.0056	0.2088	-0.1024	0.1827	0.0377	0.1541	-0.0171	0.0210
RMSE	0.1273	0.5504	0.1594	0.2030	0.1202	0.3883	0.0978	0.0720
	Clayton Copula							
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.8099	1.1439	0.4236	0.2555	0.8441	1.2234	0.4948	0.1192
S.D	0.1309	0.5236	0.1412	0.0913	0.1134	0.3611	0.0999	0.0611
Bias	0.0099	0.0439	-0.0764	0.1488	0.0441	0.1234	-0.0053	0.0126
RMSE	0.1312	0.5254	0.1605	0.1746	0.1217	0.3816	0.1001	0.0624
	Gumbel Copula							
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.7892	1.0326	0.4650	0.2373	0.8484	1.2692	0.4900	0.1259
S.D	0.1333	0.5297	0.1338	0.0986	0.1142	0.3646	0.0986	0.0645
Bias	-0.0108	-0.0674	-0.0350	0.1307	0.0484	0.1692	-0.0099	0.0193
RMSE	0.1337	0.5340	0.1383	0.1637	0.1241	0.4019	0.0991	0.0673

* The true DGP marginal distributions are the mixture of normals.

Table 1.3: Copula and Marginals Misspecification 1 ($n = 500$)

	Parametric Estimation				Semiparametric Estimation			
				Frank Copula				
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.8140	1.3080	0.3775	0.2916	0.8463	1.3514	0.4499	0.1351
S.D	0.1257	0.4899	0.1202	0.0862	0.1137	0.3502	0.0964	0.0686
Bias	0.0140	0.2080	-0.1225	0.1849	0.0463	0.2514	-0.0501	0.0285
RMSE	0.1265	0.5322	0.1716	0.2040	0.1227	0.4311	0.1087	0.0743
				Clayton Copula				
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.8244	1.5699	0.3691	0.3176	0.8534	1.4386	0.4945	0.1586
S.D	0.1271	0.6609	0.1697	0.0999	0.1128	0.3398	0.1044	0.0734
Bias	0.0244	0.4699	-0.1308	0.2110	0.0534	0.3386	-0.0054	0.0520
RMSE	0.1294	0.8109	0.2143	0.2335	0.1248	0.4797	0.1046	0.0899
				Gumbel Copula				
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.7981	1.0706	0.4232	0.2448	0.8546	1.2025	0.4697	0.1137
S.D	0.1281	0.5795	0.1519	0.1077	0.1118	0.3611	0.1027	0.0600
Bias	-0.0019	-0.0294	-0.0767	0.1382	0.0546	0.1025	-0.0302	0.0070
RMSE	0.1281	0.5802	0.1702	0.1752	0.1244	0.3754	0.1070	0.0604

* The true DGP copula and marginals are the Gaussian and mixture of normals, respectively.

Table 1.4: Copula and Marginals Misspecification 2 ($n = 500$)

	Parametric Estimation				Semiparametric Estimation			
Gaussian Copula								
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.7992	1.1673	0.4517	0.2527	0.8500	1.1788	0.5173	0.1192
S.D	0.1342	0.6901	0.1680	0.1179	0.1158	0.3602	0.1000	0.0652
Bias	-0.0008	0.0673	-0.0483	0.1461	0.0500	0.0788	0.0173	0.0126
RMSE	0.1342	0.6934	0.1748	0.1877	0.1262	0.3687	0.1015	0.0664
Clayton Copula								
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.8235	1.6132	0.3870	0.3184	0.8484	1.3679	0.5212	0.1548
S.D	0.1329	0.7039	0.1670	0.1018	0.1188	0.3416	0.1012	0.0755
Bias	0.0235	0.5132	-0.1130	0.2118	0.0484	0.2679	0.0212	0.0482
RMSE	0.1350	0.8711	0.2017	0.2350	0.1283	0.4341	0.1034	0.0896
Gumbel Copula								
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.8001	1.1697	0.4202	0.2564	0.8485	1.1059	0.4997	0.1071
S.D	0.1347	0.6697	0.1608	0.1165	0.1161	0.3548	0.0997	0.0601
Bias	0.0001	0.0697	-0.0798	0.1498	0.0485	0.0059	-0.0003	0.0005
RMSE	0.1347	0.6733	0.1795	0.1897	0.1258	0.3548	0.0997	0.0601

* The true DGP copula and marginal distributions are the Frank copula and mixture of normals, respectively.

Table 1.5: Copula and Marginals Misspecification 3 ($n = 500$)

	Parametric Estimation				Semiparametric Estimation			
				Gaussian Copula				
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.7986	1.0471	0.4017	0.2392	0.8533	1.1780	0.4493	0.1076
S.D	0.1346	0.6366	0.1731	0.1181	0.1164	0.3438	0.1033	0.0569
Bias	-0.0014	-0.0529	-0.0983	0.1325	0.0533	0.0780	-0.0508	0.0009
RMSE	0.1346	0.6388	0.1991	0.1775	0.1281	0.3525	0.1151	0.0569
				Frank Copula				
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.8083	1.1559	0.3611	0.2712	0.8412	1.2404	0.4199	0.1160
S.D	0.1318	0.4453	0.1143	0.0856	0.1166	0.3408	0.0965	0.0611
Bias	0.0083	0.0559	-0.1389	0.1646	0.0412	0.1404	-0.0802	0.0094
RMSE	0.1321	0.4488	0.1799	0.1855	0.1237	0.3686	0.1255	0.0619
				Gumbel Copula				
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.8046	1.1937	0.3316	0.2680	0.8542	1.1610	0.4148	0.1046
S.D	0.1355	0.6663	0.1748	0.1220	0.1166	0.3283	0.1032	0.0557
Bias	0.0046	0.0937	-0.1684	0.1613	0.0542	0.0610	-0.0852	-0.0020
RMSE	0.1356	0.6728	0.2427	0.2022	0.1285	0.3339	0.1339	0.0557

* The true DGP copula and marginal distributions are the Clayton copula and mixture of normals, respectively.

Table 1.6: Copula and Marginals Misspecification 4 ($n = 500$)

	Parametric Estimation				Semiparametric Estimation			
	Gaussian Copula							
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.7978	1.1488	0.4658	0.2523	0.8609	1.3801	0.4957	0.1460
S.D	0.1304	0.6489	0.1598	0.1117	0.1132	0.3749	0.1052	0.0730
Bias	-0.0022	0.0488	-0.0342	0.1456	0.0609	0.2801	-0.0042	0.0393
RMSE	0.1304	0.6508	0.1634	0.1835	0.1286	0.4679	0.1053	0.0829
	Frank Copula							
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.8140	1.4128	0.3834	0.3064	0.8532	1.4755	0.4543	0.1611
S.D	0.1290	0.5211	0.1184	0.0867	0.1177	0.3466	0.0969	0.0752
Bias	0.0140	0.3128	-0.1166	0.1998	0.0532	0.3755	-0.0457	0.0545
RMSE	0.1297	0.6078	0.1662	0.2178	0.1292	0.5110	0.1072	0.0929
	Clayton Copula							
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.8276	1.8999	0.3208	0.3614	0.8603	1.6010	0.4823	0.1960
S.D	0.1321	0.7365	0.1753	0.0986	0.1172	0.3103	0.1065	0.0799
Bias	0.0276	0.7999	-0.1791	0.2548	0.0603	0.5010	-0.0177	0.0894
RMSE	0.1350	1.0873	0.2506	0.2732	0.1318	0.5893	0.1079	0.1199

* The true DGP copula and marginal distributions are the Gumbel copula and mixture of normals, respectively.

Table 1.7: Correctly Specified Models ($n = 1,000$)

	Parametric Estimation				Semiparametric Estimation			
Gaussian Copula								
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.3643	0.8000	1.1000	0.5000	0.3643
Estimate	0.8025	1.1165	0.4996	0.3632	0.8026	1.1205	0.5031	0.3596
S.D	0.0654	0.2737	0.1081	0.0656	0.0655	0.2939	0.1092	0.0668
Bias	0.0025	0.0165	-0.0004	-0.0011	0.0026	0.0205	0.0031	-0.0048
RMSE	0.0655	0.2742	0.1081	0.0656	0.0655	0.2946	0.1092	0.0670
Frank Copula								
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.3643	0.8000	1.1000	0.5000	0.3643
Estimate	0.8017	1.1188	0.5010	0.3635	0.8007	1.1164	0.5042	0.3594
S.D	0.0658	0.2605	0.1023	0.0620	0.0652	0.2663	0.1066	0.0652
Bias	0.0017	0.0188	0.0010	-0.0009	0.0007	0.0164	0.0042	-0.0049
RMSE	0.0658	0.2612	0.1023	0.0620	0.0652	0.2668	0.1067	0.0653
Clayton Copula								
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.3643	0.8000	1.1000	0.5000	0.3643
Estimate	0.8030	1.1055	0.5007	0.3621	0.8029	1.1100	0.5035	0.3572
S.D	0.0658	0.2329	0.0958	0.0566	0.0659	0.2524	0.0964	0.0560
Bias	0.0030	0.0055	0.0007	-0.0023	0.0029	0.0100	0.0035	-0.0071
RMSE	0.0659	0.2330	0.0958	0.0567	0.0660	0.2526	0.0965	0.0565
Gumbel Copula								
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.3643	0.8000	1.1000	0.5000	0.3643
Estimate	0.8022	1.1192	0.4963	0.3644	0.8025	1.1240	0.4986	0.3626
S.D	0.0668	0.2655	0.1057	0.0635	0.0665	0.2818	0.1086	0.0684
Bias	0.0022	0.0192	-0.0037	0.0001	0.0025	0.0240	-0.0014	-0.0017
RMSE	0.0669	0.2662	0.1057	0.0635	0.0665	0.2829	0.1086	0.0684

* The true DGP marginal distributions are the standard normal distribution.

Table 1.8: Marginal Misspecfication ($n = 1,000$)

	Parametric Estimation				Semiparametric Estimation			
				Gaussian Copula				
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.7941	1.0549	0.4496	0.2447	0.8641	1.3030	0.4778	0.1262
S.D	0.0911	0.4256	0.1156	0.0807	0.0778	0.2576	0.0721	0.0463
Bias	-0.0059	-0.0451	-0.0504	0.1381	0.0641	0.2030	-0.0222	0.0195
RMSE	0.0913	0.4279	0.1261	0.1599	0.1008	0.3279	0.0755	0.0502
				Frank Copula				
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.8044	1.3066	0.3940	0.2919	0.8525	1.2802	0.4777	0.1291
S.D	0.0899	0.3876	0.0966	0.0684	0.0837	0.2577	0.0690	0.0500
Bias	0.0044	0.2066	-0.1060	0.1853	0.0525	0.1802	-0.0223	0.0225
RMSE	0.0901	0.4392	0.1434	0.1975	0.0988	0.3145	0.0725	0.0549
				Clayton Copula				
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.8065	1.1207	0.4240	0.2553	0.8547	1.2669	0.4851	0.1219
S.D	0.0906	0.3704	0.1047	0.0677	0.0801	0.2622	0.0706	0.0456
Bias	0.0065	0.0207	-0.0761	0.1487	0.0547	0.1669	-0.0150	0.0153
RMSE	0.0908	0.3710	0.1294	0.1634	0.0969	0.3108	0.0722	0.0481
				Gumbel Copula				
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.7849	1.0104	0.4606	0.2391	0.8618	1.2980	0.4791	0.1268
S.D	0.0893	0.3566	0.0950	0.0695	0.0781	0.2516	0.0684	0.0463
Bias	-0.0151	-0.0896	-0.0393	0.1325	0.0618	0.1980	-0.0208	0.0201
RMSE	0.0906	0.3677	0.1028	0.1496	0.0996	0.3202	0.0715	0.0504

* The true DGP marginal distributions are the mixture of normals.

Table 1.9: Copula and Marginals Misspecification 1 ($n = 1,000$)

	Parametric Estimation				Semiparametric Estimation			
				Frank Copula				
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.8086	1.3159	0.3652	0.2975	0.8549	1.3936	0.4376	0.1371
S.D	0.0897	0.3636	0.0927	0.0650	0.0830	0.2548	0.0689	0.0506
Bias	0.0086	0.2159	-0.1347	0.1909	0.0549	0.2936	-0.0623	0.0305
RMSE	0.0901	0.4229	0.1636	0.2017	0.0995	0.3887	0.0929	0.0591
				Clayton Copula				
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.8193	1.5478	0.3661	0.3205	0.8613	1.4684	0.4886	0.1574
S.D	0.0906	0.4574	0.1217	0.0705	0.0812	0.2351	0.0710	0.0514
Bias	0.0193	0.4478	-0.1338	0.2139	0.0613	0.3684	-0.0113	0.0508
RMSE	0.0927	0.6401	0.1809	0.2252	0.1018	0.4370	0.0719	0.0722
				Gumbel Copula				
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.7930	1.0391	0.4210	0.2453	0.8620	1.2302	0.4574	0.1157
S.D	0.0911	0.4010	0.1070	0.0771	0.0790	0.2554	0.0709	0.0439
Bias	-0.0070	-0.0609	-0.0789	0.1386	0.0620	0.1302	-0.0426	0.0090
RMSE	0.0914	0.4056	0.1330	0.1586	0.1004	0.2867	0.0827	0.0449

* The true DGP copula and marginal distributions are the Gaussian copula and mixture of normals, respectively.

Table 1.10: Copula and Marginals Misspecification 2 ($n = 1,000$)

	Parametric Estimation				Semiparametric Estimation			
Gaussian Copula								
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.7935	1.0825	0.4653	0.2465	0.8601	1.1832	0.5145	0.1196
S.D	0.0926	0.4333	0.1152	0.0803	0.0768	0.2641	0.0723	0.0450
Bias	-0.0065	-0.0175	-0.0347	0.1399	0.0601	0.0832	0.0145	0.0130
RMSE	0.0929	0.4336	0.1203	0.1613	0.0976	0.2769	0.0738	0.0468
Clayton Copula								
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.8188	1.5580	0.3941	0.3173	0.8583	1.3743	0.5200	0.1542
S.D	0.0919	0.4621	0.1194	0.0708	0.0794	0.2439	0.0718	0.0526
Bias	0.0188	0.4580	-0.1059	0.2106	0.0583	0.2743	0.0200	0.0476
RMSE	0.0938	0.6506	0.1595	0.2222	0.0985	0.3671	0.0746	0.0709
Gumbel Copula								
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.7954	1.0843	0.4327	0.2496	0.8596	1.1105	0.4959	0.1082
S.D	0.0927	0.4252	0.1119	0.0796	0.0765	0.2578	0.0708	0.0413
Bias	-0.0046	-0.0157	-0.0673	0.1429	0.0596	0.0105	-0.0041	0.0016
RMSE	0.0928	0.4255	0.1306	0.1636	0.0970	0.2580	0.0709	0.0413

* The true DGP copula and marginal distributions are the Frank copula and mixture of normals, respectively.

Table 1.11: Copula and Marginals Misspecification 3 ($n = 1,000$)

	Parametric Estimation				Semiparametric Estimation			
				Gaussian Copula				
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.7928	0.9952	0.4102	0.2370	0.8618	1.2015	0.4441	0.1097
S.D	0.0929	0.4262	0.1233	0.0837	0.0764	0.2527	0.0737	0.0411
Bias	-0.0072	-0.1048	-0.0898	0.1303	0.0618	0.1015	-0.0559	0.0030
RMSE	0.0932	0.4389	0.1525	0.1549	0.0983	0.2723	0.0925	0.0412
				Frank Copula				
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.8048	1.1667	0.3480	0.2754	0.8510	1.2695	0.4101	0.1152
S.D	0.0910	0.3362	0.0918	0.0649	0.0825	0.2578	0.0701	0.0453
Bias	0.0048	0.0667	-0.1520	0.1688	0.0510	0.1695	-0.0899	0.0086
RMSE	0.0911	0.3428	0.1776	0.1808	0.0970	0.3085	0.1140	0.0461
				Gumbel Copula				
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.8046	1.1937	0.3316	0.2680	0.8594	1.1883	0.4090	0.1054
S.D	0.1355	0.6663	0.1748	0.1220	0.0784	0.2373	0.0727	0.0412
Bias	0.0046	0.0937	-0.1684	0.1613	0.0594	0.0883	-0.0911	-0.0013
RMSE	0.1356	0.6728	0.2427	0.2022	0.0984	0.2532	0.1165	0.0412

* The true DGP copula and marginal distributions are the Clayton copula and mixture of normals, respectively.

Table 1.12: Copula and Marginals Misspecification 4 ($n = 1,000$)

	Parametric Estimation				Semiparametric Estimation			
	Gaussian Copula							
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.7905	1.1059	0.4669	0.2520	0.8660	1.4046	0.4893	0.1428
S.D	0.0896	0.4412	0.1167	0.0815	0.0775	0.2644	0.0723	0.0508
Bias	-0.0095	0.0059	-0.0330	0.1454	0.0660	0.3046	-0.0107	0.0362
RMSE	0.0901	0.4412	0.1213	0.1667	0.1018	0.4034	0.0730	0.0624
	Frank Copula							
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.8123	1.4374	0.3701	0.3149	0.8628	1.5142	0.4473	0.1582
S.D	0.0901	0.3917	0.0930	0.0651	0.0817	0.2377	0.0697	0.0545
Bias	0.0123	0.3374	-0.1299	0.2083	0.0628	0.4142	-0.0526	0.0515
RMSE	0.0910	0.5169	0.1597	0.2182	0.1030	0.4776	0.0874	0.0750
	Clayton Copula							
	γ	δ_1	ρ_{sp}	ATE	γ	δ_1	ρ_{sp}	ATE
True Values	0.8000	1.1000	0.5000	0.1066	0.8000	1.1000	0.5000	0.1066
Estimate	0.8228	1.8913	0.3197	0.3656	0.8645	1.6249	0.4851	0.1894
S.D	0.0927	0.5234	0.1336	0.0714	0.0808	0.2084	0.0742	0.0550
Bias	0.0228	0.7913	-0.1803	0.2589	0.0645	0.5249	-0.0149	0.0828
RMSE	0.0955	0.9488	0.2244	0.2686	0.1034	0.5648	0.0757	0.0994

* The true DGP copula and marginal distributions are the Gumbel copula and mixture of normals, respectively.

1.6 Conclusions

In this paper, we propose semiparametric estimation and inference methods for generalized bivariate probit models. Specifically, we develop the asymptotic theory for the sieve ML estimators of semiparametric copula-based triangular systems with binary endogenous variables. It is shown that the sieve ML estimators are consistent and that their smooth functionals are $\sqrt{n}$-asymptotically normal under some regularity conditions. This semiparametric estimation approach allows the flexibility

of the models and thus provides robustness in estimation and inference.

We conduct a sensitivity analysis to examine how sensitive the estimation results are to model specifications. From this analysis, we find that overall the semiparametric estimators perform well in terms of both bias and variance. When the marginal distributions are misspecified, the semiparametric estimates significantly outperform parametric estimates and the latter exhibit substantial bias. In particular, we find that the estimates of the parameters involving the misspecified marginal distributions, such as the ATE, are very misleading. When the model is correctly specified, we show that the performance of the semiparametric estimators are comparable to that of the parametric ones. When the copula is also misspecified, the distortion of the parametric estimates under marginal misspecification becomes even more severe, whereas the semiparametric estimates do not seem to be affected by this misspecification as long as the copula of the true DGP is within the stochastic ordering class. A related interesting question is how the results would change when the data are not generated from this class of copulas.

We also formally show that the exclusion restriction is not only sufficient but also necessary for identification. Without exclusion restriction, the model parameters are not identified or, under the normality assumption, are at best weakly identified. Some empirical studies ignore the exclusion restriction when estimating the model, and our non-identification result provides a caveat for practitioners.

Chapter 2

Nonparametric Tests for Conditional Quantile Independence with Duration Outcomes

2.1 Introduction

Since the seminal work of Koenker and Bassett (1978), quantile regression (QR) models have received much attention from both theoretical and applied econometrics. Due to many appealing properties of QR, conditional quantile models have become a good alternative to conditional mean models and thus increasingly gained attention. One of the desirable features of the QR is that it can provide information on the distribution of the dependent variable conditional on covariates which allows one to capture heterogeneous effects of the covariates on the dependent variable across quantiles. In the context of the treatment effects literature, even if the average treatment effect is the most common used measure of treatment, the information on heterogeneity of treatment effects, if it exists, is likely to be missed as the mean effectively integrates out the heterogeneous factors[1]. In addition, it is well-known that the QR is less sensitive to outliers than the mean regression and places

[1] Buchinsky (1994) uses QR to describe changes in the returns to education and experience across the distribution of wage and part of his results indicates that the returns to education and experience exhibit heterogeneity across the quantiles. Bitler et al. (2006) examine the effects of policy reforms on welfare including earnings, taking the heterogeneity of the effect across the distribution into account. They find out that there is a substantial heterogeneity across the distribution from the estimation of the quantile treatment effect and that the average treatment effect may result in a misleading prediction.

less assumptions on the distribution of unobserved error terms such as existence of moments.

It is prevalent to specify a parametric conditional quantile model in empirical analysis. Viewing a parametric specification as an approximation of the true model, a parametric specification facilitates estimation and inference procedures and provides a natural way to interpret the model. However, since economic theories seldom imply parametric specifications, parametric models are vulnerable to misspecification which results in misleading implications of the models. To alleviate the sensitivity of parametric models, one can consider a fully nonparametric QR.

Even if a nonparametric model is attractive for its flexibility, it typically requires a larger sample than a parametric model to obtain estimators of reasonable precision. Moreover, the rate of convergence is very slow when the number of covariates is large due to the curse of dimensionality. There are many attempts to circumvent the curse of dimensionality in the literature. The main idea of these attempts is to impose structure on the model to improve the rate of convergence. Examples include additive separability of regression models and partially linear models (e.g. Robinson (1988); Andrews and Whang (1990); Lee (2003); Horowitz and Lee (2005)). It has been shown that these structures can improve rate of convergence and consequently bypass the efficiency issue of fully nonparametric models. However, those models with additional structures are also not free from misspecification.

This paper considers nonparametric tests for a null hypothesis that a subset of the entire covariates is jointly significant. Once the model is nonparametrically estimated, rejection of the null hypothesis is not an indication of misspecification

but a suggestion that there are omitted variables. The results of these tests provide information on variable selection in QR and can mitigate problems caused by large dimensionality of covariates. To formulate test statistics, I characterize the null hypothesis as a conditional moment restriction and then employ the integrated conditional moment (ICM) approach that was proposed by Bierens (Bierens (1982, 1990)). Bierens demonstrate in a series of papers that a conditional moment restriction can be characterized by an infinite number of unconditional moments with an appropriately chosen weighting function. This result is used to perform a parametric specification testing[2]. Bierens and Ploberger (1997) establish the asymptotic theory for the ICM test statistic and obtain upper bounds on the critical values that guarantee the actual size of test is bounded by the nominal size specified by researchers. The tests of this paper differ from the original test of Bierens (Bierens (1982, 1990); Bierens and Ploberger (1997)) in that this paper considers a nonparametric null hypothesis, which involves infinite-dimensional parameters.

One of the desirable features of the ICM approach is that it does not require to estimate alternative models and thus reduces some computational burden in obtaining the test statistics. In nonparametric tests, if a test statistic contains nonparametric objects and directly compares the null model with alternatives, it is hard to achieve power against local alternatives at $\sqrt{n}$-rate (e.g. Hardle and Mammen (1993); Hong and White (1995); Fan and Li (1996)). The ICM approach, however, makes it possible to have non-trivial power against local alternatives at

[2]Similar questions are addressed in Stute (1997) and Koul and Stute (1999), but they use the indicator function as the weighting function to transform conditional moment restrictions into unconditional ones.

the parametric rate even if the test statistic contains infinite-dimensional parameters. Subsequent studies that combine the ICM approach with nonparametric null hypotheses include Chen and Fan (1999), Delgado and Manteiga (2001), Li et al. (2003), and Huang et al. (2016) just to name a few.

Unlike the model specification tests for conditional mean regression, the test statistics of this paper contain an indicator function which is non-smooth and this non-smoothness introduces difficulty in applying the approach used for testing conditional mean models. I employ a stochastic equicontinuity argument to obtain a stochastic expansion of the test statistics with the non-smooth function to derive the asymptotic distribution of the test statistics. The stochastic equicontinuity has been applied to both parametric and semi-/non-parametric models in, for example, Andrews (1994a,b), Newey (1994), and Chen et al. (2003). Those papers use stochastic equicontinuity to derive the asymptotic distribution of an estimator in the presence of non-smooth functions and infinite-dimensional parameters. It is hard to directly show that some processes are stochastically equicontinuous and thus I make use of empirical process theory to prove stochastic equicontinuity.

This paper also incorporates censoring for the dependent variable. In some empirical applications such as a duration or survival analysis, the dependent variable is not completely observed due to censoring. Consider, for example, the case where one may be interested in estimating the effect of unemployment insurance benefits on the unemployment duration and suppose that individuals' duration spells are only observed when they were receiving the unemployment benefit. In this case, the duration spell is not completely observed and thus it is subject to censoring.

The test statistics require one to estimate the conditional quantile function under the null hypothesis. Even if a censoring variable is (conditionally) independent of the outcome of interest, ignoring the censoring results in inconsistency of estimators. Therefore, I estimate the conditional distribution function by using a variant of the Kaplan-Meier (KM) estimator (Kaplan and Meier (1958)) then invert the estimated distribution function to obtain the conditional quantile function. To accommodate regressors, it is required that one use a conditional KM estimator (Beran (1981); Dabrowska (1989, 1992); Gonzalez-Manteiga and Cadarso-Suarez (1994); Wang and Wang (2009)). Specifically, I use the local conditional KM estimator proposed by Kong and Xia (2017), which is a local polynomial regression version of conditional KM estimator.

The test statistics in this paper are asymptotically smooth functionals of a Gaussian process and their asymptotic distributions depend on the data generating process. Therefore, it is difficult to tabulate critical values for the test statistics. To resolve this problem, I use a subsampling method to approximate the asymptotic distributions of the test statistics. It is shown that, under a set of conditions, the subsampling method yields critical values that guarantee the asymptotically correct size of test.

A closely related paper, Volgushev et al. (2013) recently proposed a nonparametric test for significance of covariates in conditional quantile models and address the same question as in this paper. The test proposed in this paper is similar to theirs in terms of the form of the test statistic. The main difference from Volgushev et al. (2013) is that this paper covers the case where the dependent variable is sub-

ject to censoring. A minor difference between Volgushev et al. (2013) and this paper is that I use different weighting functions to construct test statistics. Specifically, Volgushev et al. (2013) use the indicator function as a weighting function in the test statistic, but I consider another class of weighting functions for the test statistic. The class of functions is called *generically comprehensively revealing* (GCR, hereafter) class, which is a term coined by Stinchcombe and White (1998). Huang et al. (2016) discuss the choice of weighting function between the indicator function and the GCR class and point out that there are several advantages of the GCR class over the indicator function for testing the conditional quantile independence.

Sant'Anna (2016) is another closely related paper in that he considers tests for nonparametric models with duration outcomes. He proposes nonparametric specification tests in a treatment effect context. His tests also rely on the Bierens's ICM approach, and he suggests using a bootstrap procedure to obtain the critical values. The tests of Sant'Anna (2016) can also be used to test similar hypotheses, such as homogeneity of the conditional average treatment effect, to the hypothesis considered in this paper, but this paper focuses on QR models with duration outcomes.

The rest of this paper is organized as follows. I briefly review related studies in the literature in the following subsection. Section 2.2 formalizes the model and construct the test statistics. Section 2.3 establishes the asymptotic theory of the tests. A subsampling procedure is provided in Section 2.4. Section 2.5 concludes.

2.1.1 Related Literature

There are a lot of studies on both parametric and nonparametric QR models since Koenker and Bassett (1978) who pioneer the theory on QR using the "check function" approach and establish the asymptotic distribution of the QR estimator. Specifically, this paper proposes a specification testing of nonparametric QR models, therefore it is related to the studies on nonparametric QR models estimated by the local polynomial quantile regression. Chaudhuri (1991) develops a nonparametric estimation method for the conditional quantile function, which is similar to the local polynomial regression, and derives a Bahadur's representation. Some refinements of the representation are examined in several studies such as Kong et al. (2010); Guerre and Sabbah (2012); Lee et al. (2015); Qu and Yoon (2015). For censored QR models, different methods to estimate conditional quantile functions have been developed in the literature (e.g. Buchinsky and Hahn (1998); Chernozhukov and Hong (2002); Honore et al. (2002); Portnoy (2003); Wang and Wang (2009)). In contrast to the standard QR models, the literature on nonparametric QR with (random) censoring is relatively small and includes Beran (1981), Dabrowska (1989), Dabrowska (1992), Kong et al. (2013), and Kong and Xia (2017).

In terms of testing the significance of covariates, this paper is closely related to Fan and Li (1996) who consider a nonparametric specification testing for conditional mean regression models. They construct a test statistic only using the restricted model estimated by the kernel method and derive the asymptotic distribution of the test statistic under the null. Their test statistic is based on an equivalent conditional moment restriction and they show that the test is consistent and has

power against local alternatives at a nonparametric rate slower than $\sqrt{n}$. Chen and Fan (1999) propose a nonparametric test for more general hypotheses for conditional mean models. Their test procedure makes use of the ICM approach and they derive the asymptotic distribution of a stochastic process by using the central limit theorem for Hilbert-valued random arrays that could be serially correlated. Delgado and Manteiga (2001) address the same question to the one of Fan and Li (1996) and use the ICM approach with kernel methods. Chen and Fan (1999) and Delgado and Manteiga (2001) share some common features with the test of this paper: (i) the null hypothesis is nonparametrically or semiparametrically specified and (ii) the tests rely on the ICM approach. Whereas Chen and Fan (1999) and Delgado and Manteiga (2001) consider conditional mean models, the test in this paper focuses on conditional QR models. One can also refer to Lavergne and Vuong (2000); Lavergne and Patilea (2008); Lavergne et al. (2015) for testing significance in conditional mean models.

There are also a myriad of studies on the specification testings for QR models, but most of studies focus on testing parametric specifications. Koenker and Bassett (1982) investigate three tests for linear QR models - the Wald, the likelihood ratio (LR), and the Lagrange multiplier (LM) tests - with the focus on the significance of covariates. Zheng (1998) proposes a testing procedure for parametric specifications under the conditional quantile restriction, which is similar to the approach of Fan and Li (1996). The test of Bierens and Ginther (2001) relies on the ICM approach to testing parametric specifications of conditional quantile models. Horowitz and Spokoiny (2002) develop a test statistic and provide a resampling

method for obtaining critical values for their test statistic, but their specification test is also applicable to parametric specification testing. He and Zhu (2003) and Whang (2006a) use a similar idea of the ICM approach, but the weighting function they use is different from the ones considered in the original work of Bierens[3]. Both suggest using resampling methods to simulate the distributions of the test statistics. Whang (2006b) develops a specification test for the parametric conditional quantile model and focuses on the case where the parameters in the model are estimated by using the empirical likelihood method. As mentioned above, this paper is different from those in that the test statistic accommodates infinite-dimensional parameters as the null hypothesis is nonparametrically formulated and I also consider the issue of censoring which these papers do not.

This paper is also related to variable selection in QR models. Belloni and Chernozhukov (2011) investigate the issue on variable selection in high-dimensional sparse models. They propose l^1-penalized QR to deal with the high-dimensionality with sparsity and establish asymptotic results on the penalized QR estimators. The main difference from Belloni and Chernozhukov (2011) is that neither high-dimensional data nor the sparsity assumption are considered in this paper.

2.2 Model and Test Statistics

Let T be the dependent variable of interest and X be a vector of covariates of dimension $d_x \geq 2$. In cases where the outcome variable is censored by a variable

[3]They consider the indicator function as an alternative to the weighting function of the form in Bierens and Ginther (2001) and Bierens and Ploberger (1997).

denoted by C, researchers usually observe $W \equiv (Y, D, X') \in \mathbb{R}^{2+d_x}$, where

$$Y_i = \min(T_i, C_i), \ D_i = \mathbf{1}(T_i \leq C_i).$$

Let $\tau \in (0,1)$ be given and $F_{T|X}(t|x)$ be the conditional distribution function of T given $X = x$. Then τ-th conditional quantile function of T given $X = x$ is defined as

$$Q_{T|X}(\tau|x) \equiv \inf\{q \in \mathbb{R} : F_{T|X}(q|x) \geq \tau\}. \tag{2.2.1}$$

Suppose that X can be divided into two parts X_1 and X_2, where $X_1 \in \mathbb{R}^{d_1}$, $X_2 \in \mathbb{R}^{d_2}$, and $d_1 + d_2 = d_x$ and that it is believed that the variable X_2 is not significant for the τ-th conditional quantile of Y, conditional on X_1 (i.e. the τ-th conditional quantile of Y given X only depends on X_1, but X_2). Since economic theories do not suggest a parametric form for $F_{T|X}$ (equivalently $Q_{T|X}$) in most cases, one may need to nonparametrically estimate $Q_{T|X}$ unless there is a strong belief in a specific form of $Q_{T|X}$. However, it is well-known that nonparametric estimators suffer from the curse of dimensionality, so it would be desirable to omit such insignificant variables from regression to allow efficiency gains for estimators.

Considering the duality between conditional quantile processes and the conditional distributions, the notion of insignificance of covariates in QR is equivalent to the notion called *conditional quantile independence*. Before formalizing the null hypothesis, recall the formal definition of conditional quantile independence.

Definition 2.2.1. *Let Y, X, and Z be random variables. For a given $\tau \in (0,1)$ the*

variable Z is said to be ***conditionally τ-quantile independent*** *of Y on X if*

$$Q_{Y|X,Z}(\tau|X,Z) = Q_{Y|X}(\tau|X).$$

The concept of conditional τ-quantile independence is also related to important questions in economics and some illustrative examples are given below:

Example 2.2.1 (The effect of unemployment insurance benefit on unemployment duration)**.** *Many studies have examined factors affecting individual unemployment duration spell (e.g. Heckman and Singer (1984); Han and Hausman (1990); Katz and Meyer (1990); Meyer (1990)), and unemployment insurance has been considered as one of the determinants of unemployment duration. Let UI be the level of unemployment insurance benefits and X be other covariates. Letting T denote the unemployment duration spell, one can formulate the null hypothesis for testing the effect of UI on the quantile of T as $Q_{T|X,UI}(\tau|X,UI) = Q_{T|X}(\tau|X)$ for some $\tau \in (0,1)$. Meyer (1990) analyzes the effect of unemployment insurance benefit on unemployment duration, but he focuses on estimating the effect of UI on the hazard rather than quantile of the unemployment duration.*

Example 2.2.2 (Intergenerational association in timing of the first marriage)**.** *Berrington and Diamond (2000) investigate the effect of individual characteristics on timing of the first partnership formation. The individual characteristics can be divided into two groups: the first group includes current social characteristics and the other group consists of variables of family background. They show that education is a key factor that affects the timing, but one may be interested in examining intergenerational*

association in timing of formation of cohabitation at some quantile. Let X^s be the vector of the social characteristics of an individual and T^f be the timing of formation of his parents' partnership. Define $T(\tau|X^s, T^f)$ be the τ-th conditional quantile function of timing of the first partnership formation of the next generation. Then one can test the absence of intergenerational association at τ-th quantile by considering $T(\tau|X^s, T^f) = T(\tau|X^s)$.

Suppose that a researcher is interested in estimating $Q_{T|X}(\tau|x)$. Since the dependent variable T is subject to censoring and thus incomplete, the conditional quantile function may not be identified without additional structure. In this paper, I assume that T and C are conditionally independent given X in order to identify $Q_{T|X}$. Let $\Lambda(t|X)$ be the cumulative hazard function, then it can be written as

$$\Lambda(t|X) \equiv \int_0^t \frac{dF_{T|X}(s|X)}{1 - F_{T|X}(s|X)} = -\ln(1 - F_{T|X}(t|X)). \tag{2.2.2}$$

Equation (2.2.2) indicates that if $\Lambda(t|X)$ is identified, then $F_{T|X}(t|X)$ is identified and vice versa. Let $F_{C|X}(\cdot|X = x)$ be the conditional distribution of C given $X = x$. The following lemma shows that conditional independence of T and C given X is sufficient for identification of $F_{T|X}$ and $F_{C|X}$. All mathematical proofs are presented in Appendix.

Lemma 2.2.1. *Suppose that $T \perp C|X$. Then, the conditional distribution functions $F_{T|X}$ and $F_{C|X}$ are identified for almost all $X \in \mathcal{X}$.*

It is clear that X_2 being conditionally τ-quantile independent of Y on X_1 is equivalent to the fact the conditional τ-th quantile of Y given X depends only on

X_1. Thus, the null hypothesis is conditional τ-quantile independence of X_2 on X_1 and can be written as

$$Q_{T|X}(\tau|X) = Q_{T|X_1}(\tau|X_1) \text{ a.s.} \tag{2.2.3}$$

It is natural to measure the distance between $Q_{TX}(\tau|X)$ and $Q_{TX_1}(\tau|X_1)$ to test the null hypothesis in (2.2.3). If both $Q_{Y|X}(\tau|X)$ and $Q_{Y|X_1}(\tau|X_1)$ are estimated, then one can compare these two estimators by measuring the distance between them. In the case where models contain infinite-dimensional parameters, however, comparing a null model with an alternative model generally fails to achieve power against alternatives at the parametric rate $n^{-1/2}$. Many nonparametric tests that compare null models with alternative models suffer such a loss of power (e.g. Hardle and Mammen (1993); Hong and White (1995); Su and White (2008)) and some nonparametric tests involving nonparametric objects may fail to detect local alternatives at the rate $n^{-1/2}$ even if they require one to only estimate null models (e.g. Fan and Li (1996); Zheng (1998)[4]). This is because the nonparametric estimators have slower rates of convergence and thus a direct comparison involving some infinite-dimensional parameter would be costly. In this paper, I adopt Bierens's approach to specification testings (Bierens (1990); Bierens and Ploberger (1997)) which is known as the ICM test. The main idea of the ICM approach is to transform a conditional moment restriction into an infinite number of unconditional moment restrictions indexed by

[4]Zheng (1998) considers specification tests for parametric conditional quantile models and the models are parametrically estimated. However, his test statistic needs to be nonparametrically estimated and he uses the kernel method to construct the test statistic. Consequently, the test detects local alternatives at a slower rate than $n^{-1/2}$.

some nuisance parameter.

To utilize the ICM approach, the null hypothesis in (2.2.3) needs to be reformulated as a conditional moment restriction. By definition of the conditional quantile function, $\Pr(Y \leq Q_{Y|X}(\tau|X)|X) = \tau, \ \forall \tau \in \mathcal{T}$ and thus it is true that $\Pr(Y \leq Q_{Y|X_1}(\tau|X_1)|X) = \tau, \ \forall \tau \in \mathcal{T}$ under the null hypothesis. This observation turns out to be true under certain condition and one can formulate an equivalent null hypothesis to (2.2.3) with these additional conditions. The following lemma demonstrates that (2.2.3) can be rewritten in a different from that is equivalent under a uniqueness assumption.

Lemma 2.2.2. *Suppose that $F_{T|X}$ is identified. Let $\tau \in (0,1)$ be given. Suppose that the τ-th conditional quantile function $Q_{T|X}(\tau|X)$ is unique almost surely in X and that $\Pr(D = 1|X = x) \in (0,1]$ uniformly in $x \in \mathcal{X}$. Then equation (2.2.3) holds if and only if the moment condition*

$$\mathbb{E}[D_i\{\mathbf{1}(Y_i \leq Q_{T|X_1}(\tau|X_{1i})) - \tau\}|X_i] = 0 \tag{2.2.4}$$

holds almost surely.

The conditional moment characterization of the null hypothesis in (2.2.4) leads one to adopt Bierens's ICM approach. It is well-known that a conditional moment restriction is equivalent to infinitely many unconditional moment restrictions. However, Bierens (Bierens (1982, 1990)) develops a tractable way to handle the infinitely many unconditional moments by appropriately choosing an index set $\mathcal{I}$ and a weighting function $\psi(\cdot,\cdot)$ on $\mathcal{X} \times \mathcal{I}$ as follows. Specifically, the conditional moment

restriction in (2.2.4) is equivalent to the following unconditional moments

$$\mathbb{E}[D_i\{\mathbf{1}(Y_i \leq Q_{T|X_1}(\tau|X_{1i})) - \tau\}\psi(X_i, t)] = 0 \tag{2.2.5}$$

for all $t \in \mathcal{I}$. Therefore, one can consider testing (2.2.5) to see if the conditional moment restriction in (2.2.4) holds. To construct test statistics, define a stochastic process

$$J_n(t;\tau) \equiv \frac{1}{\sqrt{n}} \sum_i D_i\{\mathbf{1}(Y_i \leq Q_{T|X_1}(\tau|X_{1i})) - \tau\}\psi(X_i, t) \tag{2.2.6}$$

which is a sample analogue of (2.2.5). Since the moment restrictions in (2.2.5) are indexed by t and they need to hold for all $t \in \mathcal{I}$, either the Kolmogorov-Smirnov (KS) or the Cramer-von-Mises (CM) type tests can be used:

$$KS_n \equiv \sup_{t\in\mathcal{I}} |J_n(t;\tau)|, \tag{2.2.7}$$

$$CM_n \equiv \int_{\mathcal{I}} J_n(t;\tau)^2 d\mu(t), \tag{2.2.8}$$

where $\mu(\cdot)$ is a (probability) measure on $\mathcal{I}$. Note that these test statistics are continuous functionals of the stochastic process $J_n(\cdot;\tau)$.

Since the conditional quantile function $Q_{T|X_1}$ in (2.2.6) is unknown, it needs to be estimated. To do so, I estimate the conditional distribution function $F_{T|X_1}$ and then invert it to obtain the conditional quantile function. Since T is not completely observed in the data, I propose estimating the conditional distribution by using a local KM estimator. The local KM estimators are proposed in, for example, Dabrowska (1989); Gonzalez-Manteiga and Cadarso-Suarez (1994); Wang and Wang (2009). Specifically, Gonzalez-Manteiga and Cadarso-Suarez (1994) propose

an estimator of $F_{T|X}$,

$$\hat{F}_{T|X}(y|X=x) = 1 - \Pi_{j=1}^{n}\{1 - \frac{B_{nj}(x)}{\sum_{k=1}^{n}\mathbf{1}(Y_k \geq Y_j)B_{nk}(x)}\}^{b_j(y)},$$

where $b_j(y) = \mathbf{1}(Y_j \leq y, D_j = 1)$ and $\{B_{nk}(x) : k = 1, 2, ..., n\}$ is a sequence of nonnegative weights adding up to 1. While several types of weights are considered in the literature, most of them usually focus on the situation where the dimension of X is small. I use the local polynomial regression type weight in Kong and Xia (2017) that allows to incorporate multi-dimensionality of the covariates.

Under the conditional independence between T and C given X, lemma 2.2.1 shows that the conditional distribution of T given X, $F_{T|X}$ is identified and thus one can estimate the distribution function. It, however, is only required to estimate the conditional quantile function of T given X_1 to construct test statistics. Assuming further that C is independent of X, one can show that $T \perp C|X_1$, and thus the τ-th conditional quantile function given X_1 is estimated by

$$\hat{Q}_{T|X_1}(\tau|X_1) \equiv \inf\{y \in \mathbb{R} : \hat{F}_{1n}(y|X_1) \geq \tau\},$$

where $\hat{F}_{1n}(y|X_1 = x_1)$ is a local KM estimator of $F_{T|X_1}$. Finally, a feasible version of $J_n(t;\tau)$ is given by

$$\hat{J}_n(t;\tau) \equiv \frac{1}{\sqrt{n}}\sum_i D_i\{\mathbf{1}(Y_i \leq \hat{Q}_{T|X_1}(\tau|X_{1i})) - \tau\}\psi(X_i, t) \tag{2.2.9}$$

and feasible test statistics are defined as

$$\hat{KS}_n \equiv \sup_{t\in\mathcal{J}} |\hat{J}_n(t;\tau)|, \tag{2.2.10}$$

$$\hat{CM}_n \equiv \int_{\mathcal{J}} \hat{J}_n(t;\tau)d\mu(t). \tag{2.2.11}$$

Before proceeding, I introduce notation that will be used throughout the rest of this paper. Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space. For $x \in \mathbb{R}^d$,$||x||_E$ means the Euclidean norm of x in $\mathbb{R}^d$. Let $l^2(\mathcal{W})$ be the space of functions that are square-integrable on a set $\mathcal{W}$. Similarly, define $l^\infty(\mathcal{W})$ as the space of functions that are uniformly bounded on a set $\mathcal{W}$. For a generic function g on a set $\mathcal{W}$, $||g||_2 \equiv (\int_{\mathcal{W}} g^2 dP)^{1/2}$ and $||g||_\infty \equiv \sup_{w\in\mathcal{W}} |g(w)|$ are the L^2- and $\sup-$ norm, respectively. The expectation of g is denoted by $\mathbb{E}g \equiv \int g(w)dF_W(w)$, where $F_W(\cdot)$ is the distribution function of W. For a sequence of random maps $X_n : \Omega \to \mathbb{R}$ and a random variable X, $X_n \Rightarrow X$ ($X_n \xrightarrow{d} X$, resp.) indicates that X_n converges weakly (in distribution, resp.) to X in the sense of Definition 1.3.3 in van der Vaart and Wellner (1996).

2.3 Asymptotic theory

In this section, I develop the asymptotic theory for test statistics $\hat{KS}_n$ and $\hat{CM}_n$. Since the test statistics are continuous functionals of the process $\hat{J}_n(\cdot;\tau)$, I first establish the weak convergence of $\hat{J}_n(\cdot;\tau)$. Then the asymptotic distribution of the test statistics can be obtained by the continuous mapping theorem.

2.3.1 Assumptions

Let $p_0(X) \equiv \Pr(D = 1|X)$ and $U_i \equiv T_i - Q_{T|X}(\tau|X_i)$. For a real number p, denote the largest integer smaller than p by $\lfloor p \rfloor$. Let $\mathcal{C}^p(\mathcal{X})$ be the space of $\lfloor p \rfloor$-times continuously differentiable real-valued functions on $\mathcal{X}$. Denote the differential operator by $\mathcal{D}$ and let $\mathcal{D}^\omega \equiv \frac{\partial^{[\omega]}}{\partial x_1^{\omega_1} ... \partial x_d^{\omega_d}}$. Define the *Hölder norm* for $h \in \mathcal{C}^p(\mathcal{X})$ as following :

$$||h||_{\Lambda^p} \equiv \sup_{[\omega] \leq \lfloor p \rfloor, x \in \mathcal{X}} |\mathcal{D}^\omega h(x)| + \sup_{[\omega] = \lfloor p \rfloor} \sup_{x,y \in \mathcal{X}, x \neq y} \frac{|\mathcal{D}^\omega h(x) - \mathcal{D}^\omega h(y)|}{||x - y||_E^{p - \lfloor p \rfloor}} < \infty,$$

where $p - \lfloor p \rfloor \in (0, 1]$ is the *Hölder exponent*. Then the class of functions $\Lambda^p(\mathcal{X}) \equiv \{h \in \mathcal{C}^p(\mathcal{X}) : ||h||_{\Lambda^p} < \infty\}$ is called a *Hölder class*. A *Hölder ball* with radius $R > 0$ is defined by $\Lambda_R^p(\mathcal{X}) \equiv \{h \in \Lambda^p(\mathcal{X}) : ||h||_{\Lambda^p} \leq R\}$ for some $R \in (0, \infty)$. Let $\mathcal{G}$ be a class of functions and $||\cdot||$ be a norm on $\mathcal{G}$. Let $\mathcal{G}_\delta(g_0) \equiv \{g :\in \mathcal{G} : ||g - g_0|| < \delta, g_0 \in \mathcal{G}\}$. A function m on $\mathcal{G}$ is called *pathwise differentiable* at $g \in \mathcal{G}_\delta(g_0)$ in the direction $[\bar{g} - g]$ if $\{g + t(\bar{g} - g) : t \in [0, 1]\} \subset \mathcal{G}$ and the limit

$$\lim_{t \to 0} \frac{m(g + t(\bar{g} - g)) - m(g)}{t}$$

exists. I consider the following assumptions.

Assumption 2.3.1. *(i) The data* $\{W_i \equiv (Y_i, D_i, X_i')'\}_{i=1}^n$ *are i.i.d; (ii)* $(T, X') \perp C$*; (iii)* $\mathcal{X}$ *and* $\mathcal{X}_1$ *are compact and convex subsets of* $\mathbb{R}^{d_x}$ *and* $\mathbb{R}^{d_1}$*, respectively; (iv)* $p_0(x) \in (0, 1]$ *for all* $x \in \mathcal{X}$*.*

Assumption 2.3.2. *There exists* $R > 0$ *such that* $Q_{T|X}(\tau|X = \cdot) \in \Lambda_R^{p_1}(\mathcal{X}_1)$*, where* $p_1 > \frac{d_1}{2}$*.*

Assumption 2.3.3. *The conditional distribution function $F_{U|X}$ admits its density function $f_{U|X}(t|X = x)$ satisfying the following condition: (i) $f_{U|X}(t|X = \cdot) \in \Lambda^{p_2}(\mathcal{X})$ for some $p_2 > 0$, uniformly in t in a neighborhood of $t = 0$; (ii) $f_{U|X}(0|X)$ is bounded away from zero uniformly in $X \in \mathcal{X}$; (iii) The first-order derivative with respect to t is bounded and continuous, uniformly on a neighborhood of $t = 0$ and uniformly in $X \in \mathcal{X}$.*

Assumption 2.3.4. *The marginal distribution functions of X and X_1 have their own density functions f_X and f_{X_1} with the following properties: (i) $f_X, f_{X_1} \in \Lambda^{p_3}(\mathcal{X})$ for some $p_3 > 0$; (ii) $f_X(\cdot)$ and $f_{X_1}(\cdot)$ are positive on $\mathcal{X}$ and $\mathcal{X}_1$, respectively.*

Assumption 2.3.5. *(i) $K_F(\cdot)$ is a symmetric probability density function on $\mathbb{R}^{d_1}$ with finite second moments and bounded first order derivative; (ii) the bandwidth associated with K_F, h_{Fn}, satisfies the following conditions: $h_{Fn} \to 0$, $\frac{\log n}{\sqrt{n} h_{Fn}^{d_1}} \to 0$, and $n h_{Fn}^{d_1 + \frac{4}{3}(p_2+1)} \to 0$.*

Assumption 2.3.6. *The class $\Psi \equiv \{\psi(X_i, t) : t \in \mathcal{I}\}$ satisfies the following conditions: (i) The weighting function $\psi(\cdot, \cdot) : \mathcal{X} \times \mathcal{I} \to \mathbb{R}$ is uniformly bounded and $\mathcal{I}$ is a compact subset of $\mathbb{R}^{d_x}$; (ii) $\psi(X_i, t) = \mathbf{w}(X_i' t)$ for some $\mathbf{w}(\cdot)$ real analytic and non-polynomial and there exists a function $G_\Psi(\cdot)$ such that for any $t_1, t_2 \in \mathcal{X}$, $|\psi(X, t_1) - \psi(X, t_2)| \leq G_\Psi(X)||t_1 - t_2||_E$ with $\mathbb{E}[G_\Psi(X_i)^2] < \infty$.*

The condition (ii) in Assumption 2.3.1is satisfied when C is completely random, and guarantees the identification of conditional quantile functions of T given X[5]. This restriction is considered in, for example, Bang and Tsiatis (2000) and Hon-

[5]The identification of $F_{T|X}$ is based on lemma 2.2.1. To be specific, for any bounded and

ore et al. (2002). It also implies that $T \perp C|X_1$ and thus one can use a local KM estimator of $F_{T|X_1}$ to estimate the conditional quantile function $Q_{T|X_1}$. Condition (iii) is a support condition and the last condition (condition (iv)) implies that not all observations are censored. This condition is crucial for the equivalence between (2.2.3) and (2.2.4).

Assumptions 2.3.2 and 2.3.3 impose smoothness of the conditional quantile functions and the conditional density functions, respectively. Note that Assumption 2.3.3 implies that $F_{T|X_1}(\cdot|x_1)$ and $F_{T|X}(\cdot|x)$ are Lipschitz continuous in y for all $x_1 \in \mathcal{X}_1$ and $x \in \mathcal{X}$, respectively. Moreover, since $F_{T|X_1}(\cdot|x_1)$ is pathwise differentiable for all $x_1 \in \mathcal{X}_1$ under Assumption 2.3.3, one can show that for any $q \in \Lambda_R^p(\mathcal{X}_1)$ and $\delta_n \downarrow 0$,

$$\sup |F_{T|X_1}(\tilde{q}|x_1) - F_{T|X_1}(q|x_1) - (\tilde{q} - q) f_{T|X_1}(q|x_1)| = O(||\tilde{q} - q||_\infty^2), \qquad (2.3.1)$$

where the supremum is taken over $x_1 \in \mathcal{X}_1$ and $\tilde{q} \in \Lambda_R^p(\mathcal{X}_1)$ such that $||\tilde{q} - q||_\infty \leq \delta_n$. Assumption 2.3.4 considers the smoothness of the marginal density functions of X_1 and X and guarantees that the sparsity function- $1/f_X(\cdot)$ - is well-defined on $\mathcal{X}$ and imposes smoothness of the marginal density functions of X and X_1. Assumption 2.3.5 restricts the kernel function used to estimate the local KM estimator and specifies the rate of the bandwidth h_{Fn}.

Assumption 2.3.6 restricts the class of weighting functions used to construct

continuous functions g and h, one can show that $\mathbb{E}[g(T)h(C)|X] = \mathbb{E}[\mathbb{E}[g(T)h(C)|X,T]|X] = \mathbb{E}[g(T)\mathbb{E}[h(C)|X,T]|X] = \mathbb{E}[h(C)]\mathbb{E}[g(T)|X] = \mathbb{E}[h(C)|X]\mathbb{E}[g(T)|X]$. Therefore, one obtains that $T \perp C|X$, and identification is achieved by lemma 2.2.1.

the unconditional moments in (2.2.5). Bierens (1990) takes $\psi(x,t) = \exp(it'x)$, where $i^2 = -1$, and Bierens and Ploberger (1997) use $\psi(x,t) = \exp(x't)$. Stinchcombe and White (1998) show that more classes of functions can be considered and they call such functions GCR functions. Assumption 2.3.6 comes from Corollary 3.9 in Stinchcombe and White (1998) and one can choose the logistic distribution function, the normal distribution or density function, or the exponential function for $\mathbf{w}(\cdot)$. Note that, since $\mathbf{w}(\cdot)$ is assumed to be analytical and the support of X and the index set $\mathcal{I}$ are assumed to be compact, the first- and the second- order derivatives of $\psi(x,t)$ with respect to x are uniformly bounded on $\mathcal{X} \times \mathcal{I}$.

As an alternative class of weighting functions, one may choose $\Psi^I \equiv \{\mathbf{1}(X_i \leq t) : t \in \mathcal{I}\}$ as in Volgushev et al. (2013). Even if the indicator function is not GCR, the class of functions Ψ^I has been used to construct test statistics (e.g. Delgado and Manteiga (2001); Escanciano and Goh (2014); Volgushev et al. (2013)). However, there are several advantages of using the class of weighting functions in Assumption 2.3.6 over using Ψ^I (see, for example, (Huang et al., 2016, pp.1444-1445)) as the conditional distribution function of X_2 on X_1 needs to be smooth enough when using Ψ^I. Lastly, it is worth noting that one can replace the condition that $\mathbf{w}(\cdot)$ is analytical with one that $\mathbf{w}(\cdot)$ is *smooth enough* in terms of that $\mathbf{w}(\cdot)$ is just finitely-many continuously differentiable without any cost, but being analytical will be imposed throughout this paper. Lastly, I take the distribution function of X and the support of X for the measure $\mu(\cdot)$ in (2.2.8) and (2.2.11) and the index set $\mathcal{I}$, respectively.

2.3.2 Weak convergence

I first establish the weak convergence of the infeasible process $J_n(\cdot;\tau)$. The next theorem establishes that the empirical process $J_n(\cdot)$ converges weakly to a Gaussian limit under Assumptions 2.3.1 and 2.3.6:

Theorem 2.3.7. *Suppose that Assumptions 2.3.1 and 2.3.6 hold. Then*

$$J_n(\cdot;\tau) \Rightarrow \mathbb{G}(\cdot) \text{ in } l^\infty(\mathcal{I}),$$

where $\mathbb{G}(\cdot)$ is a Gaussian process with zero mean and covariance kernel

$$\Sigma^p(t_1,t_2) \equiv \mathbb{E}[\tau(1-\tau)p_0(X_i)\psi(X_i,t_1)\psi(X_i,t_2)].$$

As mentioned before, the conditional quantile function needs to be estimated and estimation of the function introduces sampling error that must be dealt with. To investigate the asymptotic behavior of the feasible process $\hat{J}_n(t;\tau)$, let $\bar{\psi}(X_{1i},t) \equiv \mathbb{E}[\psi(X_i,t)|X_{1i}]$ and consider a decomposition of $\hat{J}_n(t;\tau)$ as follows:

$$\begin{aligned}
\hat{J}_n(t;\tau) &= \frac{1}{\sqrt{n}}\sum_i D_i\{\mathbf{1}(Y_i \le \hat{Q}_{T|X_1}(\tau|X_{1i})) - \tau\}\psi(X_i,t) \\
&= J_n(t;\tau) + \frac{1}{\sqrt{n}}\sum_i D_i\{\mathbf{1}(Y_i \le \hat{q}_{1i}) - \mathbf{1}(Y_i \le q_{1i})\}\psi(X_i,t) \\
&= J_n(t;\tau) + \nu_n^p(t,\hat{q}_1;\tau) - \nu_n^p(t;q_1;\tau) + \hat{J}_{sn}(t;\tau), \qquad (2.3.2)
\end{aligned}$$

where

$$\nu_n^p(t,q;\tau) \equiv \frac{1}{\sqrt{n}}\sum_i D_i\{\mathbf{1}(Y_i \le q_i) - F_{T|X_1}(q_i|X_{1i})\}\bar{\psi}(X_{1i},t),$$
$$\hat{J}_{sn}(t;\tau) \equiv \frac{1}{\sqrt{n}}\sum_i D_i\{F_{T|X_1}(\hat{q}_{1i}|X_{1i}) - F_{T|X_1}(q_{1i}|X_{1i})\}\bar{\psi}(X_{1i},t).$$

Let

$$\xi(Y_j,D_j,y,x) \equiv [\frac{\mathbf{1}(Y_j \le y)\cdot D_j}{(1-F_{T|X_1}(Y_j|x))(1-F_{C|X_1}(Y_j|x))} - \int_0^{\min(Y_j,y)} \frac{f_{T|X_1}(s|x)ds}{(1-F_{T|X_1}(s|x))^2(1-F_{C|X_1}(s|x))}],$$

then it can be shown that $\mathbb{E}[\xi(Y_j,D_j,y,x)] = 0$ and $Var(\xi(Y_j,D_j,y,x)) < \infty$ for any y and x (cf. Gonzalez-Manteiga and Cadarso-Suarez (1994)). To establish the weak convergence of $\hat{J}_n(\cdot;\tau)$, I employ a stochastic equicontinuity argument and the theory of U-processes. To be more specific, a stochastic argument can be utilized in this way: if one can show that the process $\nu_n^p(\cdot,\cdot;\tau)$ is stochastically equicontinuous, it can be shown that $\nu_n^p(t,\hat{q}_1;\tau) - \nu_n^p(t;q_1;\tau) = o_p(1)$ uniformly in $t \in \mathcal{I}$ by stochastic equicontinuity since the estimated conditional quantile function converges to $Q_{T|X_1}(\tau|X_1 = x_1)$ uniformly in x_1 under several conditions. Moreover, the "smoothed" term $\hat{J}_{sn}(\cdot;\tau)$ can be handled as follows: one can approximate $F_{T|X_1}(\hat{q}_{1i}|X_{1i})$ up to the second-order approximation with respect to $\hat{q}_{1i}$. Since it is possible to make the second-order term $o_p(n^{-1/2})$ under some conditions on the bandwidth, the process is asymptotically equivalent to the first-order approximation of $F_{T|X_1}(\hat{q}_{1i}|X_{1i}) - F_{T|X_1}(q_{1i}|X_{1i})$. Then I use the theory of U-processes to deal with this first-order term which contains $\hat{q}_{1i} - q_{1i}$. The next theorem shows that the

feasible process $\hat{J}_n(\cdot;\tau)$ converges weakly to a Gaussian process in $l^\infty(\mathcal{I})$:

Theorem 2.3.8. *Suppose that Assumptions 2.3.1-2.3.6 hold. Define*

$$\tilde{\psi}(X_{1i},t) \equiv \mathbb{E}[p_0(X_i)\bar{\psi}(X_{1i},t)|X_{1i}],$$

$$m(W_i,t) \equiv \psi(X_i,t)D_i\{\mathbf{1}(Y_i \leq Q_{T|X_1}(\tau|X_{1i})) - \tau\} - (1-\tau)\tilde{\psi}(X_{1i},t)\xi(Y_i,D_i,q_{1i},X_{1i}).$$

Then,

$$\hat{J}_n(\cdot;\tau) \Rightarrow \hat{\mathbb{G}}(\cdot) \text{ in } l^\infty(\mathcal{I}),$$

where $\hat{\mathbb{G}}(\cdot)$ is a Gaussian process with zero mean and covariance kernel

$$\hat{\Sigma}(t_1,t_2) = E[m(W_i,t_1)m(W_i,t_2)].$$

Finally, one can derive the asymptotic distributions of the test statistics under the null hypothesis by using theorem 2.3.8 and the continuous mapping theorem. The asymptotic distributions under the null are given in the following corollary.

Corollary 2.3.9. *Suppose that the conditions in theorem 2.3.8 are satisfied. Then*

$$\hat{KS}_n \xrightarrow{d} \sup_{t\in\mathcal{I}} |\hat{\mathbb{G}}(t)|,$$
$$\hat{CM}_n \xrightarrow{d} \int \hat{\mathbb{G}}(t)^2 d\mu(t).$$

2.3.3 Power Properties

Now I examine the power properties and show that the tests have non-trivial power against local alternatives at the parametric rate. To investigate the power

properties, consider local alternatives as following:

$$Q_{T|X}(\tau|X) = Q_{T|X_1}(\tau|X_1) + \frac{1}{\sqrt{n}}\tilde{Q}(\tau|X) \tag{2.3.3}$$

for some bounded function $\tilde{Q}(\tau|X)$.

Assumption 2.3.10. *The conditional quantile function $Q_{T|X}$ under the local alternative in (2.3.3) belongs to $\Lambda^{p_A}_{\tilde{R}}(\mathcal{X})$ for some $\tilde{R} > 0$ and $p_A > \frac{d_x}{2}$, for all n.*

The following theorem demonstrates that the tests can detect local alternatives at $\sqrt{n}$-rate.

Theorem 2.3.11. *Suppose that Assumptions 2.3.1 and 2.3.3 through 2.3.10 hold. Under the local alternative in (2.3.3),*

$$\hat{J}_n(\cdot;\tau) \Rightarrow \hat{\mathbb{G}}(\cdot) - R_a(\cdot) \text{ in } l^\infty(\mathcal{J}),$$

where $R_a(t) \equiv \mathbb{E}[p_0(X_i)\psi(X_i,t)f(Q_{T|X_1}(\tau|X_{1i})|X_i)\tilde{Q}(\tau|X_i)]$.

Corollary 2.3.12. *Suppose that the conditions in theorem 2.3.11 are satisfied. Then, under the local alternative in (2.3.3),*

$$\hat{KS}_n \xrightarrow{d} \sup_{t\in\mathcal{J}} |\hat{\mathbb{G}}(t) - R_a(t)|,$$
$$\hat{CM}_n \xrightarrow{d} \int_{\mathcal{J}} |\hat{\mathbb{G}}(t) - R_a(t)|^2 d\mu(t).$$

2.4 Subsampling Approximation

Since the asymptotic distribution of the process $\hat{J}_n(\cdot)$ depends on the data generating processes of X_{1i} and X_i, it is hard to calculate critical values for the

test statistics $\hat{KS}_n$ and $\hat{CM}_n$[6] and this difficulty has drawn attention to methods of obtaining asymptotically valid critical values. Bierens and Ginther (2001) and Bierens and Ploberger (1997) propose a method to obtain critical values. Their approach is to use upper bounds on the critical values. but these bounds do not deliver asymptotically correct size. Another way to obtain critical values is to use a variant of resampling methods, and some bootstrap methods have been developed in several studies (e.g. Delgado and Manteiga (2001); Volgushev et al. (2013)).

I employ a subsampling method to approximate the asymptotic distributions of the test statistics. Subsampling is widely used to overcome difficulty in obtaining critical values of statistics. In QR models, subsampling is considered in Chernozhukov and Fernández-Val (2005), Escanciano and Velasco (2010) and Whang (2006a) as a way to mimic asymptotic distributions of statistics. One of the main reasons for using a subsampling instead of the bootstrap is that subsampling is much more effective than bootstrap in terms of its applicability. Specifically, Politis and Romano (1994) show that under mild conditions[7] subsampling can be used to approximate the asymptotic distribution of any statistic. Another reason is that the bootstrap may be problematic if a statistic is non-smooth and in that case the bootstrap needs to be carefully applied[8]. Lastly, the subsampling method proposed in

[6] Bierens and Ploberger (1997) and Chen and Fan (1999) derive the asymptotic distribution of the Cramer-von-Mises type statistic in parametric and nonparametric specification testings, respectively, and they show that the test statistic is asymptotically equivalent to an infinite sum of weighted $\chi^2(1)$ random variables.

[7] These conditions include the weak convergence of statistic and the conditions on the size of subsamples.

[8] The process $\hat{J}_n(\cdot;\tau)$ contains an indicator function, and the bootstrap may be challenging due to the non-smoothness of the indicator function.

this paper does not require one to estimate the influence function of $\hat{J}_n(t;\tau)$ and hence it is easy to implement. One can refer to Sant'Anna (2016) for the validity of the multiplier bootstrap in a similar situation[9] .

The subsampling procedure employed in this paper is the same as the one in Whang (2006a). To describe the subsampling procedure, define the distribution functions of $\hat{KS}_n$ and $\hat{CM}_n$ as following:

$$F_n^{KS}(z) \equiv \Pr(\hat{KS}_n \leq z);\ F_n^{CM}(z) \equiv \Pr(\hat{CM}_n \leq z).$$

Let $\{W_i, W_{i+1}, ..., W_{i+b-1}\}$ be a subsample from the original sample $\{W_j : j = 1, 2, ..., n\}$ of size b, where $i = 1, 2, ..., n-b+1$. Let $\hat{J}_{n,b,i}(t,\tau)$ be the process $\hat{J}_n(t,\tau)$ that is computed by only using the subsample $\{W_i, W_{i+1}, ..., W_{i+b-1}\}$ for $i = 1, 2, ..., n-b+1$. Then $\hat{KS}_{n,b,i}$ and $\hat{CM}_{n,b,i}$ are defined by the same way of (2.2.11) but with $\hat{J}_{n,b,i}(t,\tau)$. To approximate the distributions $F_n^{KS}(\cdot)$ and $F_n^{CM}(\cdot)$, consider the following objects:

$$\hat{F}_{n,b}^{KS}(z) \equiv \frac{1}{n-b+1}\sum_{i}^{n-b+1} \mathbf{1}(\hat{KS}_{n,b,i} \leq z);\ \hat{F}_{n,b}^{CM}(z) \equiv \frac{1}{n-b+1}\sum_{i}^{n-b+1} \mathbf{1}(\hat{CM}_{n,b,i} \leq z).$$

Let $c_n^{KS}(\alpha)$ and $c_n^{CM}(\alpha)$ be the $(1-\alpha)$-th quantiles of $F_n^{KS}(\cdot)$ and $F_n^{CM}(\cdot)$ under the null hypothesis. In the same way, let $\hat{c}_{n,b}^{KS}(\alpha)$ and $\hat{c}_{n,b}^{CM}(\alpha)$ be the $(1-\alpha)$-th quantiles of $\hat{F}_{n,b}^{KS}(\cdot)$ and $\hat{F}_{n,b}^{CM}(\cdot)$, respectively. The following theorem demonstrates that the subsampling provides asymptotically valid size of test under the null hypothesis.

[9] For the tests in the standard QR, one can refer to Volgushev et al. (2013) for bootstrap validity. They suggest using the bootstrap to approximate the asymptotic distribution of their test statistic.

Lastly, define

$$F^{KS}(z) \equiv \Pr(\sup_{t\in\mathcal{J}} |\hat{\mathbb{G}}(t)| \leq z);\ F^{CM}(z) \equiv \Pr(\int_{\mathcal{J}} |\hat{\mathbb{G}}(t)|^2 d\mu(t) \leq z);$$

and let $c_\infty^{KS}(\alpha)$ and c_∞^{CM} be the α-th quantiles of F^{KS} and F^{CM}, respectively.

Theorem 2.4.1. *Suppose that $b/n \to 0$ and $b \to \infty$ as $n \to \infty$.*

(i) If conditions in theorem 2.3.8 are satisfied, then under the null hypothesis,

$$c_{n,b}^{KS}(\alpha) \xrightarrow{p} c_\infty^{KS}(\alpha);\ c_{n,b}^{CM}(\alpha) \xrightarrow{p} c_\infty^{CM}(\alpha),$$

and

$$\Pr(\hat{KS}_n > c_{n,b}^{KS}(\alpha)) \to \alpha,$$
$$\Pr(\hat{CM}_n > c_{n,b}^{CM}(\alpha)) \to \alpha.$$

(ii) If conditions in theorem 2.3.11 hold, then under the local alternative in (2.3.3), then

$$\Pr(\hat{KS}_n > c_{n,b}^{KS}(\alpha)) \to \Pr(\sup_{t\in\mathcal{J}} |\hat{\mathbb{G}}(t) - R_a(t)| > c_\infty^{KS}(\alpha)),$$
$$\Pr(\hat{CM}_n > c_\infty^{CM}(\alpha)) \to \Pr(\int_{\mathcal{J}} |\hat{\mathbb{G}}(t) - R_a(t)|^2 d\mu(t) > c_\infty^{CM}(\alpha)).$$

Remark 2.4.2. *It is possible to calculate the test statistics over all $\binom{n}{b}$ subsamples to obtain $\hat{F}_{n,b}^{KS}(\cdot)$ and $\hat{F}_{n,b}^{CM}(\cdot)$, as shown in Chernozhukov and Fernández-Val (2005), but this approach is computationally much more burdensome than the procedure given above.*

2.5 Conclusion

In this paper, I propose nonparametric tests for conditional quantile independence for a class of models with duration outcomes. Duration outcomes are usually subject to censoring, therefore I use a local KM estimator to estimate the conditional quantile function. Since conditional quantile independence can be formalized as a conditional moment restriction, I adopt Bierens's ICM approach to construct test statistics. I show that the test statistics are continuous functionals of a Gaussian process under suitable conditions and that the tests have non-trivial power against local alternatives at the parametric rate even if the tests are nonparametric. Since the asymptotic distributions of test statistics depend on the data generating process, I provide a subsampling method to obtain the critical values and establish the validity of the subsampling method.

There are several areas for further work. First, one can consider tests that are uniform in the quantile index τ. The tests proposed in this paper are pointwise in the sense that they focus on conditional quantile independence at a specific $\tau \in (0, 1)$. One of advantages of uniform test is that they can partly answer the question related to arbitrariness of choice of a specific quantile. Uniformity in quantile index is an important issue and many studies have considered uniform inference in QR (Koenker and Bassett (1982), Koenker and Machado (1999), Koenker and Xiao (2002), Su and White (2012), and so on). In a similar spirit to Manski (1988), one may wonder why a variable is not significant at a specific quantile but at others. In addition, the effects of covariates tend to be continuous in the quantile index and similar within a range of quantile levels (e.g. (Koenker and Hallock, 2001, p.150)) in many situations.

Therefore, uniform tests would be more preferable to pointwise tests in the sense that they can handle arbitrariness of choice of τ and possess more empirical content. Moreover, the uniform tests are closely related to tests for conditional independence. If one is interested in the uniform test of conditional quantile independence over the quantile $\tau \in [0,1]$, this test becomes a test for conditional independence and several studies have proposed tests for conditional independence (e.g. Su and White (2008); Song (2009); Huang et al. (2016)). While pointwise conditional quantile independence and conditional independence are rather extreme hypotheses, uniform quantile independence can be regarded as an intermediate notion connecting those two extreme cases[10].

Secondly, it is also expected that the tests in this paper can be extended to the case where one is interested in conducting semiparametric or other nonparametric specification tests (e.g. tests for additive separability and partially linear structure). Related to this extension, another potential direction would be to consider testing with different estimation strategies for the conditional quantile function. Recently, Belloni et al. (2016) propose a series-based estimation method for the conditional quantile processes and establish the asymptotic theory, and Chao et al. (2016) also study quantile processes with series estimation and provide conditions under which the series estimator of a conditional quantile process converges weakly to a Gaussian process. Series estimation methods are convenient for imposing some structure, such as additive separability and partial linearity, on the model. Based on the results in

[10]The author is currently developing uniform tests.

those recent studies, one could develop some test procedures with series estimation[11].

Lastly, one can consider specification testing for conditional quantile functions with endogenous censoring. This paper assumes that censoring is independent of the covariates and the latent dependent variable, but this assumption is not likely to be satisfied in many empirical examples and several studies consider models with endogenous censoring (e.g. Khan and Tamer (2009), Khan et al. (2011), and Fan and Liu (2013)). Therefore, it would be worth extending the tests to the case of endogenous censoring.

[11] For specification tests based on series estimation, one can refer to, for example, Hong and White (1995), Donald (1997), and Li et al. (2003).

Chapter 3

Identification and Confidence Regions for Treatment Effect and its Distribution under Stochastic Dominance

3.1 Introduction

Program evaluation has been widely studied in the econometrics and statistics literature. An evaluation problem requires measuring the difference in the outcomes of possible states, and this leads us to formulate a counterfactual model to evaluate the effect of a program or a policy. Since one is only able to observe the outcome of the realized state, it is impossible to find the treatment effect without additional assumptions, which is the main obstacle to finding the treatment effect. In other words, the central question for program evaluation is, how can one identify the counterfactual outcomes?

This paper is aimed to investigate the identification of treatment effect (TE) and its distribution when the treatment is endogenous. To analyze these problems, I construct a counterfactual model containing a binary treatment and two potential outcome variables with unknown functional forms. I do not assume any structure of the outcome functions and the selection rule. Hence, the model considered in this paper is based on the treatment effects approach[1] which is commonly used in

[1] To clarify the notion of the treatment effects approach, see Heckman and Vytlacil (2007). In short, I use the term as a contrasted notion of the structural equations approach.

the statistics literature. To resolve the problem caused by endogeneity of selection, I adopt Manski's approach (Manski (1990)) which replaces unknown components with identifiable components in the worst case.

The main identification objects are the marginal distribution and quantile functions of the potential outcomes, and the distribution of the TE. Once the quantile functions of the potential outcomes are identified, one can also identify the quantile treatment effect (QTE) which is an important object in both theoretical and applied econometrics. Most of studies following the treatment effects approach have focused on average treatment effect (ATE) rather than QTE or some distributional features of the TE. The QTE, however, is more informative than the ATE. Specifically, the QTE provides much more information on the heterogeneity in the TE than does the ATE. This is because the heterogeneous factors are integrated out when calculating the ATE. Bitler et al. (2006)[2] found that the ATE missed the heterogeneous effects of the policy reforms and that the heterogeneous factors result in the difference between the ATE and the QTE. In this regard, the QTE is more informative than the ATE since the former takes the heterogeneous effects into consideration.

Identification and estimation of QTE have been considered by a number of studies. Firpo (2007) investigates identification and estimation of QTE under the unconfoundedness condition and establishes that his estimators achieve the semi-parametric efficiency bound. Donald and Hsu (2014) consider estimation of distribution functions of the potential outcomes under the unconfoundedness condition

[2]They investigated the effect of policy reforms on the welfare and compared the ATE with the QTE. The estimation results of the QTE across the distributions of earnings, transfer payments, and total measurable income indicate that the effect of welfare reforms is heterogeneous.

and propose inverse probability weighting estimators. They also provide the asymptotic theory for the quantile functions as well as the distribution functions of the potential outcomes. The main difference from those papers is that I do not impose the unconfoundedness condition nor the full exogeneity of the treatment.

For the distribution of the TE, I follow the approach introduced by Fan and Park (2010). The distribution of the TE is required to examine whether the treatment is effectively being implemented. In particular, the distribution of the TE is called for the case where the benefit from the treatment is non-transferrable (Heckman and Vytlacil (2007)). Since Fan and Park (2010) assume that the treatment is exogenously assigned, their approach does not address the selection issue. I show that one can still get a bound on the distribution of the TE even when the treatment is endogenous.

In many cases, the bounds on the objects presented above may not be informative when the bounds on the counterfactual components are too broad. To deal with this problem, many studies have utilized some distributional assumptions to get tighter bounds on the parameters of interest (Manski (1997); Manski and Pepper (2000); Blundell et al. (2007)). In this paper, I consider stochastic dominance relations between counterfactual outcome variables to tighten the bounds.

Stochastic dominance can be used not only for studies on conventional inequality measurement but also for constructing an economic model. Heckman et al. (1997) take the first and second order stochastic dominance to impose a dependence

structure between two counterfactual outcomes[3]. A more relevant example can be found in Blundell et al. (2007). Blundell et al. (2007) investigate changes in the distribution of wage while imposing first-order stochastic dominance and a median restriction. These assumptions were motivated by the standard labor supply model describing positive selection[4], and it is shown that the restrictions can tighten the bounds on the distribution of wages. I provide a general result which is directly related to Blundell et al. (2007) and also explore other versions of stochastic dominance.

This paper is expected to contribute to the literature in two ways. First, it suggests two versions of stochastic dominance which are consistent with some economic theories, and presents the identification results under these assumptions. Second, this paper also identifies the distribution of the TE in the case where the treatment is not randomly assigned.

The rest of this paper consists of five sections. I briefly review previous studies in Section 3.2 and give the identification results in Section 3.3. Section 3.4 provides consistent estimators of the bounds derived in the previous section and Section 3.5 gives an empirical example on the return to college. Section 3.6 concludes and discusses potential extensions of this paper.

[3]The dependence structure is made from the rational choice model. The rational choice model assumes that an agent participates in a program if the expected utility from participation is greater than or equal to one from non-participation. Since the utility functions are assumed to be concave, without loss of generality, this can be represented by second-order stochastic dominance.

[4]Positive selection in their paper means that wages of people employed are more likely to be higher than wages of the unemployed.

3.2 Previous Studies

There are a myriad of studies examining the identification of the TE under nonparametric models. The main identification objects and the underlying economic models vary across studies, but one can divide the literature into two groups as Heckman and Vytlacil (2007) argue. The studies in the first group assume economic models which explain the data generating process, and this is why they are called the structural equations approach. In contrast, the other group employs a counterfactual model without constructing an underlying economic model, and this approach is referred to as the treatment effects approach.

Most of the studies following the structural approach use a triangular system of equations as an underlying economic model and impose some variants of monotonicity on the model. They also assume that there exist instrumental variables which affect the outcome variables only through the endogenous variables[5]. Among the studies adopting the structural approach, Chernozhukov and Hansen (2005) and Jun et al. (2011) consider the identification of the QTE under a triangular system of equations. They identify the quantiles of the outcome equations and obtain the QTE from the identified equations [6].

Chernozhukov and Hansen (2005) show point-identification results of the

[5] There are some studies which do not introduce instrumental variables to the model. For example, Chesher (2005) considers a triangular system of equations and imposes (local) exclusion and/or exogeneity condition for the regressors instead of using an instrumental variable.

[6] Imbens and Newey (2009) also consider identification of the (quantiles of) outcome function in a triangular system by using a control function approach. Their approach, however, does not work for binary (or discrete) treatment because the endogenous variable is assumed to be continuously distributed.

quantiles of the outcome function by imposing strict monotonicity of the outcome functions in their error term and rank similarity (or rank invariance). They assume that the endogenous treatment is binary or discrete[7] and instrumental variables play a crucial role in getting rank similarity. Jun et al. (2011) provide partial identification results of the quantiles of the outcome function. They relax some restrictions imposed by Chernozhukov and Hansen (2005) such as rank similarity and strict monotonicity. The identification strategy of Jun et al. (2011) relies on the Dynkin system and the dependence structure between error terms.

For the ATE, Vytlacil and Yildiz (2007) consider a situation in which the outcome function is weakly separable and the endogenous variable is binary. They propose a Wald type estimator to find a specific value of the covariate X which compensates for the TE, holding the propensity score constant, by varying the instrumental variable. A similar model is examined by Jun et al. (2012). As in Jun et al. (2011), their identification strategy for the ATE is to use the Dynkin system.

On the other hand, studies following the treatment effects approach do not specify an underlying model and their identification strategies are mainly based on Manski (1990). Manski (1990) studied bounds on the ATE and the main idea is to bound the counterfactual components by using prior information such as logical bounds and/or other assumptions.

Manski (1997) and Manski and Pepper (2000) consider the identification of the ATE under several monotonicity assumptions. Manski (1997) introduces the

[7] They also consider the case where the treatment is a continuous random variable, but their main result is the identification with a discrete treatment.

monotone treatment response (MTR) assumption, which means that the outcome is monotone in the treatments for all individuals[8]. In Manski and Pepper (2000), the model is equipped with other monotonicity assumptions, e.g. the monotone instrumental variables (MIV) assumption and the monotone treatment selection (MTS) assumption[9]. These studies demonstrate that distributional assumptions may improve bounds on the parameters of interest in terms of informativeness.

For identification and estimation of the QTE with a potential outcome model, Abadie et al. (2002) provide identification and estimation of the QTE when a binary treatment is endogenously determined. They adapt the local average treatment effect framework (Imbens and Angrist (1994)) and achieve point-identification of the QTE.

For the distribution of the TE, Fan and Park (2010) provide an identification result for this object without introducing structural equations. They exploit copula theory to identify the distribution of the TE when the marginal distribution functions of potential outcomes are directly identified from the data. One of drawbacks of their results is that they only consider the case where the treatment is randomly assigned, which is very rare in the practice.

3.3 Identification

Let D be a binary variable that indicates whether a person gets the treatment or not, i.e. $D = 1$ if the person gets the treatment and $D = 0$ if the person does

[8]The MTR means that for given two treatments t_1 and t_2, $t_2 \geq t_1$ implies that $Y_{t_2} \geq Y_{t_1}$. The set of treatments is assumed to be ordered.

[9]The MTS means that for given two treatments t_1 and t_2, $t_2 \geq t_1$ implies that $E[Y_j|D = t_2] \geq E[Y_j|D = t_1]$ for all $j \in \mathcal{T}$, where $\mathcal{T}$ is an ordered set of treatments.

not get the treatment. Let Y_d denote the potential outcome when $D = d$, where $d \in \{0, 1\}$. It is only possible to observe (Y, D), where $Y = DY_1 + (1 - D)Y_0$. This paper focuses on identification of τ-th quantiles of Y_1 and Y_0 for some $\tau \in (0, 1)$, and the distribution function of the TE $(Y_1 - Y_0)$. For given $y_1, y_0 \in \mathbb{R}$, define the following functions:

$$
\begin{aligned}
LB_1(y_1) &\equiv \Pr(Y \leq y_1 | D = 1) \Pr(D = 1), \\
UB_1(y_1) &\equiv \Pr(Y \leq y_1 | D = 1) \Pr(D = 1) + \Pr(D = 0), \\
LB_0(y_0) &\equiv \Pr(Y \leq y_0 | D = 0) \Pr(D = 0), \\
UB_0(y_0) &\equiv \Pr(Y \leq y_0 | D = 0) \Pr(D = 0) + \Pr(D = 1).
\end{aligned}
$$

Under this counterfactual model, it can be shown that the marginal distribution functions of $Y_1$and Y_0 are partially identified.

Lemma 3.3.1. *Let $F_1(\cdot)$ and $F_0(\cdot)$ be the distribution functions of Y_1 and Y_0, respectively. Then,*

$$
\begin{aligned}
F_1(y_1) &\in [LB_1(y_1), UB_1(y_1)], && (3.3.1) \\
F_0(y_0) &\in [LB_0(y_0), UB_0(y_0)]. && (3.3.2)
\end{aligned}
$$

Since the marginal distribution functions are only partially identified, it is natural that the quantiles of the potential outcomes are also partially identified. For a subset $A \subseteq \mathbb{R}$, denote the space of cadlag functions that map from A to $\mathbb{R}$ by $\mathbb{D}(A)$. For a non-decreasing function $G \in \mathbb{D}(\mathbb{R})$, define the left-continuous inverse $G^{\leftarrow}(r) \equiv \inf\{y : G(y) \geq r\}$. For a given $\tau \in (0, 1)$, define $Q_1(\tau) \equiv F_1^{\leftarrow}(\tau)$

and $Q_0(\tau) \equiv F_0^{\leftarrow}(\tau)$ (i.e. $Q_1(\tau)$ and $Q_0(\tau)$ are the τ-th quantile of Y_1 and Y_0, respectively). Then the τ-th QTE is defined as follows.

Definition 3.3.1. *Let $\tau \in (0,1)$ be given. The τ-th QTE is defined as*

$$QTE(\tau) \equiv Q_1(\tau) - Q_0(\tau).$$

As mentioned above, another object of interest in this paper is the distribution of the treatment effect. To formally define the distribution of the treatment effect, I provide the definition of the treatment effect as follows.

Definition 3.3.2. *The treatment effect Δ is the difference between Y_1 and Y_0. That is,*

$$\Delta \equiv Y_1 - Y_0.$$

Consider the equation (3.3.1) and suppose that one is interested in τ-th quantile of Y_1, $Q_1(\tau)$. To identify this quantity, I first focus on the lower bound $LB_1(\cdot)$. Since all the components of $LB_1(\cdot)$ are identified from the data, one can find the value

$$Q_1^U(\tau) \equiv LB_1^{\leftarrow}(\tau).$$

Similarly, one can find the value

$$Q_1^L(\tau) \equiv UB_1^{\leftarrow}(\tau).$$

In a similar fashion, define

$$\begin{aligned} Q_0^U(\tau) &\equiv LB_0^{\leftarrow}(\tau), \\ Q_0^L(\tau) &\equiv UB_0^{\leftarrow}(\tau). \end{aligned}$$

Then, one can have the following results which were introduced by Manski (1994).

Lemma 3.3.2. *Let $\tau \in (0,1)$ be fixed. If the marginal distributions of Y_1 and Y_0 are partially identified as in Lemma 3.3.1, then*

$$Q_1(\tau) \in [Q_1^L(\tau), Q_1^U(\tau)], \tag{3.3.3}$$

$$Q_0(\tau) \in [Q_0^L(\tau), Q_0^U(\tau)], \tag{3.3.4}$$

$$QTE(\tau) \in [Q_1^L(\tau) - Q_0^U(\tau), Q_1^U(\tau) - Q_0^L(\tau)]. \tag{3.3.5}$$

Lemma 3.3.2 shows how one can (partially) recover the quantile of the potential outcomes from (partially) identified marginal distribution functions. The results in Lemma 3.3.2 are closely related to the identification results in Stoye (2010). He considers identification of some classes of functionals of the distribution functions of the potential outcomes. In particular, one of these classes, which is called the class of D_1-parameters, includes the quantiles of the potential outcomes as a special case. In contrast to that Stoye (2010) provides identification results under a general potential outcome framework, I extend some part of his results to the cases where stochastic dominance assumptions are imposed.

3.3.1 Identification under Stochastic Dominance

As mentioned in the previous sections, prior information on a model helps to obtain much finer identification results. In this regard, I introduce stochastic dominance assumptions. Before starting with the identification analysis, I first give the definition of first-order stochastic dominance.

Suppose that there are two random variables X and Y which have marginal distribution functions $F_X(\cdot)$ and $F_Y(\cdot)$, respectively. X first-order stochastically dominates Y if for all $t \in \mathbb{R}$, $F_X(t) \leq F_Y(t)$. Note that if X first-order stochastically dominates Y, then one can show that $E[X] \geq E[Y]$, but not vice versa.

The following assumption states that the potential outcome conditional on $D = 1$ first-order stochastically dominates the potential outcome conditional on $D = 0$.

Assumption 3.3.3. *For all $j \in \{0, 1\}$, $Y_j|D = 1$ first-order stochastically dominates $Y_j|D = 0$.*

Assumption 3.3.3 means that for given $j \in \{0, 1\}$ and for all $y \in \mathbb{R}$, $F_j(y|D = 1) \leq F_j(y|D = 0)$, where $F_j(\cdot|D = 1)$ and $F_j(\cdot|D = 0)$ are the distribution functions of $Y_j|D = 1$ and $Y_j|D = 0$, respectively. Blundell et al. (2007) applied a version of this assumption as well as a median restriction to their study. Since Assumption 3.3.3 implies that $E[Y_j|D = 1] \geq E[Y_j|D = 0]$, this assumption is a sufficient condition for the MTS assumption. Note that, however, this stochastic dominance condition does not imply that $Y_j|D = 1 \geq Y_j|D = 0$ a.s. nor that $Y_j|D = 1 < Y_j|D = 0$ a.s.,

for all $j \in \{0,1\}$. Thus, this stochastic dominance assumption is more general than the MTR assumption but stronger than the MTS assumption[10].

Example 3.3.1. *Suppose that the outcome variable is wage and that the treatment is to earn a college degree. It is likely that the more capable people are, the more likely it is for them to complete college education. As a result, one may anticipate that people with college degrees are more likely to have higher learning ability than those who did not complete college education. Many studies in labor economics literature consider learning ability as an important factor affecting wage and thus one can suppose that people with college degrees have higher wages than those without. This can be formalized by first-order stochastic dominance as in Blundell et al. (2007).*

The following theorem gives the identification results under Assumption 3.3.3.

Theorem 3.3.4. *Suppose that Assumption 3.3.3 holds. For given $y \in \mathbb{R}$, define*

$$\begin{aligned} LB_1^{FSD1}(y) &\equiv \Pr(Y_1 \le y | D=1), \\ UB_1^{FSD1}(y) &\equiv \Pr(Y_1 \le y | D=1)\Pr(D=1) + \Pr(D=0), \\ LB_0^{FSD1}(y) &\equiv \Pr(Y_0 \le y | D=0)\Pr(D=0), \\ UB_0^{FSD1}(y) &\equiv \Pr(Y_0 \le y | D=0). \end{aligned}$$

[10] Jun et al. (2011) consider a triangular model with endogenous variables, which has the forms $y = g(x,u)$ and $x = h(z,v)$, where z is an instrumental variable and both g and h are non-decreasing in u and v, respectively. They assume that the quantile of u conditional on v increases in v (Assumption D). One can observe that Assumption 3.3.3 is similar to Assumption D in Jun et al. (2011) because for given $j \in \{0,1\}$, $Q_j(\tau|D=1) \ge Q_j(\tau|D=0)$ under Assumption 3.3.3, where $Q_j(\tau|D=k)$ is the τ-th quantile of Y_j conditional on $D=k$.

Then,

$$F_1(y) \in [LB_1^{FSD1}(y), UB_1^{FSD1}(y)], \tag{3.3.6}$$

$$F_0(y) \in [LB_0^{FSD1}(y), UB_0^{FSD1}(y)]. \tag{3.3.7}$$

Remark 3.3.5. *Comparing the bounds under stochastic dominance with the bounds given in 3.3.1, one can see that some bounds are identical to the ones in 3.3.1 (i.e.* $UB_1^{FSD1}(y) = UB_1(y)$ *and* $LB_0^{FSD1}(y) = LB_0(y)$ *for given* $y \in \mathbb{R}$*). Since Assumption 3.3.3 designates only one direction of the monotonicity of the distribution functions, it is impossible to improve the lower bound on* $F_0(y)$ *and the upper bound on* $F_1(y)$*. Nevertheless, the bounds on the marginal distribution functions provided in Theorem 3.3.4 are sharper than the ones in Lemma 3.3.1.*

The following stochastic dominance assumption may be regarded as an assumption corresponding to the MTR.

Assumption 3.3.6. *For all* $j \in \{0, 1\}$, $Y_1|D = j$ *first-order stochastically dominates* $Y_0|D = j$.

Note that Assumption 3.3.6 implies that $E[Y_1] \geq E[Y_0]$, but does not imply that $Y_1 \geq Y_0$ almost surely. That is, Assumption 3.3.6 only determines the order between two distribution functions.

Example 3.3.2. *Imposing Assumption 3.3.6 on Example 3.3.1 implies that people who have college degrees are likely to be paid higher wages than when they do not. In other words, the return to college education is likely to be positive. However, it is*

not necessarily true that a person with a high school degree is paid higher wage than when she had a college degree.

Under Assumption 3.3.6, one can obtain the following theorem.

Theorem 3.3.7. *Suppose that Assumption 3.3.6 holds. For given $y \in \mathbb{R}$, define*

$$\begin{aligned}
LB_1^{FSD2}(y) &= \Pr(Y_1 \le y | D=1)\Pr(D=1), \\
UB_1^{FSD2}(y) &= \Pr(Y_1 \le y | D=1)\Pr(D=1) + \Pr(Y_0 \le y | D=0)\Pr(D=0), \\
LB_0^{FSD2}(y) &= \Pr(Y_0 \le y | D=0)\Pr(D=0) + \Pr(Y_1 \le y | D=1)\Pr(D=1), \\
UB_0^{FSD2}(y) &= \Pr(Y_0 \le y | D=0)\Pr(D=0) + \Pr(D=1).
\end{aligned}$$

Then,

$$F_1(y) \in [LB_1^{FSD2}(y), UB_1^{FSD2}(y)], \tag{3.3.8}$$

$$F_0(y) \in [LB_0^{FSD2}(y), UB_0^{FSD2}(y)]. \tag{3.3.9}$$

As Assumption 3.3.3 is not enough to narrow $UB_1(y)$ and $LB_0(y)$, one can see that the lower bound and upper bound on $F_1(y)$ and $F_0(y)$ in Theorem 3.3.7 remain the same as those in Lemma 3.3.1. If both Assumptions 3.3.3 and 3.3.6 hold, it can be shown that one can tighten the bounds on the marginal distribution functions and the result is established in the following corollary.

Corollary 3.3.8. *Suppose that Assumptions 3.3.3 and 3.3.6 hold. For given $y \in \mathbb{R}$,*

$$\begin{aligned}
F_1(y) &\in [LB_1^{FSD1}(y), UB_1^{FSD2}(y)], \\
F_0(y) &\in [LB_0^{FSD2}(y), UB_0^{FSD1}(y)].
\end{aligned}$$

Remark 3.3.9. *The identified sets for marginal distribution functions of* Y_1 *and* Y_0 *in Corollary 3.3.8 are connected, and the intersection of these sets is the boundary of each set (i.e.* $UB_1^{FSD2}(y) = LB_0^{FSD2}(y)$*).*

3.3.2 The Distribution of the Treatment Effect

In this section, I present the identification result for the distribution of the TE $Y_1 - Y_0$. The main strategy is based on the identification strategy from Fan and Park (2010), which uses the notion of a copula with marginal distribution functions.

A copula is a joint distribution function function of two uniform random variables. Sklar's Theorem (Theorem 2.3.3 in Nelsen (1999)) shows that if there are two random variables X and Y with marginal distribution functions $F_X(x)$ and $F_Y(y)$, respectively, then the joint distribution function of X and Y, defined as $F_{XY}(x, y)$, is characterized by a copula[11]. Sklar's Theorem also shows that if C is a copula, and if $F_X(\cdot)$ and $F_Y(\cdot)$ are the marginal distribution functions for X and Y, respectively, then one can define a function $F_{XY}(x, y) = C(F_X(x), F_Y(y))$ as a joint distribution of two random variables X and Y whose the marginal distribution functions are $F_X(\cdot)$ and $F_Y(\cdot)$, respectively.

To derive the bound on the distribution function of the TE Δ, let $F_\Delta(\cdot)$ be the distribution function of Δ. Theorem 2 in Williamson and Downs (1990) can be

[11] It can be proven that there exists a copula C such that $F_{XY}(x, y) = C(F_X(x), F_Y(y))$.

used to show that if the marginal distribution functions are given by F_1 and F_0, then

$$\begin{aligned} \sup_{u+v=x}\{\max[F_1(u)-F_0(-v),0]\} &\leq F_\Delta(x), \\ \inf_{u+v=x}\{\min[F_1(u)-F_0(-v),0]\}+1 &\geq F_\Delta(x). \end{aligned}$$

These bounds on the distribution function $F_\Delta(\cdot)$ are based on the *Fréchet-Hoeffding lower and upper bounds* on $F_{Y_1,Y_0}(\cdot,\cdot)$, where $F_{Y_1,Y_0}(\cdot,\cdot)$ is the joint distribution function of Y_1 and Y_0. The next theorem establishes the bound on the distribution of the TE $F_\Delta(\cdot)$ when the marginal distributions are partially identified.

Theorem 3.3.10. *Suppose that, for all $y \in \mathbb{R}$, the identified sets of $F_1(y)$ and $F_0(y)$ are given by $[\tilde{LB}_1(y), \tilde{UB}_1(y)]$ and $[\tilde{LB}_0(y), \tilde{UB}_0(y)]$, respectively. For given $\delta \in \mathbb{R}$, define*

$$LB_\Delta(\delta) = \sup_y\{\max[\tilde{LB}_1(y)-\tilde{UB}_0(y-\delta),0]\}, \tag{3.3.10}$$

$$UB_\Delta(\delta) = \inf_y\{\min[\tilde{UB}_1(y)-\tilde{LB}_0(y-\delta),0]\}+1. \tag{3.3.11}$$

Then,

$$F_\Delta(\delta) \in [LB_\Delta(\delta), UB_\Delta(\delta)].$$

Remark 3.3.11. *Fan and Park (2010) consider randomized experiments so that the marginal distribution functions are directly point-identified from data. Since this paper does not rule out situations where the treatment is endogenous and the structure of the model is inadequate to fully identify the marginal distributions, the identified set of $F_\Delta(\delta)$ in Theorem 3.3.10 is broader than the one provided by Fan and Park*

(2010).

3.4 Estimation and Confidence Regions for Identified Sets

In this section, I provide consistent estimators of the bounds on marginal distribution functions and the distribution function of the TE, which are presented in Lemma 3.3.1 and Theorem 3.3.10, respectively. For a given identified set $\Theta_I(\theta_0)$ of a parameter θ_0, I also construct a confidence region for that identified set. Let $F_{jk}(y)$ and p^* denote $\Pr(Y_{ji} \leq y | D_i = k)$ and $\Pr(D = 1)$, respectively. Then one can estimate p^* by its sample analogue $\hat{p}_n \equiv \frac{1}{n}\sum_i^n D_i$. The asymptotic theory in this section mostly focuses on the identification regions without stochastic dominance assumptions. I impose some assumptions on the data generating process to establish the asymptotic theory.

Assumption 3.4.1. $\{W_i \equiv (Y_{1i}, Y_{0i}, D_i)' : i = 1, 2, ..., n\}$ *is a random sample.*

Assumption 3.4.2. *There exists a small* $\epsilon_0 > 0$ *such that* $p^* \in [\epsilon_0, 1 - \epsilon_0]$.

Assumption 3.4.1 means that the observed data $\{(Y_i, D_i)' : i = 1, 2, ..., n\}$ are i.i.d. Assumption 3.4.2 implies that there exists a number $\lambda_0 \in (0, \infty)$ such that $\frac{\frac{1}{n}\sum_i D_i}{1 - \frac{1}{n}\sum_i D_i} \to \lambda_0$ as $n \to \infty$[12]. Under Assumptions 3.4.1 and 3.4.2, one can consistently estimate the bounds in 3.3.1 with the following objects:

[12]Instead of Assumption 3.4.2, Fan and Park (2010) consider this condition.

$$\hat{LB}_{1n}(y) \equiv \frac{1}{n}\sum_{i}^{n} D_i \mathbf{1}(Y_i \leq y), \tag{3.4.1}$$

$$\hat{UB}_{1n}(y) \equiv \frac{1}{n}\sum_{i}^{n} \{D_i \mathbf{1}(Y_i \leq y) + 1 - D_i\}, \tag{3.4.2}$$

$$\hat{LB}_{0n}(y) \equiv \frac{1}{n}\sum_{i}^{n} (1 - D_i)\mathbf{1}(Y_i \leq y), \tag{3.4.3}$$

$$\hat{UB}_{0n}(y) \equiv \frac{1}{n}\sum_{i}^{n} \{(1 - D_i)\mathbf{1}(Y_i \leq y) + D_i\}. \tag{3.4.4}$$

Consequently, the identification regions of $Q_1(\tau)$ and $Q_0(\tau)$ can be estimated by taking left-continuous inverse of the quantities in equations (3.4.1)-(3.4.4). Since all of the summands are binary variables which have finite second moments and $\hat{p}_n \xrightarrow{p} p^*$, one can show that these are consistent estimators of the true parameters by applying the law of large numbers. Furthermore, these estimators are $\sqrt{n}$-asymptotically normal for given $y \in \mathbb{R}$ and will be used to construct confidence regions for the identified sets of the marginal distributions and the quantiles of the potential outcomes. The confidence regions considered in this paper are *confidence regions for identified sets that are pointwise consistent in level,* and the term is used by Romano and Shaikh (2010). To define the confidence regions, let $\Theta_I(\theta_0)$ be an identification region of a parameter θ_0 and $\alpha \in (0, 1)$ be given. A confidence region for $\Theta_I(\theta_0)$ that is pointwise consistent in level α, denoted by $\mathcal{C}_n(\alpha; \theta_0)$, is a random set such that

$$\liminf_{n\to\infty} \Pr(\Theta_I(\theta_0) \subseteq \mathcal{C}_n(\alpha; \theta_0)) \geq \alpha.$$

It is worth noting that such confidence regions are conservative in a sense

that their coverage probability is greater than or equal to a given level α even asymptotically. I first construct confidence regions for the marginal distribution functions $F_1(y)$ and $F_0(y)$ and then consider those for the quantiles of the potential outcomes. I only provide results for the general case (i.e. the identification regions of $F_1(y)$ and $F_0(y)$ are given in Lemma 3.3.1), but the results can be modified to construct confidence regions of identification regions under stochastic dominance assumptions in this paper.

Before proceeding, I introduce notation that will be used to establish confidence regions. Suppose that there is a consistent estimator of θ_0, $\hat{\theta}_n$. I denote the variance of $\hat{\theta}_n$ by $\sigma^2(\hat{\theta}_n)$. Let $\Phi(\cdot)$ and $\phi(\cdot)$ be the distribution and density functions of the standard normal random variable, respectively. I denote τ-th quantile of the standard normal random variable by z_τ(i.e. $\Phi(z_\tau) = \tau$).

It is straightforward to see that, under Assumption 3.4.1,

$$\begin{aligned}
&\sigma_F^2(\sqrt{n}\hat{LB}_{1n}(y)) = p^* \cdot F_{11}(y) \cdot (1 - p^* F_{11}(y)), \\
&\sigma_F^2(\sqrt{n}\hat{UB}_{1n}(y)) = p^* \cdot (1 - F_{11}(y)) \cdot \{1 - p^* \cdot (1 - F_{11}(y))\}, \\
&\sigma_F^2(\sqrt{n}\hat{LB}_{0n}(y)) = (1 - p^*) F_{00}(y) \cdot (1 - (1 - p^*) \cdot F_{00}(y)), \\
&\sigma_F^2(\sqrt{n}\hat{UB}_{0n}(y)) = (1 - p^*)(1 - F_{00}(y)) \{1 - (1 - p^*) \cdot (1 - F_{00}(y))\}.
\end{aligned}$$

The following theorem provides confidence regions for $\Theta_I(F_1(y))$ and $\Theta_I(F_0(y))$ in the general case.

Theorem 3.4.3. *Let $y \in \mathbb{R}$ and $\alpha \in (0,1)$ be given. Suppose that the identification regions of $F_1(y)$ and $F_0(y)$ are given by $\Theta_I(F_1(y)) = [LB_1(y), UB_1(y)]$ and*

$\Theta_I(F_0(y)) = [LB_0(y), UB_0(y)]$, *respectively. Define*

$$\mathcal{C}_n(\alpha; F_1(y)) \equiv [\hat{LB}_{1n}(y) - C^L_{F1n}(\alpha; y), \hat{UB}_{1n}(y) + C^U_{F1n}(\alpha; y)], \tag{3.4.5}$$

$$\mathcal{C}_n(\alpha; F_0(y)) \equiv [\hat{LB}_{0n}(y) - C^L_{F0n}(\alpha; y), \hat{UB}_{0n}(y) + C^U_{F0n}(\alpha; y)], \tag{3.4.6}$$

where

$$C^L_{F1n}(\alpha; y) \equiv z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{LB}_{1n}(y))}{\sqrt{n}},$$
$$C^U_{F1n}(\alpha; y) \equiv z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{UB}_{1n}(y))}{\sqrt{n}},$$
$$C^L_{F0n}(\alpha; y) \equiv z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{LB}_{0n}(y))}{\sqrt{n}},$$
$$C^U_{F0n}(\alpha; y) \equiv z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{UB}_{0n}(y))}{\sqrt{n}}.$$

If Assumptions 3.4.1 and 3.4.2 are satisfied, then

$$\liminf_{n\to\infty} \Pr(\Theta_I(F_1(y)) \subseteq \mathcal{C}_n(\alpha; F_1(y))) \geq \alpha$$

and

$$\liminf_{n\to\infty} \Pr(\Theta_I(F_0(y)) \subseteq \mathcal{C}_n(\alpha; F_0(y))) \geq \alpha.$$

Now I consider confidence regions for the identified sets of quantiles of potential outcomes. Recall that the quantiles of the potential outcomes can be identified by considering the left-continuous inverse of the lower and upper bounds on their marginal distribution functions. I use the functional-delta method (Theorem 3.9.4 in van der Vaart and Wellner (1996)) to construct confidence regions for $\Theta_I(Q_1(\tau))$ and $\Theta_I(Q_0(\tau))$. I impose additional assumptions on the distribution functions F_{11}

and F_{00} to construct confidence regions for these quantiles.

Assumption 3.4.4. *(i) The conditional distributions functions $F_{11}(y)$ and $F_{00}(y)$ admit their density functions, denoted by $f_{11}(y)$ and $f_{00}(y)$, respectively; (ii) The density functions $f_{11}(y)$ and $f_{00}(y)$ are bounded and continuously differentiable, and their first-order derivatives $f_{11}^{'}$ and $f_{00}^{'}$ are uniformly bounded; (iii) There exists a small $\eta_0 > 0$ such that, for given $\tau \in [\eta_0, 1-\eta_0]$, $f_{11}(Q_1^U(\tau))$, $f_{11}(Q_1^L(\tau))$, $f_{00}(Q_0^U(\tau))$, and $f_{00}(Q_0^L(\tau))$ are bounded away from zero.*

Assumption 3.4.4 imposes smoothness of the conditional distribution functions $F_{11}(y)$ and $F_{00}(y)$. This assumption allows us to use the functional-delta method to establish the asymptotic normality of the bounds on quantiles of the potential outcomes. Let $\tau \in [\eta_0, 1-\eta_0]$ be given. The next theorem provides confidence regions for $\Theta_I(Q_1(\tau))$ and $\Theta_I(Q_0(\tau))$ that are pointwise consistent in level α.

Theorem 3.4.5. *Suppose that Assumptions 3.4.1-3.4.4 are satisfied and let $\tau \in [\eta_0, 1-\eta_0]$ be given. For given $\alpha \in (0,1)$, define*

$$\mathcal{C}_n(\alpha; Q_1(\tau)) \equiv [Q_{1n}^L(\tau) - C_{q1n}^L(\alpha;\tau), Q_{1n}^U(\tau) + C_{q1n}^U(\alpha;\tau)], \tag{3.4.7}$$

$$\mathcal{C}_n(\alpha; Q_0(\tau)) \equiv [Q_{0n}^L(\tau) - C_{q0n}^L(\alpha;\tau), Q_{0n}^U(\tau) + C_{q0n}^U(\alpha;\tau)], \tag{3.4.8}$$

where

$$C^L_{q1n}(\alpha;\tau) \equiv z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{UB}_{1n}(Q^L_1(\tau)))}{\sqrt{n}p^* f_{11}(Q^L_1(\tau))},$$
$$C^U_{q1n}(\alpha;\tau) \equiv z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{LB}_{1n}(Q^U_1(\tau)))}{\sqrt{n}p^* f_{11}(Q^U_1(\tau))},$$
$$C^L_{q0n}(\alpha;\tau) \equiv z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{UB}_{0n}(Q^L_0(\tau)))}{\sqrt{n}(1-p^*) f_{00}(Q^L_0(\tau))},$$
$$C^U_{q0n}(\alpha;\tau) \equiv z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{LB}_{0n}(Q^U_0(\tau)))}{\sqrt{n}(1-p^*) f_{00}(Q^U_0(\tau))}.$$

Then

$$\liminf_{n\to\infty} \Pr(\Theta_I(Q_1(\tau)) \subseteq \mathcal{C}_n(\alpha; Q_1(\tau))) \geq \alpha$$

and

$$\liminf_{n\to\infty} \Pr(\Theta_I(Q_0(\tau)) \subseteq \mathcal{C}_n(\alpha; Q_0(\tau))) \geq \alpha.$$

Now I consider constructing confidence regions for the identification region of $F_\Delta(\delta)$ for given $\delta \in \mathbb{R}$. For each $\delta \in \mathbb{R}$, define the following objects:

$$y^{sup}(\delta) = \arg\sup_y \{LB_1(y) - UB_0(y-\delta)\},$$
$$\hat{y}^{sup}_n(\delta) = \arg\sup_y \{\hat{LB}_{1n}(y) - \hat{UB}_{0n}(y-\delta)\},$$
$$y^{inf}(\delta) = \arg\inf_y \{UB_1(y) - LB_0(y-\delta)\},$$
$$\hat{y}^{inf}_n(\delta) = \arg\inf_y \{\hat{UB}_{1n}(y) - \hat{LB}_{0n}(y-\delta)\}.$$

Then $\hat{y}^{sup}_n(\delta)$ and $\hat{y}^{inf}_n(\delta)$ are natural estimators of $y^{sup}(\delta)$ and $y^{inf}(\delta)$, respectively. I impose the following assumptions to construct a confidence region for $\Theta_I(F_\Delta(\delta))$.

Assumption 3.4.6. *For all $\delta \in \mathbb{R}$, $y^{sup}(\delta)$ and $y^{inf}(\delta)$ are unique and interior*

points.

Assumption 3.4.7. *The supports of Y_1 and Y_0 are compact subsets in $\mathbb{R}$.*

Assumption 3.4.6 guarantees consistency of $\hat{y}_n^{sup}(\delta)$ and $\hat{y}_n^{inf}(\delta)$. Assumption 3.4.7 implies that the support of the observed outcome variable Y is also a compact subset in $\mathbb{R}$. To characterize a confidence region for $\Theta_I(F_\Delta(\delta))$, define

$$m_i^L(y;\delta) \equiv D_i \mathbf{1}(Y_i \leq y) - \{(1-D_i)\mathbf{1}(Y_i \leq y-\delta) + D_i\},$$
$$m_i^U(y;\delta) \equiv \{D_i \mathbf{1}(Y_i \leq y) + (1-D_i)\} - (1-D_i)\mathbf{1}(Y_i \leq y-\delta).$$

The next theorem provides a confidence region for $\Theta_I(F_\Delta(\delta))$ that is pointwise consistent in level.

Theorem 3.4.8. *Let $\alpha \in (0,1)$ and $\delta \in \mathbb{R}$ be given. Suppose that the marginal distributions of Y_1 and Y_0 are identified as Lemma 3.3.1 and that Assumptions 3.4.1, 3.4.2, 3.4.4, 3.4.6, and 3.4.7 are satisfied. Let*

$$\hat{LB}_{\Delta n}(\delta) \equiv \sup_y \{\max[\hat{LB}_{1n}(y) - \hat{UB}_{0n}(y-\delta), 0]\},$$
$$\hat{UB}_{\Delta n}(\delta) \equiv \inf_y \{\min[\hat{UB}_{1n}(y) - \hat{LB}_{0n}(y-\delta), 0]\} + 1.$$

Define

$$\mathcal{C}_n(\alpha; F_\Delta(\delta)) \equiv [\hat{LB}_{\Delta n}(\delta) - c_{\frac{\alpha+1}{2}}(\delta), \hat{UB}_{\Delta n}(\delta) + \tilde{c}_{\frac{\alpha+1}{2}}(\delta)],$$

where for given $\tau \in (0,1)$, $c_\tau(\delta)$ and $\tilde{c}_\tau(\delta)$ are τ-th quantiles of the random variables $C(\delta) \equiv \max[N(0, Var(m_i^L(y^{sup}(\delta);\delta)), 0]$ and $\tilde{C}(\delta) \equiv \min[N(0, Var(m_i^U(y^{inf}(\delta);\delta)), 0] + 1$, respectively.

If

$$p^* f_{11}^{'}(y^{sup}(\delta)) - (1-p^*) f_{00}^{'}(y^{sup}(\delta) - \delta) < 0$$

and

$$(1-p^*) f_{00}^{'}(y^{inf}(\delta)) - p^* f_{11}^{'}(y^{inf}(\delta) - \delta) > 0$$

hold, then

$$\lim_{n\to\infty} \inf \Pr(\Theta_I(F_\Delta(\delta)) \subseteq \mathcal{C}_n(\alpha; F_\Delta(\delta))) \geq \alpha,$$

where $\Theta_I(F_\Delta(\delta)) = [\sup_y\{\max[LB_1(y) - UB_0(y-\delta), 0]\}, \inf_y\{\min[UB_1(y) - LB_0(y-\delta), 0]\} + 1]$.

The confidence regions provided in this section are not feasible as they contain unknown quantities. However, it is relatively straightforward to construct feasible confidence regions for the identified sets of the marginal distributions and quantile functions of the potential outcomes. One can replace the unknown quantities with their sample analogues or nonparametric estimators to construct feasible confidence regions for these identified sets. In contrast, the confidence region for the identified set of $F_\Delta(\delta)$ involves critical values that are from some non-standard distributions. One may think of resampling methods to simulate these distributions to obtain critical values, but the validity of such a resampling method needs to be proven. Fan and Park (2010) provide a bootstrap scheme to obtain the critical values $c_\tau(\delta)$ and $\tilde{c}_\tau(\delta)$. Related to resampling methods for a general class of partially identified models, Bugni (2010) introduces a bootstrap procedure that can be used for inference for some class of partially identified models. I leave this issue for future work.

3.5 Application to the Return to College

Labor economists have often tried to examine the return to schooling. Education level of an individual is a choice variable[13], and this fact results in the endogeneity of educational attainment. Specifically, completion of a college education entirely depends on individual's decision making process. In this regard, I analyze an empirical problem of measuring the return to college by defining the treatment as earning a bachelor's degree.

I take the ability of an individual as a source of endogeneity of the education level. Hendricks and Leukhina (2014) recognize that the rate of completing college education is quite low in spite of a big difference in earnings between college graduates and high school graduates, and they infer this gap comes from the difference in the ability. Since in general it is believed that people's ability is positively correlated with wage and education level, which coincides with what Example 3.3.1 illustrates, one may apply Assumption 3.3.3 to this empirical question.

Both theoretical and empirical studies on the return to schooling have suggested that more-educated people are in better labor status in terms of wage than less-educated people[14], and such observations can be rationalized by viewing education as human capital. This implies that for any given education level, the potential wage that would have been paid for college graduates is likely to be higher than the potential wage that would have been paid for non-college graduates. Therefore, it

[13] The presence of a compulsory school attendance law may make it difficult to classify the educational level as a choice variable. Nevertheless, when it comes to post-high school education, it is harmless to define the education level as a choice variable.

[14] See Card (1999) for more details.

seems perfectly plausible to impose Assumption 3.3.6 on the model.

3.5.1 Data

I extract variables from Integrated Public Use Microdata Series (IPUMS)[15] to estimate bounds. Considering the financial crisis during 2007-2008, which may have caused a drastic change in economic conditions in the U.S., I use the data from 2005[16]. I restrict the sample to white males in the age group of 23 to 40 years old. Moreover, the sample in this study only contains heads of households who are U.S. citizens, and I drop individuals who are not in the labor force or are self-employed. As a result, I am able to obtain a sample of 93,742 observations.

Table 3.1 summarizes the descriptive statistics of some variables. AGE is the variable indicating the age of each individual, and $EDUC$ is the educational attainment. $EDUCD$ contains more specific information on the educational attainment. In particular, it distinguishes people who have bachelor's degree or higher from others while $EDUC$ merely shows how many years of education. The variable, $EDUCD$, is a dummy variable indicating whether an individual is treated or not. $INCWAGE$ is the annual income from wage, measured in dollars, and $WEEKWAGE$ is the weekly income from wage. IPUMS does not provide data on weekly income and thus I obtain $WEEKWAGE$ by dividing $INCWAGE$ by the number of weeks worked. The dependent variable is the log transformation of the weekly earnings,

[15] Ruggles et al. (2010).

[16] Another reason I use the dataset is that the information on the number of weeks worked ($WKSWORK1$) is available only up to 2007. Taking positive correlation between the education level and working hours into consideration (see, for example, Card (1999)), this variable is required to generate the weekly earning.

denoted by *log(WEEKWAGE)*. The probability of earning a college degree is about 39% and the mean of weekly earning is approximately 1,681 dollars.

The average log of weekly earnings of the treated and the untreated are about 7.022 and 6.567, respectively. A t-test confirms that the difference in the mean between two groups is statistically significant at the 1% level.

3.5.2 Estimation Results

I first estimate the bounds on the QTE of the college degree on the log of (weekly) wage for given $\tau \in (0,1)$, and the results are presented in Table 3.2. I implicitly assume that the supports of Y_1 and Y_0 are the same, and use the realized values of these variables to calculate the empirical distribution functions.

The first panel shows the lower and upper bounds on the QTE for given $\tau \in (0,1)$ without any distributional assumptions. The second panel reveals the estimation results of the bounds under Assumption 3.3.3, and the last two columns provide the results under Assumption 3.3.6. I do not report the estimated bounds when Assumptions 3.3.3 and 3.3.6 are imposed together, but one can find these bounds from Figure 3.3[17]. If any distributional restrictions are not imposed, the bounds on the QTE are barely informative. In particular, the bounds yield a broader interval for the QTE when τ is small.

On the other hand, it is shown that imposing the restrictions improves the bounds so that the identified sets become more informative. The upper bound on

[17] As mentioned earlier, the lower bound coincides with the lower bound under Assumption 3.3.6 and the upper bound is identical to the one under Assumption 3.3.3.

the QTE under Assumption 3.3.3 decreases and thus one obtains a narrower bound on the QTE than the one without prior information. Panel 3 shows that the lower bounds on the QTE can be tightened if Assumption 3.3.6 is imposed. The lower bounds for all τ are identical to 0 and this is because Assumption 3.3.6 implies that Y_1 first-order stochastically dominates Y_0, which also implies that the τ-th quantile of Y_1 is always greater than or equal to the one of Y_0. Note that the lower bound under Assumption 3.3.3 and the upper bound under Assumption 3.3.6 do not have contributions in terms of tightening the bounds in the general case. However, Figure 3.3 shows that combining these two assumptions gives a much narrower interval for the QTE.

Figures 3.4 through 3.6 illustrate the estimation results of the bounds on distribution of the TE under the assumptions. Without the assumptions, the upper bound and the lower bound on the distribution function of the TE are constant functions which have the values of 1 and 0, respectively. Figure 3.4 compares these bounds to the ones derived under Assumption 3.3.3. The upper bound under Assumption 3.3.3 is identical to the upper bound in the general case, but the lower bound under Assumption 3.3.3 is more informative than the one in the general case. Similarly, one can see that Assumption 3.3.6 improves the upper bound of the distribution of the TE and this is verified by Figure 3.5. As the case of the QTE, one can obtain a much greater identifying power when combining two assumptions as Figure 3.6 illustrates.

From the bounds on the distribution of the TE, the quantiles of the TE

are also partially identified and Table 3.3 shows the results[18]. Since the bounds on the distribution for the general case do not give any information, it is impossible to obtain any instructive bounds on the quantile of the TE. In contrast to the general case, Panel 2 and Panel 3 in Table 3.3 demonstrate how the stochastic dominance assumptions help these bounds be tightened. As the previous results for other bounds, combining the two assumptions yields much narrower bounds on the quantiles of the TE[19] and thus the identified sets become very informative.

Table 3.1: Descriptive Statistics

	Mean	S.D	Min	Max
AGE	32.836	4.968	23	40
EDUC	7.924	2.059	0	11
WKSWORK	49.619	7.123	1	52
INCWAGE	52177.04	44952.22	4	629000
WEEKWAGE	1092.747	1681.379	0.077	209666.7
log (INCWAGE)	10.623	0.726	1.386	13.352
log(WEEKWAGE)	6.743	0.670	-2.565	12.253
Treatment	0.387	0.487	0	1

[18] I restrict the support of the TE to $[-5, 5]$.

[19] The lower bound and the upper bound are equal to the lower bound under Assumption 3.3.6 and the upper bound under Assumption 3.3.3, respectively.

Table 3.2: Estimation Results of the Bounds on the QTEs

τ	Panel 1: General		Panel 2: Assumption 3.3.3		Panel 3: Assumption 3.3.6	
	Lower	Upper	Lower	Upper	Lower	Upper
0.1	-8.657	9.210	-8.657	0.388	0.000	9.210
0.2	-8.923	9.616	-8.923	0.376	0.000	9.616
0.3	-9.159	10.012	-9.159	0.368	0.000	10.012
0.4	-9.339	7.025	-9.339	0.386	0.000	7.025
0.5	-9.616	6.119	-9.616	0.441	0.000	6.119
0.6	-10.222	5.856	-10.222	0.452	0.000	5.856
0.7	-5.644	5.633	-5.644	0.445	0.000	5.633
0.8	-5.254	5.426	-5.254	0.490	0.000	5.426
0.9	-4.878	5.202	-4.878	0.539	0.000	5.202

Figure 3.1: Bounds on the QTE under Assumption 3.3.3

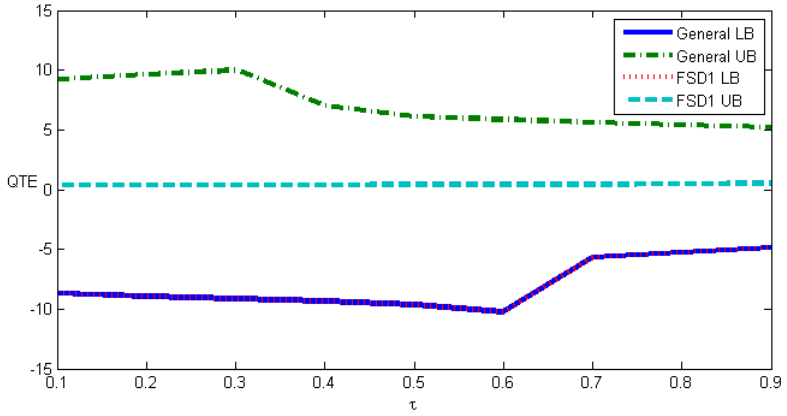


Figure 3.2: Bounds on the QTE under Assumption 3.3.6

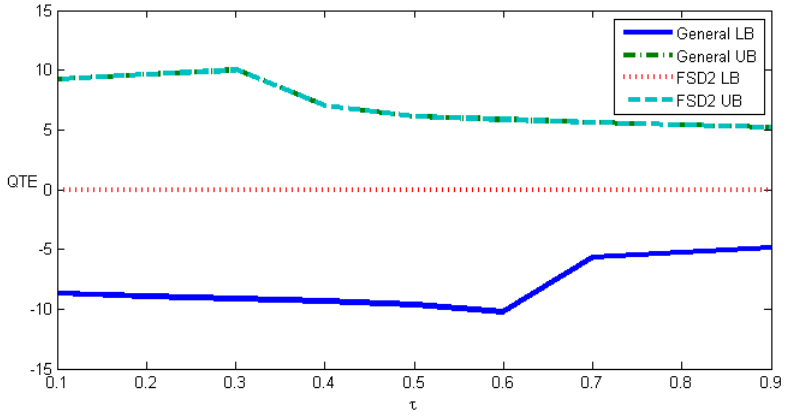


Figure 3.3: Bounds on the QTE under Assumptions 3.3.3 and 3.3.6

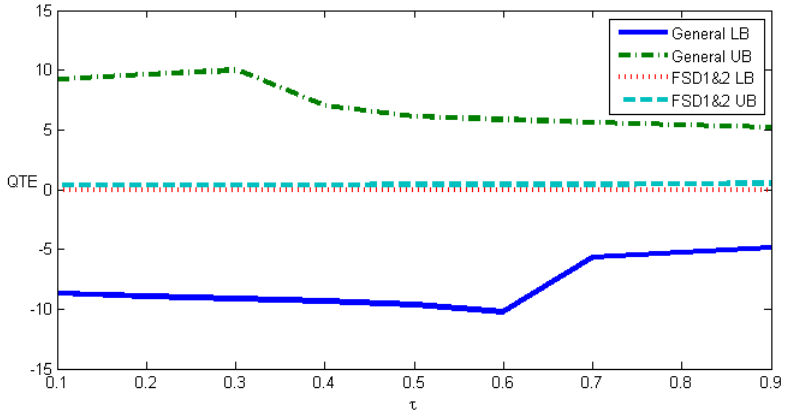


Figure 3.4: Bounds on the Distribution of the TE under Assumption 3.3.3

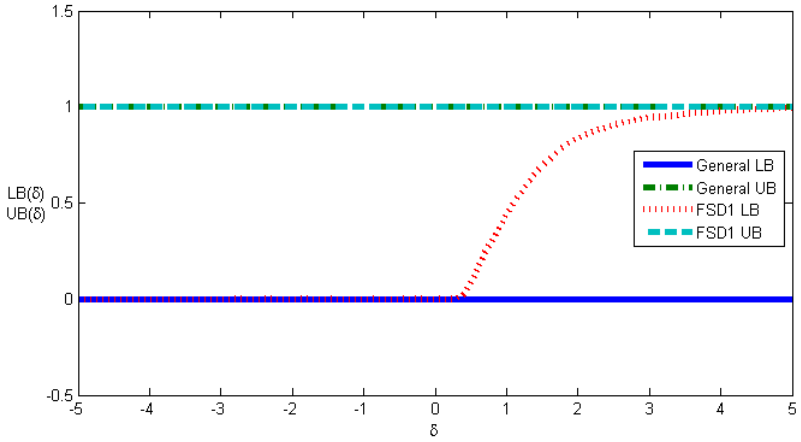


Figure 3.5: Bounds on the Distribution of the TE under Assumption 3.3.6

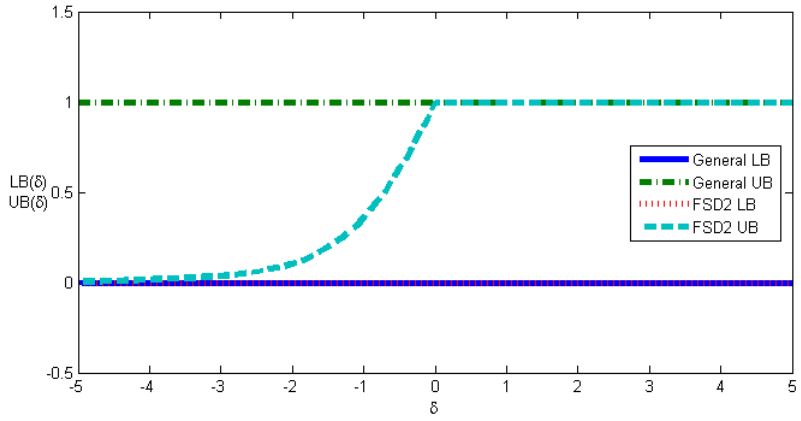


Figure 3.6: Bounds on the Distribution of the TE under Assumptions 3.3.3 and 3.3.6

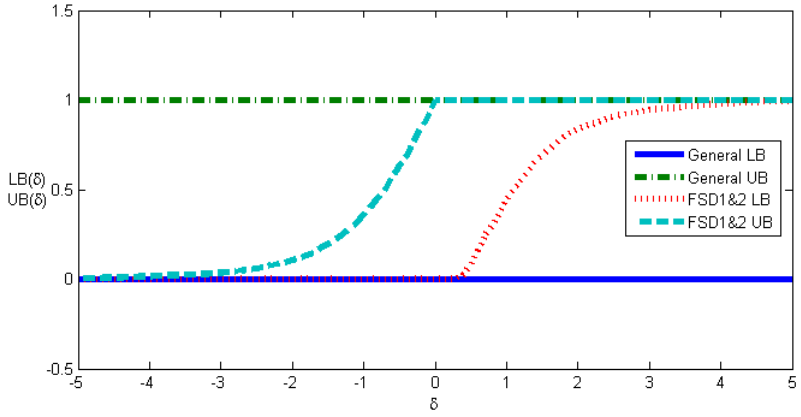


Table 3.3: Bounds on Quantiles of the TE

τ	Panel 1: General		Panel 2: Assumption 3.3.3		Panel 3: Assumption 3.3.6	
	Lower	Upper	Lower	Upper	Lower	Upper
0.1				0.6	-2	
0.2				0.7	-1.4	
0.3				0.8	-1.1	
0.4				1	-0.9	
0.5	$-\infty$	∞	$-\infty$	1.1	-0.6	∞
0.6				1.3	-0.5	
0.7				1.6	-0.3	
0.8				1.9	-0.2	
0.9				2.5	-0.1	

3.6 Conclusions

In this paper, I partially identify the QTEs and the distribution of the TE when the treatment is endogenous. To tighten the bounds, I consider several versions of stochastic dominance which seem reasonable in many situations. I adopt the approach of Fan and Park (2010) to identify the distribution of the TE $Y_1 - Y_0$.

It is shown that without additional assumptions, these bounds on the distribution function of the TE are broader than the ones given by Fan and Park (2010), and that stochastic dominance assumptions help to tighten the bounds. I apply the stochastic dominance assumptions for examining the return to college example, and the empirical evidence confirms that the distributional assumptions increase the identifying power.

There are several extensions one could consider. First, one can consider the structural approach instead of the treatment effects approach. Admittedly, the treatment effects approach has advantages over the structural approach in terms of robustness and/or credibility of the results. However, the approach is not capable of answering some important questions related to program evaluation. Heckman and Vytlacil (2007) describe three classes of questions in an economic policy evaluation, which entail evaluating and forecasting the impacts of a policy. The treatment effects approach is sufficient for (P1) evaluating the effects of a policy, but inadequate for (P2) forecasting the impact in a different environment or (P3) predicting the anticipated effects of a policy never performed in some environments[20]. The structural approach, however, can handle all three classes, thus enabling us to answer much broader classes of questions. In this sense, it is well worth considering other distributional assumptions and/or economic models such as a triangular system.

Second, it is worth considering different types of confidence intervals (or regions) for different objects. For example, one may be interested in inference for quantile processes, QTE process, or the distribution of the TE over the support of

[20]For details, see Heckman and Vytlacil (2007).

the TE. This requires the development of uniform asymptotic theory. Considering such an issue, one may develop an asymptotic theory for uniform inference for partially identified models. In addition, the confidence regions given in this paper are for the identified sets, not for the parameters of interest. Imbens and Manski (2004) and Stoye (2009) investigate how to construct asymptotically valid confidence intervals for partially identified parameters instead for the identified set and illustrate their approaches with the example on means with missing data. Since one may be interested in inference for parameters of interest themselves rather than for identified sets, it would be fruitful to provide asymptotically valid confidence intervals for the parameters considered in this paper.

Third, one can consider identification and estimation of the joint distribution of the potential outcomes under stochastic dominance assumptions. Related to this issue, Fan et al. (2014) provide identification and confidence sets for functionals of the joint distribution of the potential outcomes. The joint distribution can incorporate many other parameters that are important and relevant to the program evaluation, and thus it would be worth investigating this issue.

Lastly, this paper does not incorporate covariates. In many empirical situations, however, covariates are important to control for some heterogeneity. In the presence of covariates, one can adapt methods used in the literature on (conditional) moment inequality models (see, for example, Chernozhukov et al. (2007); Andrews and Soares (2010); Andrews and Shi (2013)) or the approach developed by Chernozhukov et al. (2013) to perform inference for parameters of interest. I leave these potential extensions for future work.

Appendices

Appendix A

Chapter 1 Appendix

A.1 Proof of Lemma 1.2.1

The proof of Lemma 1.2.1 is a slight modification of the proof of Theorem 2.14 of (Joe, 1997, p. 44). Suppose $C_{2|1} \prec_S \tilde{C}_{2|1}$. Let $(U_1, U_2) \sim C$, $(\tilde{U}_1, \tilde{U}_2) \sim \tilde{C}$, with $U_j \stackrel{d}{=} \tilde{U}_j$, $j = 1, 2$. By Theorem 2.9 of (Joe, 1997, p. 40), $(U_1, U_2) \stackrel{d}{=} (\tilde{U}_1, \psi(U_1, U_2))$ with $\psi(u_1, u_2) = \tilde{C}_{2|1}^{-1}(C_{2|1}(u_2|u_1)|u_1)$. Since $C_{2|1} \prec_S \tilde{C}_{2|1}$, ψ is increasing in u_1 and u_2. We consider two cases:

- Case 1: Suppose that u_1 and u_2 are such that $\psi(u_1, u_2) \leq u_2$. Then

$$\begin{aligned}
\tilde{C}(u_1, u_2) &= \Pr[\tilde{U}_1 \leq u_1, \tilde{U}_2 \leq u_2)] \\
&= \Pr[\tilde{U}_1 < u_1, \tilde{U}_2 < u_2)] \\
&= \Pr[U_1 < u_1, \psi(U_1, U_2) < u_2)] \\
&\geq \Pr[U_1 < u_1, \psi(u_1, U_2) < u_2] \\
&> \Pr[U_1 < u_1, U_2 < u_2)] = C(u_1, u_2)
\end{aligned}$$

  where the strict inequality holds since $U_2 < u_2$ implies $\psi(u_1, U_2) \leq \psi(u_1, u_2) \leq u_2$ (but not vice versa since $\psi(u_1, U_2) \leq u_2$ and $\psi(u_1, u_2) \leq u_2$ does not necessarily imply $U_2 < u_2$ and $\Pr[\psi(u_1, u_2) < \psi(u_1, U_2)] = \Pr[u_2 < U_2] \neq$

0), and the second last inequality holds since, given $U_1 < u_1$, $\psi(U_1, U_2) \leq \psi(u_1, U_2) < u_2$.

- Case 2: Suppose that u_1 and u_2 are such that $\psi(u_1, u_2) > u_2$. Then

$$\begin{aligned} u_2 - C(u_1, u_2) &= \Pr[U_1 > u_1, U_2 < u_2)] \\ &> \Pr[U_1 > u_1, \psi(u_1, U_2) \leq u_2)] \\ &\geq \Pr[U_1 > u_1, \psi(U_1, U_2) \leq u_2)] \\ &= \Pr[\tilde{U}_1 > u_1, \tilde{U}_2 < u_2] = u_2 - \tilde{C}(u_1, u_2) \end{aligned}$$

where the strict inequality holds since $U_2 > u_2$ implies $\psi(u_1, U_2) \geq \psi(u_1, u_2) > u_2$ or $\psi(u_1, U_2) \leq u_2$ implies $U_2 \leq u_2$ (but not vice versa).

Therefore in both cases, $C(u_1, u_2) < \tilde{C}(u_1, u_2)$ for any u_1 and u_2.

A.2 Proof of Theorem 1.2.11

Continued from the main text, we prove that there exist $(t_0, t_1, \delta_1, \rho)$ and $(t_0^*, t_1^*, \delta_1^*, \rho^*)$ such that the equation (1.2.14) holds. To show this, we choose further specifications. We assume a normal copula.[1] We choose $\rho = 0$, $\rho^* = 1$, $q_0 = t_0 = 1/3$, and $q_1 = t_1 = 2/3$. Since (U_1, U_2) are jointly uniform, note that when $\rho = 0$, the probability of the quadrant in $[0, 1]^2$ specified by each of (1.2.6), (1.2.8), (1.2.10), and (1.2.12) equals the volume of the quadrant. When $\rho^* = 1$, all the probability mass lies on the 45 degree line in $[0, 1]^2$ and no where else, so the probability of a quadrant

[1] This choice is not critical except that we can have ρ reach to 1.

specified by each of (1.2.7), (1.2.9), (1.2.11), and (1.2.13) equals the length of the 45 line which intersects with that quadrant. Suppose that the following observational equivalence holds:

$$\Pr[u_1 \leq t_0, u_2 \geq q_0; \rho] = \Pr[u_1 \leq t_0^*, u_2 \geq q_0; \rho^*] = 2/9,$$
$$\Pr[u_1 \leq t_0, u_2 \leq q_0; \rho] = \Pr[u_1 \leq t_0^\dagger, u_2 \leq q_0; \rho^*] = 1/9,$$
$$\Pr[u_1 \leq t_1, u_2 \geq q_1; \rho] = \Pr[u_1 \leq t_1^*, u_2 \geq q_1; \rho^*] = 2/9,$$
$$\Pr[u_1 \leq t_1, u_2 \leq q_1; \rho] = \Pr[u_1 \leq t_1^\dagger, u_2 \leq q_1; \rho^*] = 4/9.$$

One can easily show that these equations yield that $t_0^* = 5/9$, $t_0^\dagger = 1/9$, $t_1^* = 8/9$, and $t_1^\dagger = 4/9$. Consider the equation (1.2.14), which can be rewritten as $t_1^\dagger = t_1^* + t_0^\dagger - t_0^*$ or $t_1^\dagger - t_1^* = t_0^\dagger - t_0^*$. Then, note that we have $t_1^\dagger - t_1^* = t_0^\dagger - t_0^* = -4/9$, which is, in fact, the value of δ_1^*. In sum, the values of parameters that give the observationally equivalent fitted probabilities are

$$(t_0, t_1, q_0, q_1, \delta_1, \rho) = \left(\frac{1}{3}, \frac{2}{3}, \frac{1}{3}, \frac{2}{3}, 0, 0\right), \tag{A.2.1}$$
$$(t_0^*, t_1^*, q_0, q_1, \delta_1^*, \rho^*) = \left(\frac{5}{9}, \frac{8}{9}, \frac{1}{3}, \frac{2}{3}, -\frac{4}{9}, 1\right). \tag{A.2.2}$$

This argument can be made slightly more general, and thus the counterexample more realistic, by relaxing $F_{\tilde{\varepsilon}} \sim Unif(0,1)$ and $\rho^* = 1$. We show that a similar argument goes through with $F_{\tilde{\varepsilon}}$ being a general distribution function with a symmetric density function, and $-1 \leq \rho^* \leq 1$ as long as the copula density is symmetric around $u_2 = u_1$ (i.e., the 45 degree line) and $u_2 = 1 - u_1$. Let $F \equiv F_{\tilde{\varepsilon}}$ be a general distribution whose density function is symmetric. Then there exists a

solution $s_0^\dagger = s_0^\dagger(t_0, q_0, \rho, \rho^*)$ such that

$$\begin{aligned} C(F(F^{-1}(t_0)+0), q_0; \rho) &= \Pr[u_1 \le t_0, u_2 \le q_0; \rho] \\ &= \Pr[u_1 \le s_0^\dagger, u_2 \le q_0; \rho^*] \\ &= C(s_0^\dagger, q_0; \rho^*). \end{aligned}$$

Then, by letting $\delta_1^* = F^{-1}(s_0^\dagger) - F^{-1}(t_0^*)$, we have $s_0^\dagger = F(F^{-1}(t_0^*) + \delta_1^*)$ and therefore $(t_0, q_0, \delta_1, \rho)$ and $(t_0^*, q_0, \delta_1^*, \rho^*)$ result in $p_{11,x} = C(F(F^{-1}(t_0)+0), q_0; \rho) = C(F(F^{-1}(t_0^*) + \delta_1^*), q_0; \rho^*)$. Suppose that $\delta_1 = 0$. Then there exists a solution $s_1^\dagger = s_1^\dagger(t_1, q_1, \rho, \rho^*)$ such that

$$\begin{aligned} C(F(F^{-1}(t_1)+0), q_1; \rho) &= \Pr[u_1 \le t_1, u_2 \le q_1; \rho] \\ &= \Pr[u_1 \le s_1^\dagger, u_2 \le q_1; \rho^*] \\ &= C(s_1^\dagger, q_1; \rho^*). \end{aligned}$$

Then, if we can show that

$$F^{-1}(s_1^\dagger) = F^{-1}(t_1^*) + \delta_1^*,$$

then $s_1^\dagger = F(F^{-1}(t_1^*) + \delta_1^*)$ and therefore $(t_1, q_1, \delta_1, \rho)$ and $(t_1^*, q_1, \delta_1^*, \rho^*)$ result in $\tilde{p}_{11,1} = C(F(F^{-1}(t_1)+0), q_1; \rho) = C(F(F^{-1}(t_1^*) + \delta_1^*), q_1; \rho)$. Note $F^{-1}(s_1^\dagger) = F^{-1}(t_1^*) + \delta_1^*$ can be rewritten as $F^{-1}(s_1^\dagger) = F^{-1}(t_1^*) + F^{-1}(s_0^\dagger) - F^{-1}(t_0^*)$ or

$$F^{-1}(s_1^\dagger) - F^{-1}(t_1^*) = F^{-1}(s_0^\dagger) - F^{-1}(t_0^*). \tag{A.2.3}$$

But note that since the density of F is symmetric, any two values s and $\tilde{s}$ in $(0,1)$

that are symmetric around $u_1 = 1/2$ will satisfy

$$F^{-1}(s) = -F^{-1}(\tilde{s}).$$

Therefore, since in our example $s_0^\dagger$ and t_1^* are symmetric around $u_1 = 1/2$, and so are $s_1^\dagger$ and t_0^*, we have the desired result (A.2.3), and the counterexample (A.2.1)–(A.2.2) remains valid. Note that the symmetry of the density function of F plays a key role here; the uniform distribution trivially satisfies the condition as does the normal distribution.

The above counter-example to identification involves a parameter on the boundary of the parameter space ($\rho^* = 1$), while the identification results in the paper assume that the parameter space is open and thus that $\rho \in (-1, 1)$. We now show that the key idea of the argument remains the same with $-1 < \rho^* < 1$. Suppose that the copula density is symmetric around $u_2 = u_1$ and $u_2 = 1 - u_1$. The normal copula satisfies this condition for any $\rho \in (-1, 1)$. Because of this condition, the symmetry of $s_0^\dagger$ and t_1^* (and of $s_1^\dagger$ and t_0^*) around $u_1 = 1/2$ does not break at a different value of ρ^*, even though the values of $s_0^\dagger$, t_1^*, $s_1^\dagger$, and t_0^* themselves change. Therefore, (A.2.3) continues to hold with $\rho^* \neq 1$.

A.3 Proof of Theorem 1.4.7

The following proposition is a modification of Theorem 3.1 in Chen (2007) and it establishes the consistency of sieve M-estimator [2].

[2] See also Remark 3.3 in Chen (2007).

Proposition A.3.1. *Let $\hat{\theta}_n$ be the sieve extremum estimator defined in Equation (1.4.2). Suppose that the following conditions hold :*

(i) $Q_0(\theta)$ is uniquely maximized at θ_0 in Θ and $Q_0(\theta_0) > -\infty$;

(ii) Θ is compact under $d_c(\cdot,\cdot)$, and $Q_0(\theta)$ is upper semicontinuous on Θ under $d_c(\cdot,\cdot)$;

(iii) The sieve spaces, Θ_n, is compact under $d_c(\cdot,\cdot)$;

(iv) $\Theta_k \subseteq \Theta_{k+1} \subseteq \Theta$ for all $k \geq 1$, and there exists a sequence $\pi_k\theta_0 \in \Theta_k$ such that $d_c(\theta_0, \pi_k\theta_0) \to 0$ as $k \to \infty$;

(v) For all $k \geq 1$, $p\lim_{n\to\infty} \sup_{\theta\in\Theta} |Q_n(\theta) - Q_0(\theta)| = 0$.

Then, $d_c(\hat{\theta}_n, \theta_0) = o_p(1)$.

We show that the conditions in Theorem 1.4.7 imply those in this proposition to prove consistency of the sieve estimator. We first need to verify that (i) the true parameter θ_0 is the unique maximizer of $Q_0(\cdot)$ over Θ and that (ii) the sample log-likelihood function $Q_n(\cdot)$ uniformly converges to $Q_0(\cdot)$ over the sieve space in probability to establish the consistency of the sieve ML estimator. The following lemma shows that if the model with unknown marginal distributions are identified and additional conditions are satisfied, then the true parameter θ_0 is the unique maximizer of $Q_0(\cdot)$ over Θ.

Lemma A.3.1. *Suppose that Assumptions 1.2.1-1.2.6, 1.2.9, 1.4.1 and 1.4.2 are satisfied. Then the condition (i) in Proposition A.3.1 is satisfied.*

Proof. By Theorem 1.2.10, the model is identified. Under Assumption 1.4.2, we can see that for any $\theta \in \Theta$, $|Q_0(\theta)| \leq \mathbb{E}|l(\theta, W_i)| \leq \sum_{y,d\in\{0,1\}} \mathbb{E}|\log(p_{yd,XZ}(\theta))| < \infty$, and thus the function $Q_0(\theta)$ is well-defined on Θ and $Q_0(\theta) > -\infty$ for all $\theta \in \Theta$; hence $Q_0(\theta_0) > -\infty$. Since the model is identified, it implies that for $\theta \neq \theta_0$, there exists a set $E \subset Supp(X, Z)$ such that $\int_E dP_{XZ} > 0$ and for some $y, d \in \{0, 1\}$, $\frac{p_{yd,xz}(\theta)}{p_{yd,xz}(\theta_0)} \neq 1$ on E, where P_{XZ} is the distribution function of (X, Z). Thus, we have

$$\begin{aligned} Q_0(\theta) - Q_0(\theta_0) &= \int \sum_{y,d\in\{0,1\}} p_{yd,xz}(\theta_0) \log(\frac{p_{yd,xz}(\theta)}{p_{yd,xz}(\theta_0)}) dP_{XZ} \\ &< \log(\int_E \sum_{y,d\in\{0,1\}} p_{yd,xz}(\theta) dP_{XZ}) \leq 0, \end{aligned}$$

where the strict inequality holds by the fact that $p_{yd,xz}(\theta) \neq p_{yd,xz}(\theta_0)$ on E and Jensen's inequality. Hence, θ_0 is the unique maximizer of $Q_0(\cdot)$. □

For any $\omega > 0$, let $N(\omega, \Theta_n, d_c)$ be the covering numbers without bracketing of Θ_n w.r.t the pseudo-metric d_c. We now establish the uniform convergence of $Q_n(\cdot)$ to Q_0 over the sieve space.

Lemma A.3.2. *Suppose that Assumptions 1.2.1-1.2.6, 1.2.9 are satisfied. If Assumptions 1.4.1 through 1.4.6 hold, then*

$$\sup_{\theta\in\Theta_n} |Q_n(\theta) - Q_0(\theta)| \xrightarrow{p} 0$$

for all $n \geq 1$.

Proof. We verify Condition 3.5M in Chen (2007). Let B stand for a generic constant and it can be different in each place. By Assumptions 1.4.2 and 1.4.3, the first

condition in Condition 3.5M is satisfied. Let $n \geq 1$ be a natural number and $\theta, \tilde{\theta} \in \Theta_n$. Define $R_1(\theta) = F_\epsilon(X'\beta + \delta_1)$, $R_0(\theta) = F_\epsilon(X'\beta)$, and $S(\theta) = F_\nu(X'\alpha + Z'\gamma)$. Similarly, we define $R_1(\tilde{\theta}) = \tilde{F}_\epsilon(X'\tilde{\beta} + \tilde{\delta}_1)$, $R_0(\tilde{\theta}) = \tilde{F}_\epsilon(X'\tilde{\beta})$, and $S(\tilde{\theta}) = \tilde{F}_\nu(X'\tilde{\alpha} + Z'\tilde{\gamma})$. For the simplicity of the notations, we denote $R_j(\theta) = R_j$, $R_j(\tilde{\theta}) = \tilde{R}_j$, $S(\theta) = S$, and $S(\tilde{\theta}) = \tilde{S}$ for all $j = 0, 1$. Observe that

$$\begin{aligned}
|p_{11,XZ}(\theta) - p_{11,XZ}(\tilde{\theta})| &= |C(R_1, S; \rho) - C(\tilde{R}_1, \tilde{S}; \tilde{\rho})| \\
&\leq |C(R_1, S; \rho) - C(\tilde{R}_1, \tilde{S}; \rho)| + |C(\tilde{R}_1, \tilde{S}; \rho) - C(\tilde{R}_1, \tilde{S}; \tilde{\rho})| \\
&\leq |R_1 - \tilde{R}_1| + |S - \tilde{S}| + |C_\rho(\tilde{R}_1, \tilde{S}; \hat{\rho})||\rho - \tilde{\rho}| \\
&\leq |R_1 - \tilde{R}_1| + |S - \tilde{S}| + B|\rho - \tilde{\rho}|
\end{aligned}$$

where $C_\rho(\cdot, \cdot; \cdot)$ is the partial derivative of $C(\cdot, \cdot; \cdot)$ with respect to ρ and $\hat{\rho}$ is between ρ and $\tilde{\rho}$ and $B < \infty$. Note that the last inequality holds due to a generic property of copulas (see, e.g. Theorem 2.2.4 in Nelsen (1999)) and the mean value theorem. We also have

$$\begin{aligned}
|R_1 - \tilde{R}_1| &= |F_\epsilon(X'\beta + \delta_1) - \tilde{F}_\epsilon(X'\tilde{\beta} + \tilde{\delta}_1)| \\
&\leq |F_\epsilon(X'\beta + \delta_1) - F_\epsilon(X'\tilde{\beta} + \tilde{\delta}_1)| + |F_\epsilon(X'\tilde{\beta} + \tilde{\delta}_1) - \tilde{F}_\epsilon(X'\tilde{\beta} + \tilde{\delta}_1)| \\
&\leq |f_\epsilon(X'\hat{\beta} + \hat{\delta}_1)| \cdot |X'(\beta - \tilde{\beta}) + (\delta_1 - \tilde{\delta}_1)| + \int_0^{G(X'\tilde{\beta} + \tilde{\delta}_1)} |h_\epsilon(t) - \tilde{h}_\epsilon(t)| dt \\
&\leq \sup_{x \in \mathbb{R}} |h_\epsilon(G(x))g(x)| \times ||(X', 1)'||_E \cdot ||\psi - \tilde{\psi}||_E + ||h_\epsilon - \tilde{h}_\epsilon||_\infty \\
&\leq B \times ||(X', 1)'||_E \times ||(\beta', \delta_1)' - (\tilde{\beta}', \tilde{\delta}_1)'||_E + ||h_\epsilon - \tilde{h}_\epsilon||_\infty, \qquad \text{(A.3.1)}
\end{aligned}$$

for some constant $B < \infty$. Similarly, we can show that

$$|R_0 - \tilde{R}_0| \leq B \times ||X||_E \times ||\beta - \tilde{\beta}||_E + ||h_\epsilon - \tilde{h}_\epsilon||_\infty \tag{A.3.2}$$

and

$$|S - \tilde{S}| \leq B \times ||(X^{'}, Z^{'})^{'}||_E \times ||(\alpha^{'}, \gamma^{'})^{'} - (\tilde{\alpha}^{'}, \tilde{\gamma}^{'})^{'}||_E + ||h_\nu - \tilde{h}_\nu||_\infty. \tag{A.3.3}$$

Note that, for any comparable subvectors ψ_s and $\tilde{\psi}_s$ of ψ and $\tilde{\psi}$, respectively, we have $||\psi_s - \tilde{\psi}_s||_E \leq ||\psi - \tilde{\psi}||_E$ and that, for any subvector W_s of W, we have $||W_S||_E \leq ||W||_E$ a.s. Thus we have

$$\begin{aligned} |p_{11,XZ}(\theta) - p_{11,XZ}(\tilde{\theta})| &\leq B||(X^{'}, 1)^{'}||_E \cdot ||\psi - \tilde{\psi}||_E + ||h_\epsilon - \tilde{h}_\epsilon||_\infty \\ &\leq B||(X^{'}, 1)^{'}||_E d_c(\theta, \tilde{\theta}) \end{aligned}$$

Consequently,

$$\begin{aligned} |p_{10,XZ}(\theta) - p_{10,XZ}(\tilde{\theta})| &\leq |R_0 - \tilde{R}_0| + |C(R_0, S; \rho) - C(\tilde{R}_0, \tilde{S}; \tilde{\rho})| \\ &\leq 2|R_0 - \tilde{R}_0| + |S - \tilde{S}| + B|\rho - \tilde{\rho}| \\ &\leq B\{||X||_E||\beta - \tilde{\beta}||_E + ||(X^{'}, Z^{'})^{'}||_E||(\alpha^{'}, \gamma^{'})^{'} - (\tilde{\alpha}^{'}, \tilde{\gamma}^{'})^{'}||_E \\ &+ ||h_\epsilon - \tilde{h}_\epsilon||_\infty + ||h_\nu - \tilde{h}_\nu||_\infty + |\rho - \tilde{\rho}|\} \\ &\leq B \cdot ||(X^{'}, Z^{'}, 1)^{'}||_E d_c(\theta, \tilde{\theta}), \end{aligned}$$

$$
\begin{aligned}
|p_{01,XZ}(\theta) - p_{01,XZ}(\tilde{\theta})| &\leq 2|S - \tilde{S}| + |R_1 - \tilde{R}_1| + B|\rho - \tilde{\rho}| \\
&\leq B||(X^{'}, Z^{'}, 1)^{'}||_E d_c(\theta, \tilde{\theta}),
\end{aligned}
$$

and

$$
\begin{aligned}
&|p_{00,XZ}(\theta) - p_{00,XZ}(\tilde{\theta})| \\
\leq &|p_{11,XZ}(\theta) - p_{11,XZ}(\tilde{\theta})| + |p_{10,XZ}(\theta) - p_{10,XZ}(\tilde{\theta})| + |p_{01,XZ}(\theta) - p_{01,XZ}(\tilde{\theta})| \\
\leq &B||(X^{'}, Z^{'}, 1)^{'}||_E d_c(\theta, \tilde{\theta}).
\end{aligned}
$$

In all, we have

$$
\begin{aligned}
|l(\theta, W_i) - l(\tilde{\theta}, W_i)| &\leq \sum_{y,d=0,1} \mathbf{1}_{yd}(Y_i, D_i) \cdot |\log p_{yd}(X_i, Z_i; \theta) - \log p_{yd}(X_i, Z_i; \tilde{\theta})| \\
&\leq \frac{1}{\underline{p}(X_i, Z_i)} \sum_{y,d=0,1} \mathbf{1}_{yd}(Y_i, D_i)|p_{yd}(X_i, Z_i; \theta) - p_{yd}(X_i, Z_i; \tilde{\theta})| \\
&\leq \frac{B}{\underline{p}(X_i, Z_i)}||(X_i^{'}, Z_i^{'}, 1)^{'}||_E d_c(\theta, \tilde{\theta}) \\
&\equiv U(W_i) d_c(\theta, \tilde{\theta}), \qquad \text{(A.3.4)}
\end{aligned}
$$

where $\mathbb{E}[U(W_i)^2] < \infty$ by Assumptions 1.4.2 and 1.4.3. This results in

$$
\sup_{\theta, \tilde{\theta} \in \Theta_n, d_c(\theta, \tilde{\theta}) \leq \epsilon_0} |l(\theta, W_i) - l(\tilde{\theta}, W_i)| \leq U(W_i)\epsilon_0 \qquad \text{(A.3.5)}
$$

and thus the second condition in Condition 3.5M is satisfied with $s = 1$.

For the last condition in Condition 3.5M, note that for any $\omega > 0$, we have

$$N(\omega, \Theta_n, d_c) \leq N(\frac{\omega}{2}, \Psi, ||\cdot||_E) \cdot N(\frac{\omega}{4}, \mathcal{H}_{en}, ||\cdot||_\infty) \cdot N(\frac{\omega}{4}, \mathcal{H}_{\nu n}, ||\cdot||_\infty).$$

By Lemma 2.5 in van de Geer (2000), we have $\log N(\frac{\omega}{4}, \mathcal{H}_{en}, ||\cdot||_\infty) \leq k_n \log(1+\frac{32R}{\omega})$ under Assumption 1.4.5-(i); and hence

$$\begin{aligned} \log N(\omega, \Theta_n, d_c) &\leq const. \times k_n \times \log(1 + \frac{32R}{\omega}) \\ &= o(n) \end{aligned}$$

if $k_n/n \to 0$. Since the condition $k_n/n = o(1)$ is imposed by Assumption 1.4.5-(i), the last condition in Condition 3.5M is also satisfied. In all, we have the uniform convergence of Q_n to Q_0 over Θ_n. □

To finish proving Theorem 1.4.7, we verify the conditions in Proposition A.3.1. By Lemmas A.3.1 and A.3.2, the conditions (i) and (v) in Proposition A.3.1 are satisfied. Using Equation (A.3.4) and Jensen's inequality, we can see that, for any $\theta, \tilde{\theta} \in \Theta$,

$$\begin{aligned} |Q_0(\theta) - Q_0(\tilde{\theta})| &\leq E|l(\theta, W_i) - l(\tilde{\theta}, W_i)| \\ &\leq E[U(W_i)]d_c(\theta, \tilde{\theta}) \\ &= B \cdot d_c(\theta, \tilde{\theta}) \end{aligned}$$

for some $B < \infty$. Thus, $Q_0(\cdot)$ is continuous with respect to d_c. As mentioned before, the parameter space Θ is compact under d_c and thus the conditions (ii) and (iii) are satisfied with the specified parameter space and the norm. Since the condition (iv)

is directly imposed, we have $d(\hat{\theta}_n, \theta_0) = o_p(1)$ by Proposition A.3.1.

A.4 Proof of Theorem 1.4.9

We derive the convergence rate of the sieve M-estimator w.r.t. the norm $||\cdot||_2$ by checking the conditions in Theorem 3.2 in Chen (2007). Since $\{W_i\}_{i=1}^n$ is assumed to be i.i.d by Assumption 1.4.3, Condition 3.6 in Chen (2007) is satisfied. For Condition 3.7 in Chen (2007), we note that for a small $\epsilon_1 > 0$ and for any $\theta \in \Theta_n$ such that $||\theta - \theta_0|| \leq \epsilon_1$, we have

$$\begin{aligned}
&Var(l(\theta, W_i) - l(\theta_0, W_i)) \\
\leq &\mathbb{E}[l(\theta, W_i) - l(\theta_0, W_i)]^2 \\
\leq &\mathbb{E}[\frac{1}{\underline{p}(X_i, Z_i)^2} \sum_{y,d=0,1} \mathbf{1}_{yd}(Y_i, D_i) |p_{yd}(X_i, Z_i; \theta) - p_{yd}(X_i, Z_i; \theta_0)|^2] \\
\leq &\mathbb{E}[\frac{1}{\underline{p}(X_i, Z_i)^2} \sum_{y,d\in\{0,1\}} |p_{yd}(X_i, Z_i; \theta) - p_{yd}(X_i, Z_i; \theta_0)|^2].
\end{aligned}$$

By the same logic in Equation (A.3.4), we have

$$Var(l(\theta, W_i) - l(\theta_0, W_i)) \leq \mathbb{E}[U(W_i)^2] d_c(\theta, \theta_0)^2.$$

Note that

$$\begin{aligned}
d_c(\theta, \theta_0)^2 &= (||\psi - \psi_0||_E + ||h_\epsilon - h_{\epsilon 0}||_\infty + ||h_\nu - h_{\nu 0}||_\infty)^2 \\
&\leq 4(||\psi - \psi_0||_E^2 + ||h_\epsilon - h_{\epsilon 0}||_\infty^2 + ||h_\nu - h_{\nu 0}||_\infty^2).
\end{aligned}$$

By Lemma 2 in Chen and Shen (1998), we have

$$||h_j - h_{j0}||_\infty^2 \leq ||h_j - h_{j0}||_2^{\frac{4p}{2p+1}} \tag{A.4.1}$$

for all $j \in \{\epsilon, \nu\}$. Since $\frac{4p}{2p+1} > 1$ under Assumption 1.4.4, we can show that

$$\sup_{\{\theta \in \Theta_n : ||\theta - \theta_0||_2 \leq \epsilon_1\}} Var(l(\theta, W_i) - l(\theta_0, W_i)) \leq B_1 \epsilon_1^2$$

with $\epsilon_1 \leq 1$ and some constant B_1, and thus Condition 3.7 in Chen (2007) is satisfied.

We recall Equation (A.3.4) to verify Condition 3.8 in Chen (2007). Let $\epsilon_2 > 0$ be given and consider

$$\begin{aligned}
&|l(\theta, W_i) - l(\theta_0, W_i)| \\
\leq& U(W_i)\{||\psi - \psi_0||_E + ||h_\epsilon - h_{\epsilon 0}||_\infty + ||h_\nu - h_{\nu 0}||_\infty\} \\
\leq& U(W_i)\{||\psi - \psi_0||_E + ||h_\epsilon - h_{\epsilon 0}||_2^{\frac{2p}{2p+1}} + ||h_\nu - h_{\nu 0}||_2^{\frac{2p}{2p+1}}\} \\
\leq& U(W_i)\{||\psi - \psi_0||_E^{\frac{2p+1}{2p}} + ||h_\epsilon - h_{\epsilon 0}||_2 + ||h_\nu - h_{\nu 0}||_2\}^{\frac{2p}{2p+1}} \\
\leq& U(W_i)\{||\psi - \psi_0||_E \times (\sup_{\psi \in \Psi} ||\psi|| + ||\psi_0||)^{\frac{1}{2p}} + ||h_\epsilon - h_{\epsilon 0}||_2 + ||h_\nu - h_{\nu 0}||_2\}^{\frac{2p}{2p+1}} \\
\leq& \tilde{U}(W_i)\{||\psi - \psi_0||_E + ||h_\epsilon - h_{\epsilon 0}||_2 + ||h_\nu - h_{\nu 0}||_2\}^{\frac{2p}{2p+1}},
\end{aligned} \tag{A.4.2}$$

where $\tilde{U}(W_i) = \max\{1, (\sup_{\psi \in \Psi} ||\psi|| + ||\psi_0||)^{\frac{1}{2p}}\} \times U(W_i)$. Since the parameter space for ψ, Ψ, is compact under Assumption 1.4.1, $E[\tilde{U}(W_i)^2] < \infty$. Thus, we have

$$\sup_{\{\theta \in \Theta_n : ||\theta - \theta_0|| \leq \epsilon_2\}} |l(\theta, W_i) - l(\theta_0, W_i)| \leq \epsilon_2^{\frac{2p}{2p+1}} \tilde{U}(W_i)$$

with $\mathbb{E}[\tilde{U}_i(W_i)^2] < \infty$ and this implies that, under Assumption 1.4.4, Condition 3.8 in Chen (2007) is satisfied with $s = \frac{2p}{2p+1} \in (0, 2)$ and $\gamma = 2$.

Let $\mathcal{L}_n \equiv \{l(\theta_0, W_i) - l(\theta, W_i) : \theta \in \Theta_n, ||\theta - \theta_0||_2 \leq \epsilon_2\}$. We now need to calculate κ_n which is defined as

$$\kappa_n \equiv \inf\{\kappa \in (0,1) : \frac{1}{\sqrt{n}\kappa^2} \int_{b\kappa^2}^{\kappa} \sqrt{H_{[]}(\omega, \mathcal{L}_n, ||\cdot||_{L^2})} d\omega \leq const.\},$$

where, for $f \in \mathcal{L}_n$, $||f(\theta, W_i)||_{L^2}^2 \equiv E[f(\theta, W_i)^2]$ is the L^2-norm on $\mathcal{L}_n$ and $H_{[]}(\omega, \mathcal{L}_n, ||\cdot||_{L^2})$ is the L_2-metric entropy with bracketing of the class $\mathcal{L}_n$ (see van der Vaart and Wellner (1996) or van de Geer (2000) for the definition of L_2-metric entropy with bracketing). Let $B_0 = \mathbb{E}[U(W_i)^2]$, where $U(W_i)$ is the same to the one in Equation (A.3.4). By Theorem 2.7.11 in van der Vaart and Wellner (1996) and Equation (A.3.4), we can show that

$$\begin{aligned} N_{[]}(\omega, \mathcal{L}_n, ||\cdot||_{L^2}) &\leq N(\frac{\omega}{2B_0}, \Theta_n, d_c) \\ &\leq N(\frac{\omega}{4B_0}, \Psi, ||\cdot||_E) \cdot N(\frac{\omega}{8B_0}, \mathcal{H}_{\epsilon n}, ||\cdot||_\infty) \cdot N(\frac{\omega}{8B_0}, \mathcal{H}_{\nu n}, ||\cdot||_\infty), \end{aligned}$$

and this leads to

$$\begin{aligned} H_{[]}(\omega, \mathcal{L}_n, ||\cdot||_{L^2}) &= \log(N_{[]}(\omega, \mathcal{L}_n, ||\cdot||_{L^2})) \\ &\leq \log(N(\frac{\omega}{4B_0}, \Psi, ||\cdot||_E)) + \log(N(\frac{\omega}{8B_0}, \mathcal{H}_{\epsilon n}, ||\cdot||_\infty)) \\ &+ \log(N(\frac{\omega}{8B_0}, \mathcal{H}_{\epsilon n}, ||\cdot||_\infty)) \\ &\leq const. \times k_n \times \log(1 + \frac{64B_0R}{\omega}). \end{aligned}$$

In all, κ_n solves

$$\begin{aligned}\frac{1}{\sqrt{n}\kappa_n^2}\int_{b\kappa_n^2}^{\kappa_n}\sqrt{H_{[]}(\omega,\mathcal{L}_n,||\cdot||_{L^2})}d\omega &\leq \frac{const.}{\sqrt{n}\kappa_n^2}\int_{b\kappa_n^2}^{\kappa_n}\sqrt{k_n\cdot\log(1+\frac{64B_0R}{\omega})}d\omega \\ &\leq \frac{const.}{\sqrt{n}\kappa_n^2}\sqrt{k_n}\int_{b\kappa_n^2}^{\kappa_n}\sqrt{\frac{1}{\omega}}d\omega \\ &\leq const.\times\frac{1}{\sqrt{n}\kappa_n^2}\sqrt{k_n}\kappa_n \leq const.\end{aligned}$$

and thus $\kappa_n \propto \sqrt{\frac{k_n}{n}}$.

Lastly, since $||\theta_0-\pi_n\theta_0||_2 \leq ||\theta_0-\pi_n\theta_0||_c = O(k_n^{-p})$ by Lorentz (1966), we have

$$||\hat{\theta}_n-\theta_0||_2 = O_p(\max\{\sqrt{\frac{k_n}{n}},k_n^{-p}\})$$

by Theorem 3.2 in Chen (2007). By choosing $k_n \propto n^{\frac{1}{2p+1}}$, we have

$$||\hat{\theta}_n-\theta_0||_2 = O_p(n^{-\frac{p}{2p+1}}).$$

A.5 Proof of Proposition 1.4.1

Note that since the sieve ML estimator $\hat{\theta}_n$ is consistent w.r.t the pseudo-metric d_c by Theorem 1.4.7, it is consistent with respect to the norm $||\cdot||_2$ and thus with respect to the Fisher norm by Equation (1.4.7). We also point out that $||\hat{\theta}_n-\theta_0|| = O_p(n^{-\frac{p}{2p+1}})$ by Equation (1.4.7) and Theorem 1.4.9 under the given set of Assumptions. We follow the proof of Theorem 1 in CFT06. Assumptions 1 and 2 in CFT06 are implied by Assumption 1.2.1-1.2.6, 1.2.9-1.4.2, and 1.4.10. The first two parts in Assumption 1.4.11 correspond to Assumption 3 in CFT06 . Since $p > 1/2$ by Assumption 1.4.4, $||\hat{\theta}_n-\theta_0|| = o_p(n^{-1/4})$ by Theorem 1.4.9 and this

implies that $||\hat{\theta}_n - \theta_0|| \times ||\pi_n v^* - v^*|| = o(n^{-1/2})$ under Assumption 1.4.11 (iii). In addition, since $w > 1 + \frac{1}{2p}$, $\delta_n^w = o(n^{-1/2})$ by that $||\hat{\theta}_n - \theta_0|| = O_p(n^{-\frac{p}{2p+1}})$. Hence, Assumptions 3 and 4 in CFT06 are satisfied.

Define $r[\theta, \theta_0, W_i] \equiv l(\theta, W_i) - l(\theta_0, Z_i) - \frac{\partial l(\theta_0, W_i)}{\partial \theta'}[\theta - \theta_0]$ and $\xi_0 = 2\xi_1 + \xi_2$. Let ζ_n be a positive sequence with $\zeta_n = o(n^{-1/2})$ and $(\delta_n)^{3-(2\xi_1+\xi_2)} = \zeta_n o(n^{-1/2})$. Then we have

$$\begin{aligned}
0 &\leq \frac{1}{n}\sum_{i=1}^{n} l(\hat{\theta}_n, W_i) - l(\hat{\theta}_n \pm \zeta_n \pi_n v^*, W_i) \\
&\leq \mp \zeta_n \frac{1}{n}\sum_{i=1}^{n} \frac{\partial l(\theta_0, W_i)}{\partial \theta'}[\pi_n v^*] \\
&+ \mu_n(r[\hat{\theta}_n, \theta_0, W_i] - r[\hat{\theta}_n \pm \zeta_n \pi_n v^*, \theta_0, W_i]) \\
&+ \mathbb{E}[r[\hat{\theta}_n, \theta_0, W_i] - r[\hat{\theta}_n \pm \zeta_n \pi_n v^*, \theta_0, W_i]]. \qquad \text{(A.5.1)}
\end{aligned}$$

We first note that, by Assumption 1.4.11 (iii),

$$\begin{aligned}
\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} \frac{\partial l(\theta_0, W_i)}{\partial \theta'}[\pi_n v^* - v^*]]^2 &\leq \frac{1}{n}\mathbb{E}[\{\frac{\partial l(\theta_0, W_i)}{\partial \theta'}[\pi_n v^* - v^*]\}^2] \\
&= \frac{1}{n}||\pi_n v^* - v^*||^2 \\
&= o(n^{-1}), \qquad \text{(A.5.2)}
\end{aligned}$$

and hence $\frac{1}{n}\sum_{i=1}^{n} \frac{\partial l(\theta_0, W_i)}{\partial \theta'}[\pi_n v^* - v^*] = o_p(n^{-1/2})$.

Observe that, by the mean value theorem,

$$\begin{aligned}\mathbb{E}[r[\theta,\theta_0,W_i]] &= \mathbb{E}[l(\theta,W_i)-l(\theta_0,W_i)-\frac{\partial l(\theta_0,W_i)}{\partial\theta'}[\theta-\theta_0]] \\ &= \mathbb{E}[\frac{1}{2}\frac{\partial^2 l(\theta_0,W_i)}{\partial\theta\partial\theta'}[\theta-\theta_0,\theta-\theta_0]] \\ &+ \frac{1}{2}\mathbb{E}[\frac{\partial^2 l(\tilde{\theta},W_i)}{\partial\theta\partial\theta'}[\theta-\theta_0,\theta-\theta_0]-\frac{\partial^2 l(\theta_0,W_i)}{\partial\theta\partial\theta'}[\theta-\theta_0,\theta-\theta_0]],\end{aligned} \tag{A.5.3}$$

where $\theta,\tilde{\theta}\in\Theta_n$ and $\tilde{\theta}$ is between θ and θ_0. In addition, for any $v=(v_\psi',v_\epsilon,v_\nu)'\in\mathbb{V}$ and $\tilde{\theta}\in\Theta_n$ with $||\tilde{\theta}-\theta_0||=O(\delta_n)$, we have

$$\begin{aligned}\mathbb{E}[\frac{\partial^2 l(\tilde{\theta},W_i)}{\partial\theta\partial\theta'}[v,v]-\frac{\partial^2 l(\theta_0,W_i)}{\partial\theta\partial\theta'}[v,v]] &= v_\psi'\mathbb{E}[\frac{\partial^2 l(\tilde{\theta},W_i)}{\partial\psi\partial\psi'}-\frac{\partial^2 l(\theta_0,W_i)}{\partial\psi\partial\psi'}]v_\psi \\ &+ \sum_{j\in\{\epsilon,\nu\}} 2v_\theta'\mathbb{E}[(\frac{\partial^2 l(\tilde{\theta},W_i)}{\partial\psi\partial h_j}[v_j]-\frac{\partial^2 l(\theta_0,W_i)}{\partial\psi\partial h_j}[v_j])] \\ &+ \sum_{k,j\in\{\epsilon,\nu\}}\mathbb{E}[\frac{\partial^2 l(\tilde{\theta},W_i)}{\partial h_k\partial h_j}[v_k,v_j]-\frac{\partial^2 l(\theta_0,W_i)}{\partial h_k\partial h_j}[v_k,v_j]],\end{aligned}$$

and this term can be controlled under Assumption 1.4.13 in the same way of CFT06. This leads us to that

$$\begin{aligned}&\mathbb{E}[r[\hat{\theta}_n,\theta_0,W_i]-r[\hat{\theta}_n\pm\zeta_n\pi_n v^*,\theta_0,W_i]] \\ &= -\frac{1}{2}(||\hat{\theta}_n-\theta_0||^2-||\hat{\theta}_n\pm\zeta_n\pi_n v^*-\theta_0||)+\zeta_n o(n^{-1/2}) \\ &= \pm\zeta_n\times<\hat{\theta}_n-\theta_0,v^*>+\zeta_n o(n^{-1/2})\end{aligned} \tag{A.5.4}$$

because we have $<\hat{\theta}_n-\theta_0,\pi_n v^*-v^*>=o_p(n^{-1/2})$ and $||\pi_n v^*||^2\to||v^*||^2<\infty$.

We also point out that

$$\begin{aligned}
&\mu_n(r[\hat{\theta}_n, \theta_0, W_i] - r[\hat{\theta}_n \pm \zeta_n \pi_n v^*, \theta_0, W_i]) \\
=&\mu_n(l(\hat{\theta}_n, W_i) - l(\hat{\theta}_n \pm \zeta_n \pi_n v^*, W_i) - \frac{\partial l(\theta_0, W_i)}{\partial \theta'}[\mp \zeta_n \pi_n v^*]) \\
=&\mp \zeta_n \cdot \mu_n(\frac{\partial l(\tilde{\theta}, W_i)}{\partial \theta'}[\pi_n v^*] - \frac{\partial l(\theta_0, W_i)}{\partial \theta'}[\pi_n v^*]),
\end{aligned}$$

where $\tilde{\theta} \in \Theta_n$ is between $\hat{\theta}_n$ and $\hat{\theta}_n \pm \zeta_n \pi_n v^*$. By Assumption 1.4.14, we have

$$\mu_n(r[\hat{\theta}_n, \theta_0, W_i] - r[\hat{\theta}_n \pm \zeta_n \pi_n v^*, \theta_0, W_i]) = o_p(\zeta_n n^{-1/2}). \tag{A.5.5}$$

Combining Equations (A.5.1) through (A.5.5) with the fact that $\mathbb{E}[\frac{\partial l(\theta_0, W_i)}{\partial \theta'}[v^*]] = 0$, we have

$$\begin{aligned}
0 \leq& \frac{1}{n}\sum_{i=1}^{n} l(\hat{\theta}_n, W_i) - l(\hat{\theta}_n \pm \zeta_n \pi_n v^*, W_i) \\
=& \mp \zeta_n \cdot \mu_n(\frac{\partial l(\theta_0, W_i)}{\partial \theta'}[v^*]) \pm \zeta_n < \hat{\theta}_n - \theta_0, v^* > + \zeta_n \cdot o_p(n^{-1/2}),
\end{aligned}$$

and this results in that

$$\sqrt{n} < \hat{\theta}_n - \theta_0, v^* > = \sqrt{n}\mu_n(\frac{\partial l(\theta_0, W_i)}{\partial \theta'}[v^*]) + o_p(1) \xrightarrow{d} N(0, ||v^*||^2).$$

By Assumption 1.4.11, we have

$$\sqrt{n}(T(\hat{\theta}_n) - T(\theta_0)) = \sqrt{n} < \hat{\theta}_n - \theta_0, v^* > \xrightarrow{d} N(0, ||v^*||^2)$$

by the same way in CFT06.

A.6 Proof of Theorem 1.4.16

Take any $\lambda \in \mathbb{R}^{d_\psi} - \{0\}$. Assumption 1.4.11-(i) is satisfied with $w = \infty$ in the case of $T(\theta) = \lambda^{'}\psi$ and Assumption 1.4.15 implies Assumption 1.4.11-(ii). Hence, by Proposition 1.4.1, we have

$$\sqrt{n}(\lambda^{'}\hat{\psi}_n - \lambda^{'}\psi_0) \Rightarrow N(0, \lambda^{'}\mathfrak{I}_*(\psi_0)^{-1}\lambda).$$

Since λ was arbitrary, we obtain the result by Cramer-Wold device.

A.7 Hölder ball

Suppose that $h \in \Lambda_R^p([0,1])$, where $p = m + \zeta$, $m \geq 0$ is an integer and $\zeta \in (0,1]$ is the Hölder exponent. We want to show that $h^2 \in \Lambda_{\tilde{R}}^p([0,1])$, where $\tilde{R} = R^2 2^{m+1}$. We note that $||h||_\infty \leq R$ and thus $\sup_x |\mathcal{D}^\omega h(x)| \leq R$ for all $\omega \leq m$. By Leibniz's formula, we have

$$\begin{aligned}
|\mathcal{D}^\omega h^2(x)| &= |\mathcal{D}^\omega (h \cdot h)| \\
&= |\sum_{\iota \leq \omega} \binom{\omega}{\iota} \mathcal{D}^\iota h \mathcal{D}^{\omega - \iota} h| \\
&\leq R^2 \sum_{\iota \leq \omega} \binom{\omega}{\iota} = R^2 2^\omega \leq K^2 2^m < \infty
\end{aligned}$$

for all $\omega \leq m$. Observe that, by Leibniz's formula, for any $x, y \in [0,1]$ with $x \neq y$,

$$\begin{aligned}
|\mathcal{D}^m h^2(x) - \mathcal{D}^m h^2(y)| &= |\sum_{\omega \leq m} \binom{m}{\omega} \mathcal{D}^\omega h(x) \mathcal{D}^{m-\omega} h(x) - \sum_{\omega \leq m} \binom{m}{\omega} \mathcal{D}^\omega h(y) \mathcal{D}^{m-\omega} h(y)| \\
&\leq |\sum_{\omega \leq m} \binom{m}{\omega} \mathcal{D}^\omega h(x) \mathcal{D}^{m-\omega} h(x) - \sum_{\omega \leq m} \binom{m}{\omega} \mathcal{D}^\omega h(y) \mathcal{D}^{m-\omega} h(x)| \\
&+ |\sum_{\omega \leq m} \binom{m}{\omega} \mathcal{D}^\omega h(y) \mathcal{D}^{m-\omega} h(x) - \sum_{\omega \leq m} \binom{m}{\omega} \mathcal{D}^\omega h(y) \mathcal{D}^{m-\omega} h(y)| \\
&\leq 2 \times \{\sup_{\omega \leq m} \sup_x |\mathcal{D}^\omega h(x)|\} \times |\sum_{\omega \leq m} \binom{m}{\omega} \{\mathcal{D}^\omega h(x) - \mathcal{D}^\omega h(y)\}| \\
&\leq 2R \sum_{\omega \leq m} \binom{m}{\omega} |\mathcal{D}^\omega h(x) - \mathcal{D}^\omega h(y)|.
\end{aligned}$$

We also have that, for all $\omega < m$,

$$\begin{aligned}
\frac{|\mathcal{D}^\omega h(x) - \mathcal{D}^\omega h(y)|}{|x-y|^\zeta} &= \frac{|\mathcal{D}^\omega h(x) - \mathcal{D}^\omega h(y)|}{|x-y|} |x-y|^{1-\zeta} \\
&= |\mathcal{D}^{\omega+1} h(\tilde{x})| |x-y|^{1-\zeta} \\
&\leq R,
\end{aligned}$$

where $\tilde{x}$ is between x and y. Note that $\zeta \in (0,1]$ and thus $|x-y|^{1-\zeta} \leq 1$ for all $x, y \in [0,1]$. Since $h \in \Lambda_R^p([0,1])$, we have $\frac{|\mathcal{D}^m h(x) - \mathcal{D}^m h(y)|}{|x-y|^\zeta} \leq R$. Hence,

$$\begin{aligned}
\frac{|\mathcal{D}^m h^2(x) - \mathcal{D}^m h^2(y)|}{|x-y|^\zeta} &\leq 2R \sum_{\omega \leq m} \binom{m}{\omega} \frac{|\mathcal{D}^\omega h(x) - \mathcal{D}^\omega h(y)|}{|x-y|^\zeta} \\
&\leq 2R^2 \sum_{\omega \leq m} \binom{m}{\omega} = R^2 2^{m+1} < \infty,
\end{aligned}$$

and this implies that $h^2 \in \Lambda_{\tilde{R}}^p([0,1])$ with $\tilde{R} = R^2 2^{m+1}$.

Appendix B

Chapter 2 Appendix

Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space. For $x \in \mathbb{R}^d$,$||x||_E$ means the Euclidean norm of x in $\mathbb{R}^d$. Let $l^2(\mathcal{W})$ be the space of functions that are square-integrable on a set $\mathcal{W}$. Similarly, define $l^\infty(\mathcal{W})$ as the space of functions that are uniformly bounded on a set $\mathcal{W}$. For a generic function g on a set $\mathcal{W}$, $||g||_2 \equiv (\int_{\mathcal{W}} g^2 dP)^{1/2}$ and $||g||_\infty \equiv \sup_{w \in \mathcal{W}} |g(w)|$ are the L^2- and $\sup-$ norm, respectively. The expectation of g is denoted by $\mathbb{E}g \equiv \int g(w) dF_W(w)$, where $F_W(\cdot)$ is the distribution function of W. For a sequence of random maps $X_n : \Omega \to \mathbb{R}$ and a random variable X, $X_n \Rightarrow X$ ($X_n \xrightarrow{d} X$, resp.) indicates that X_n converges weakly (in distribution, resp.) to X in the sense of Definition 1.3.3 in van der Vaart and Wellner (1996). For any real numbers a and b, let $a \wedge b \equiv \min(a, b)$ and $a \vee b \equiv \max(a, b)$. For any real sequences (a_n) and (b_n), $a_n \lesssim b_n$ means that there is a constant $C > 0$ such that $|a_n| \leq C \cdot |b_n|$ for all $n \in \mathbb{N}$. For a set A, denote the interior of A by $int(A)$.

B.1 Proof of Lemma 2.2.1

Proof. Define

$$H_1(t|X) \equiv \Pr(Y \leq t, D = 1|X),$$
$$H_0(t|X) \equiv \Pr(Y \leq t, D = 0|X),$$
$$\tilde{H}(t|X) \equiv \Pr(Y > t|X).$$

Then, one can show that

$$\begin{aligned} H_1(t|X) &= \mathbb{E}[\mathbf{1}(Y \leq t, D = 1)|X] \\ &= \mathbb{E}[\mathbb{E}[\mathbf{1}(Y \leq t, D = 1)|X, T]|X] \\ &= \mathbb{E}[\mathbb{E}[\mathbf{1}(Y \leq t)|X, T, D = 1] \Pr(D = 1|X, T)|X] \\ &= \mathbb{E}[\mathbf{1}(T \leq t)(1 - F_{C|X}(T|X))|X] \\ &= \int_0^t (1 - F_{C|X}(s|X))dF_{T|X}(s|X), \end{aligned} \tag{B.1.1}$$

and hence $dH_1(t|X) = (1 - G(t|X))dF_{T|X}(t|X)$. Observing that the event $\{Y \geq t\}$ is equivalent to the event $\{T \geq t, C \geq t\}$, one can show that

$$\begin{aligned} \tilde{H}(t|X) &= \mathbb{E}[\mathbf{1}(Y > t)|X] \\ &= \mathbb{E}[\mathbf{1}(T > t, C > t)|X] \\ &= (1 - F_{T|X}(t|X))(1 - F_{C|X}(t|X)) \end{aligned}$$

by the conditional independence of T and C given X. Therefore,

$$\begin{aligned}\int_0^t \frac{dH_1(s|X)}{\tilde{H}(s|X)} &= \int_0^t \frac{(1-F_{C|X}(s|X))dF_{T|X}(s|X)}{(1-F_{T|X}(s|X))(1-F_{C|X}(s|X))} \\ &= \int_0^t \frac{dF_{T|X}(s|X)}{(1-F_{T|X}(s|X))} \\ &= -\ln(1-F_{T|X}(t|X)). \end{aligned} \tag{B.1.2}$$

Since H_1 and $\tilde{H}$ are identified from data, Equation (B.1.2) implies that $F_{T|X}(t|X)$ is identified. For identification of $F_{C|X}$, one can show that $H_0(t|X) = \int_0^t (1 - F_{T|X}(s|X)dF_{C|X}(s|X)$ by a similar way to (B.1.1), and thus $dH_0(t|X) = (1 - F_{T|X}(t|X))dF_{C|X}(t|X)$. Therefore,

$$\int_0^t \frac{dH_0(s|X)}{\tilde{H}(s|X)} = -\ln(1-F_{C|X}(t|X))$$

and this leads to identification of $F_{C|X}$ by the same logic above. □

B.2 Proof of Lemma 2.2.2

Proof. Suppose that $Q_{T|X}(\tau|X_i) = Q_{T|X_1}(\tau|X_{1i})$ almost surely. Note that

$$\begin{aligned}&\mathbb{E}[D_i\{\mathbf{1}(Y_i \le Q_{T|X_1}(\tau|X_{1i})) - \tau\}|X_i] \\ =&\mathbb{E}[\mathbf{1}(Y_i \le Q_{T|X_1}(\tau|X_{1i})) - \tau|D_i = 1, X_i]\Pr(D_i = 1|X_i) \\ =&\mathbb{E}[\mathbf{1}(T_i \le Q_{T|X_1}(\tau|X_{1i})) - \tau|X_i]\Pr(D_i = 1|X_i) \\ =&0 \cdot p_0(X_i) = 0\end{aligned}$$

under the null hypothesis. Conversely, suppose that the null hypothesis in (2.2.4) holds. Then one can show that

$$\begin{aligned}
0 &= \mathbb{E}[D_i\{\mathbf{1}(Y_i \leq Q_{T|X_1}(\tau|X_{1i})) - \tau\}|X_i] \\
&= \mathbb{E}[\mathbf{1}(Y_i \leq Q_{T|X_1}(\tau|X_{1i})) - \tau|D_i = 1, X_i] \Pr(D_i = 1|X_i) \\
&= \mathbb{E}[\mathbf{1}(T_i \leq Q_{T|X_1}(\tau|X_{1i}) - Q_{T|X}(\tau|X_i) + Q_{T|X}(\tau|X_i)) - \tau|X_i] \Pr(D_i = 1|X_i) \\
&= \{F_{T|X}(Q_{T|X_1}(\tau|X_{1i}) - Q_{T|X}(\tau|X_i) + Q_{T|X}(\tau|X_i)|X_i) - \tau\} \Pr(D_i = 1|X_i).
\end{aligned}$$

Since $p_0(x) > 0$ uniformly in $x \in \mathcal{X}$, the above equation implies that

$$F_{T|X}(Q_{T|X_1}(\tau|X_{1i}) - Q_{T|X}(\tau|X_i) + Q_{T|X}(\tau|X_i)|X_i) = \tau.$$

Since the conditional quantile is assumed to be unique, $Q_{T|X_1}(\tau|X_{1i}) - Q_{T|X}(\tau|X_i) + Q_{T|X}(\tau|X_i) = Q_{T|X}(\tau|X_i)$ and this leads to Equation (2.2.3). □

B.3 Proof of Theorem 2.3.7

Proof. Define $g_i(t;\tau) \equiv D_i\{\mathbf{1}(Y_i \leq Q_{T|X_1}(\tau|X_{1i})) - \tau\}\psi(X_i, t)$ and $\mathcal{G} \equiv \{g_i(t;\tau) : t \in \mathcal{I}\}$. For any $t_1, t_2 \in \mathcal{I}$, it can be shown that

$$|g_i(t_1) - g_i(t_2)| \leq |D_i\{\mathbf{1}(Y_i \leq Q_{T|X_1}(\tau|X_{1i})) - \tau\}| \cdot G_\Psi(X_i)||t_1 - t_2||_E.$$

Since $\mathbb{E}[\{\mathbf{1}(Y_i \leq Q_{T|X_1}(\tau|X_{1i})) - \tau\}D_i G_\Psi(X_i)]^2 < \infty$, it follows that $\mathcal{G}$ is a *type IV* class defined in Andrews (1994b) and thus $\mathcal{G}$ satisfies Ossiander's L^2 entropy condition by Theorem 5 in Andrews (1994b). Applying Theorem 3.1 in Ossiander (1987) yields that $\mathcal{F}$ is Donsker and thus $J_n(\cdot) \Rightarrow \mathbb{G}(\cdot)$ in $l^\infty(\mathcal{I})$. □

B.4 Proof of Theorem 2.3.8

To handle the term $\nu_n^p(t, \hat{q}_1; \tau) - \nu_n^p(t; q_1; \tau)$ in (2.3.2), I prove that $\nu_n^p(\cdot, \cdot; \tau)$ is stochastically equicontinuous. The following lemma is used to verify one of conditions of Theorem 2.11.9 in van der Vaart and Wellner (1996).

Lemma B.4.1. *Let (Θ, ρ) be a pseudo-metric space with $\Theta \equiv \Psi \times \mathcal{F}$ and $\rho(\theta_1, \theta_2) \equiv ||\psi_1 - \psi_2||_2 + ||q_1 - q_2||_{\mathcal{F}}$ for some function space $\mathcal{F}$ equipped with a norm $||\cdot||_\infty$ (that is, $||\cdot||_{\mathcal{F}} = ||\cdot||_\infty$). If Assumption 2.3.6 is satisfied and $(\mathcal{F}, ||\cdot||_{\mathcal{F}})$ is totally bounded, then (Θ, ρ) is totally bounded.*

Proof. For a pseudo-metric space $(\mathcal{M}, \rho_{\mathcal{M}})$, denote the open ball of radius $\kappa > 0$, centered at $m \in \mathcal{M}$, by $B_\kappa(m)$. To show that the pseudo-metric space $(\Psi, ||\cdot||_2)$ is totally bounded, take any $\epsilon > 0$. Let $C \equiv \mathbb{E}[G_\Psi(X_i)^2]$, where $G_\Psi(\cdot)$ is defined in Assumption 2.3.6, and $\delta \equiv \frac{\epsilon}{\sqrt{C}} > 0$. Since $\mathcal{I}$ is assumed to be compact in $\mathbb{R}^{d_x}$, it is compact and thus totally bounded. That is, there exists a set $\{t_i \in \mathcal{X} : i = 1, 2, ..., N(\delta)\}$ such that $\mathcal{X} \subseteq \cup_{i=1}^{N(\delta)} B_\delta(t_i)$. It suffices to show that $\Psi \subseteq \cup_{i=1}^{N(\delta)} B_\epsilon(\psi(X, t_i))$. Let $h(X)$ be an arbitrary element of Ψ (i.e. $h(X) = \psi(X, t)$ for some $t \in \mathcal{I}$). Since $\mathcal{I}$ is totally bounded, there exists $t_{i_0} \in \{t_i \in \mathcal{X} : i = 1, 2, ..., N(\delta)\}$ such that $||t - t_{i_0}||_E < \delta$. This implies that

$$||h(X) - \psi(X, t_{i_0})||_2^2 \leq \mathbb{E}[G(X)^2]||t - t_{i_0}||_E^2 \leq C\delta^2 = \epsilon^2$$

and thus that $\Psi \subseteq \cup_{i=1}^{N(\delta)} B_\epsilon(\psi(X, t_i))$. Since $N(\delta) < \infty$ and ϵ was arbitrary, $(\Psi, ||\cdot||_2)$ is totally bounded. Since it is assumed that $(\mathcal{F}, ||\cdot||_{\mathcal{F}})$ is totally bounded, the product of Ψ and $\mathcal{F}$ is totally bounded with respect to ρ. □

For the simplicity of notation, I abbreviate a generic $(\bar{\psi}, q) \in \bar{\Psi} \times \Lambda_R^{p_1}(\mathcal{X}_1)$ to $\theta \in \Theta$, where $\Theta \equiv \bar{\Psi} \times \Lambda_R^{p_1}(\mathcal{X}_1)$.

Lemma B.4.2. *Suppose that Assumptions 2.3.1, 2.3.2, 2.3.3, and 2.3.6 hold. Then, $\nu_n^p(\cdot, \cdot; \tau)$ is stochastically equicontinuous with respect to the semi-norm $\rho(\theta, \tilde{\theta}) \equiv ||\bar{\psi} - \tilde{\bar{\psi}}||_2 + ||q - \tilde{q}||_\infty$.*

Proof. Define

$$Z_{ni}(\psi, q) \equiv \frac{1}{\sqrt{n}} D_i \bar{\psi}(X_{1i}, t)\{\mathbf{1}(Y_i \leq q_i) - F_{T|X_1}(q_i|X_{1i})\}$$

and let $\mathcal{F} \equiv \{Z_{ni}(\bar{\psi}, q) : (\bar{\psi}, q) \in \bar{\Psi} \times \Lambda_R^{p_1}(\mathcal{X}_1)\}$. Noting that q is a function of only X_1, it is clear that $\mathbb{E}Z_{ni}(\bar{\psi}, q) = 0$ for any $(\bar{\psi}, q) \in \Theta$. I verify the conditions of Theorem 2.11.9 in van der Vaart and Wellner (1996). Let $||Z_{ni}||_{\mathcal{F}} \equiv \sup_{(\bar{\psi}, q) \in \bar{\Psi} \times \Lambda_R^{p_1}(\mathcal{X}_1)} |Z_{ni}(\bar{\psi}, q)|$ and $\rho((\bar{\psi}, q), (\tilde{\bar{\psi}}, \tilde{q})) \equiv ||\bar{\psi} - \tilde{\bar{\psi}}||_2 + ||q - \tilde{q}||_\infty$. Since $D_i\{\mathbf{1}(Y_i \leq q_i) - F_{T|X_1}(q_i|X_{1i})\}\bar{\psi}(X_{1i}, t)$ is uniformly bounded, one obtains that for any $\eta > 0$,

$$\begin{aligned}
&\mathbb{E}[||Z_{ni}||_{\mathcal{F}} \mathbf{1}(||Z_{ni}||_{\mathcal{F}} > \eta)] \\
\leq &\frac{C}{\sqrt{n}} \Pr(||Z_{ni}||_{\mathcal{F}} > \eta) \\
\leq &\frac{C}{\sqrt{n}} \frac{\mathbb{E}||D_i\{\mathbf{1}(Y_i \leq q_i) - F_{T|X_1}(q_i|X_{1i})\}\bar{\psi}(X_{1i}, t)||_{\mathcal{F}}}{n\eta^2} \\
\lesssim &\frac{1}{n\sqrt{n}\eta} = o(n).
\end{aligned} \tag{B.4.1}$$

This implies that $\sum_i \mathbb{E}[||Z_{ni}||_{\mathcal{F}} \mathbf{1}(||Z_{ni}||_{\mathcal{F}} > \eta)] = o(1)$ and thus that the first condition of the theorem is met.

Take any $(\bar{\psi}, q) \in \bar{\Psi} \times \Lambda_R^{p_1}(\mathcal{X}_1)$ and $\eta > 0$. Note that for any $\tilde{\theta} = (\tilde{\psi}, \tilde{q})$ such that $\rho(\theta, \tilde{\theta}) \leq \eta$,

$$
\begin{aligned}
&\mathbb{E}[Z_{ni}(\bar{\psi}, q) - Z_{ni}(\tilde{\psi}, \tilde{q})]^2 \\
\lesssim &\frac{1}{n}\mathbb{E}[D_i\{\mathbf{1}(Y_i \leq q_i) - F_{T|X_1}(q_i|X_{1i})\}\bar{\psi}(X_{1i}, t) \\
&- D_i\{\mathbf{1}(Y_i \leq \tilde{q}_i) - F_{T|X_1}(\tilde{q}_i|X_{1i})\}\bar{\psi}(X_{1i}, \tilde{t})]^2 \\
\lesssim &\frac{1}{n}\mathbb{E}[p_0(X_i)\{\mathbf{1}(T_i \leq q_i) - \mathbf{1}(T_i \leq \tilde{q}_i) - (F_{T|X_1}(q_i|X_{1i}) - F_{T|X_1}(\tilde{q}_i|X_{1i}))\}^2] \\
&+ \frac{1}{n}\mathbb{E}[\bar{\psi}(X_{1i}, t) - \bar{\psi}(X_{1i}, \tilde{t})]^2 \\
\lesssim &\frac{1}{n}\mathbb{E}[|\mathbf{1}(T_i \leq q_i) - \mathbf{1}(T_i \leq \tilde{q}_i)|] + \frac{1}{n}\mathbb{E}[(F_{T|X_1}(q_i|X_{1i}) - F_{T|X_1}(\tilde{q}_i|X_{1i}))^2] + \frac{1}{n}||\bar{\psi} - \tilde{\bar{\psi}}||_2^2.
\end{aligned} \tag{B.4.2}
$$

Using the argument of (Chen et al., 2003, p.1600), it can be shown that

$$
\begin{aligned}
&\sup_{||q-\tilde{q}|| \leq \eta} \frac{1}{n}\mathbb{E}[|\mathbf{1}(T_i \leq q_i) - \mathbf{1}(T_i \leq \tilde{q}_i)|] \\
\leq &\frac{1}{n}\mathbb{E}[\mathbf{1}(T_i \leq q_i + \eta) - \mathbf{1}(T_i \leq q_i - \eta)] \\
\leq &\frac{1}{n}\mathbb{E}[F_{T|X_1}(q_i + \eta|X_{1i}) - F_{T|X_1}(q_i - \eta|X_{1i})] \\
\leq &\frac{1}{n}\sup_{t \in \mathbb{R}, x_1 \in \mathcal{X}_1} |f_{T|X_1}(t|x_1)| \cdot 2\eta \lesssim \frac{1}{n}\eta
\end{aligned} \tag{B.4.3}
$$

by Assumption 2.3.3. By the same way, it is straightforward to show that, under Assumption 2.3.3,

$$
\sup_{||q-\tilde{q}|| \leq \eta} \frac{1}{n}\mathbb{E}[(F_{T|X_1}(q_i|X_{1i}) - F_{T|X_1}(\tilde{q}_i|X_{1i}))^2] \lesssim \frac{1}{n}\eta^2. \tag{B.4.4}
$$

Under Assumption 2.3.6, one can show that

$$\begin{aligned}||\bar{\psi}(X_{1i},t_1)-\bar{\psi}(X_{1i},t_2)||_2^2 &= \mathbb{E}[|\bar{\psi}(X_{1i},t_1)-\bar{\psi}(X_{1i},t_2)|]^2 \\ &\leq \mathbb{E}[\mathbb{E}[|\psi(X_i,t_1)-\psi(X_i,t_2)||X_{1i}]^2] \\ &\leq \mathbb{E}[\mathbb{E}[G_\Psi(X_i)\cdot||t_1-t_2||_E|X_{1i}]^2] \\ &\leq ||t_1-t_2||_E^2\mathbb{E}[\bar{G}_\Psi^2(X_{1i})], \end{aligned} \tag{B.4.5}$$

where $\bar{G}_\Psi(X_{1i}) \equiv \mathbb{E}[G_\Psi(X_i)|X_{1i}]$. By Jensen's inequality, $\bar{G}_\Psi$ is square-integrable as G_Ψ is square-integrable. In all, combining (B.4.2), (B.4.3) and (B.4.4)together yields that

$$n\cdot \sup_{\rho(\theta,\tilde{\theta})\leq\alpha_n} \mathbb{E}[Z_{ni}(\psi,q)-Z_{ni}(\tilde{\psi},\tilde{q})]^2 \lesssim \alpha_n+\alpha_n^2 = o(1) \tag{B.4.6}$$

for any positive sequence α_n such that $\alpha_n \downarrow 0$, and thus the second condition of the theorem is also satisfied.

Lastly, I calculate the bracketing L^2-entropy of $\Theta = \bar{\Psi}\times\Lambda_R^{p_1}(\mathcal{X}_1)$. Take any $\epsilon>0$. Since $N(\epsilon,\Theta,||\cdot||_2)\leq N_{[]}(\epsilon,\Theta,||\cdot||_2)$, it suffices to calculate the L^2-bracketing number. By the definition of the bracketing number, one can show that

$$N_{[]}(\epsilon,^-\times\Lambda_R^{p_1}(\mathcal{X}_1),||\cdot||_2)\leq N_{[]}(\frac{\epsilon}{2},\bar{\Psi},||\cdot||_2)\cdot N_{[]}(\frac{\epsilon}{2},\Lambda_R^{p_1}(\mathcal{X}_1),||\cdot||_2). \tag{B.4.7}$$

Since the weighting function $\bar{\psi}(x,t)$ is Lipschitz in t as shown in (B.4.5), it follows that $N_{[]}(\frac{\epsilon}{2},\bar{\Psi},||\cdot||_2)\leq N(\frac{\epsilon}{4||G_\Psi||_2},\mathcal{I},||\cdot||_E)$ by Theorem 2.7.11 in van der Vaart and Wellner (1996). Since $\mathcal{I}$ is assumed to be compact in $\mathbb{R}^{d_x}$, the covering number is

bounded by $C \cdot \epsilon^{-d_x}$ for some constant $C > 0$. By Corollary 2.7.2 in van der Vaart and Wellner (1996) with Assumption 2.3.2, the L^2-bracketing number, one obtains that

$$\log N_{[]}(\frac{\epsilon}{2}, \Lambda_R^{p_1}(\mathcal{X}_1), ||\cdot||_2) \lesssim \epsilon^{-\frac{d_1}{p_1}},$$

and this implies that

$$\log N_{[]}(\epsilon, \Theta, ||\cdot||_2) \leq \log N_{[]}(\frac{\epsilon}{2}, \Psi, ||\cdot||_2) + \log N_{[]}(\frac{\epsilon}{2}, \Lambda_R^{p_1}(\mathcal{X}_1), ||\cdot||_\infty) \lesssim \epsilon^{-\frac{d_1}{p_1}}.$$

Therefore, for any positive sequence $\{\alpha_n\}$ such that $\alpha_n \downarrow 0$,

$$\int_0^{\alpha_n} \sqrt{\log N_{[]}(\epsilon, \Theta, ||\cdot||_2)} d\epsilon \lesssim \alpha_n^{1-\frac{d_1}{2p_1}} = o(1) \tag{B.4.8}$$

by Assumption 2.3.2.

Note that a Hölder ball is compact under the sup-norm by the embedding theorem (e.g. Theorems 1 and 2 in Freyberger and Masten (2015)), and thus $(\Lambda_R^{p_1}(\mathcal{X}_1), ||\cdot||_\infty)$ is totally bounded. Thus, (Θ, ρ) is a totally bounded pseudo-metric space by lemma B.4.1 and equations (B.4.1), (B.4.6), and (B.4.8) together establish that all conditions of Theorem 2.11.9 in van der Vaart and Wellner (1996) are satisfied. In all, $\nu_n^p(\cdot, \cdot; \tau)$ is asymptotically tight. Since $\mathbb{E}[n \cdot Z_{ni}(\theta)^2] < \infty$ for any $\theta \in \Theta$, all finite-dimensional marginals of $\nu_n^p(\cdot, \cdot; \tau)$ converge in distribution to a multivariate normal distribution by the multivariate central limit theorem. Since $\nu_n^p(\cdot, \cdot; \tau)$ is asymptotically tight and all of its finite-dimensional marginals converge in distribution to a random vector, $\nu_n^p(\cdot, \cdot; \tau)$ converges weakly to a tight limit in $l^\infty(\Theta)$ and this leads to that $\nu_n^p(\cdot, \cdot; \tau)$ is stochastically equicontinuous (e.g. (Andrews, 1994b,

p.2251)). □

Let $q_{1i} \equiv Q_{T|X_1}(\tau|X_{1i})$ and $\hat{q}_{1i} \equiv \hat{Q}_{T|X_1}(\tau|X_{1i})$ and recall the smoothed term

$$\begin{aligned}\hat{J}_{sn}(t;G) &= \frac{1}{\sqrt{n}}\sum_i D_i\{F_{T|X_1}(\hat{q}_{1i}|X_{1i}) - F_{T|X_1}(q_{1i}|X_{1i})\}\bar{\psi}(X_{1i},t) \\ &= \frac{1}{\sqrt{n}}\sum_i D_i f_{T|X_1}(q_{1i}|X_{1i})(\hat{q}_{1i}-q_{1i})\bar{\psi}(X_{1i},t) + O_p(||\hat{q}_1 - q_1||_\infty^2).\end{aligned}$$

By the construction of the conditional quantile function, the leading term can be rewritten as

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_i D_i f_{T|X_1}(q_{1i}|X_{1i})(\hat{q}_{1i}-q_{1i})\bar{\psi}(X_{1i},t) \\ =&\frac{1}{\sqrt{n}}\sum_i D_i f_{T|X_1}(q_{1i}|X_{1i})\{\hat{F}_{1n}^{-1}(\tau|X_{1i}) - F_1^{-1}(\tau|X_{1i})\}\bar{\psi}(X_{1i},t).\end{aligned}$$

Since the inverse map is Hadamard differentiable[1], one can show that for $F(\cdot|\cdot)$ such that $||F - F_{T|X_1}||_\infty$ is small,

$$\begin{aligned}F^{-1}(\tau|x_1) - F_{T|X_1}^{-1}(\tau|x_1) = &\frac{F_{T|X_1}(Q_{T|X_1}(\tau|x_1)|x_1) - F(Q_{T|X_1}(\tau|x_1)|x_1)}{f_{T|X_1}(Q_{T|X_1}(\tau|x_1)|x_1)} \\ &+ O(||F(\cdot|\cdot) - F_{T|X_1}(\cdot|\cdot)||_\infty^2)\end{aligned}$$

by, for example, Van der Vaart (1998); Kong and Xia (2017). Since $||\hat{F}_{1n} - F_{T|X_1}||_\infty = o_p(1)$ and $||F(\cdot|\cdot) - F_{T|X_1}(\cdot|\cdot)||_\infty^2 = O(\frac{\log n}{n h_{Fn}^{d_1}}) = o(n^{-1/2})$ by Lemma 1 in Kong and

[1] One can refer to, for example, van der Vaart and Wellner (1996); Van der Vaart (1998); Kosorok (2008) for the definition of Hadamard differentiability.

Xia (2017), it can be shown that for large enough n,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_i D_i f_{T|X_1}(q_{1i}|X_{1i})\{\hat{F}_{1n}^{-1}(\tau|X_{1i}) - F_{T|X_1}^{-1}(\tau|X_{1i})\}\bar{\psi}(X_{1i},t) \\
=&-\frac{1}{\sqrt{n}}\sum_i D_i\bar{\psi}(X_{1i},t)\{\hat{F}_{1n}(q_{1i}|X_{1i}) - F_{T|X_1}(q_{1i}|X_{1i})\} + \sqrt{n}O_p(\frac{\log n}{nh_{Fn}^{d_1}}), \quad \text{(B.4.9)}
\end{aligned}$$

where the equality comes from Lemma 1 in Kong and Xia (2017). Note that under the conditions on the bandwidth h_{Fn}, the remainder term is $o_p(1)$. Therefore, one needs to investigate the leading term in (B.4.9) and the following lemma establishes that the leading term admits an asymptotic linear representation.

Lemma B.4.3. *Suppose that Assumptions 2.3.1 and 2.3.3 through 2.3.6 hold. Then,*

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_i D_i\bar{\psi}(X_{1i},t)\{\hat{F}_{1n}(q_{1i}|X_{1i}) - F_{T|X_1}(q_{1i}|X_{1i})\} \\
=&\frac{1}{\sqrt{n}}\sum_i(1-\tau)\tilde{\psi}(X_{1i},t)\xi(Y_i,D_i,q_{1i},X_{1i}) + o_p(1),
\end{aligned}$$

where

$$\begin{aligned}
\xi(Y_j,D_j,y,x) \equiv [&\frac{\mathbf{1}(Y_j\le y)\cdot D_j}{(1-F_{T|X_1}(Y_j|x))(1-G(Y_j|x))} \\
&-\int_0^{\min(Y_j,y)}\frac{f_1(s|x)ds}{(1-F_{T|X_1}(s|x))^2(1-G(s|x))}], \\
\tilde{\psi}(X_{1i},t) \equiv &\mathbb{E}[\bar{\psi}(X_{1i},t)p_0(X_i)|X_{1i}],
\end{aligned}$$

uniformly in $t\in\mathcal{I}$.

Proof. Using the asymptotic representation given in Lemma 1 in Kong and Xia

(2017), it is straightforward to see that

$$\hat{F}_{1n}(y|x) - F_{T|X_1}(y|x) = \frac{1}{nh_{Fn}^{d_1}}(1 - F_{T|X_1}(y|x)) \sum_j \tilde{B}_{h_{Fn}}(X_{1j}; x)\xi(Y_j, D_j, y, x) + O((\frac{\log n}{nh_{Fn}^{d_1}})^{3/4}), \tag{B.4.10}$$

where

$$\tilde{B}_{h_{Fn}}(X_{1j}; x) = e_1' \Omega_1^{-1} \mu(\frac{X_{1j} - x}{h_{Fn}}) K_F(\frac{X_{1j} - x}{h_{Fn}}) / f_{X_1}(x).$$

Plugging (B.4.10) into (B.4.9) yields that uniformly y and $x \in \mathcal{X}_1$,

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_i D_i \bar{\psi}(X_{1i}, t)\{\hat{F}_{1n}(y|x) - F_{T|X_1}(y|x)\} \\
=&\frac{(1 - F_{T|X_1}(y|x))}{n\sqrt{n}} \sum_i \sum_j D_i \bar{\psi}(X_{1i}, t) \frac{1}{h_{Fn}^{d_1}} \tilde{B}_{h_{Fn}}(X_{1j}; x)\xi(Y_j, D_j, y, x) \\
&+ \sqrt{n} O_p((\frac{\log n}{nh_{Fn}^{d_1}})^{3/4}),
\end{aligned}$$

where the remainder term is $o_p(1)$ under the conditions on the bandwidth h_{Fn}. To analyze the leading term in the above equation, I utilize the theory of U-processes. Let

$$p_{Fn}^{\dagger}(W_i, W_j; y, t) \equiv D_i \bar{\psi}(X_{1i}, t) \frac{1}{h_{Fn}^{d_1}} \tilde{B}_{h_{Fn}}(X_{1j}; X_{1i})\xi(Y_j, D_j, y, X_{1i}).$$

To make the U-statistic kernel $p_{Fn}^{\dagger}(\cdot, \cdot; t)$ symmetric, define $p_{Fn}(W_i, W_j; y, t) \equiv$

$\frac{1}{2}\{p^{\dagger}_{Fn}(W_i, W_j; y, t) + p^{\dagger}_{Fn}(W_j, W_i; y, t)\}$. Then one obtains that

$$\begin{aligned}
&\frac{1}{n\sqrt{n}} \sum_i \sum_j D_i \bar{\psi}(X_{1i}, t) \frac{1}{h_{Fn}^{d_1}} \tilde{B}_{h_{Fn}}(X_{1j}; X_{1i}) \xi(Y_j, D_j, y, X_{1i}) \\
=&\frac{1}{n\sqrt{n}} \sum_i \sum_{j \neq i} p^{\dagger}_{Fn}(W_i, W_j; y, t) + \frac{1}{n\sqrt{n}} \sum_i p^{\dagger}_{Fn}(W_i, W_i; y, t) \\
=&\sqrt{n} \frac{n-1}{n} \binom{n}{2}^{-1} \sum_i \sum_{j>i} p_{Fn}(W_i, W_j; y, t) + \frac{1}{n\sqrt{n}} \sum_i p^{\dagger}_{Fn}(W_i, W_i; y, t).
\end{aligned}$$

By a similar reasoning of (Kong et al., 2013, p.966), $\frac{1}{n\sqrt{n}} \sum_i p^{\dagger}_{Fn}(W_i, W_i; y, t) = o_p(1)$. On the other hand, the leading term can be analyzed by using the theory of U-processes. To keep the simplicity of notations, let $p^{\dagger}_{Fn}(W_i, W_j; t) \equiv p^{\dagger}_{Fn}(W_i, W_j; q_{1i}, t)$ and $p_{Fn}(W_i, W_j; t) \equiv p_{Fn}(W_i, W_j; q_{1i}, t)$.

Let $\mathcal{U}_n(y, t) \equiv \binom{n}{2}^{-1} \sum_i \sum_{j>i} p_{Fn}(W_i, W_j; y, t)$ and consequently denote $\mathcal{U}_n(q_{1i}, t)$ by $\mathcal{U}_n(t)$. I also define the following objects:

$$\begin{aligned}
r_{Fn}(W_i; y, t) &\equiv \mathbb{E}[p_{Fn}(W_i, W_j; y, t) | W_i], \\
\theta_{Fn}(y, t) &\equiv \mathbb{E}[p_{Fn}(W_i, W_j; y, t)], \\
\tilde{p}_{Fn}(W_i, W_j; y, t) &\equiv p_{Fn}(W_i, W_j; y, t) - r_{Fn}(W_i; y, t) - r_{Fn}(W_j; y, t) + \theta_{Fn}(y, t),
\end{aligned}$$

and $r_{Fn}(W_i; t)$, $\theta_{Fn}(t)$, and $\tilde{p}_{Fn}(W_i, W_j; t)$ are defined by the same way above. Then, applying Hoeffding's decomposition to $\mathcal{U}_n(y, t)$ yields that

$$\begin{aligned}
\mathcal{U}_n(y, t) &= \theta_{Fn}(y, t) + \frac{2}{n} \sum_i r_{Fn}(W_i; y, t) + \binom{n}{2}^{-1} \sum_i \sum_{j>i} \tilde{p}_{Fn}(W_i, W_j; y, t) \\
&\equiv \theta_{Fn}(y, t) + \frac{2}{n} \sum_i r_{Fn}(W_i; y, t) + R_{Fn}(y, t)
\end{aligned}$$

First, I show that $R_{Fn}(t) = o_p(1)$ where $R_{Fn}(t) \equiv R_{Fn}(q_{1i}, t)$. It is obvious that

$\mathbb{E}R_{Fn}(t) = 0$ for any $t \in \mathcal{I}$. To calculate the variance of $R_{Fn}(t)$ for given $t \in \mathcal{I}$, I refer to Powell et al. (1989). Note that

$$Var(R_{Fn}(t)) \leq \binom{n}{2}^{-2} \sum_i \sum_{j>i} \mathbb{E}\tilde{p}_{Fn}(W_i, W_j; t)^2 = \binom{n}{2}^{-2} O(n^2) \mathbb{E}\tilde{p}_{Fn}(W_i, W_j; t)^2.$$

Moreover, one can further show that

$$\begin{aligned}
\mathbb{E}\tilde{p}_{Fn}(W_i, W_j; t)^2 &= \mathbb{E}[p_{Fn}(W_i, W_j; t) - r_{Fn}(W_i, t) - r_{Fn}(W_j, t) + \theta_{Fn}(t)]^2 \\
&\lesssim \mathbb{E}[p_{Fn}(W_i, W_j; t) - \theta_{Fn}(t)]^2 + \mathbb{E}[r_{Fn}(W_i; t) - \theta_{Fn}(t)]^2 \\
&\leq 2\mathbb{E}[p_{Fn}(W_i, W_j; t) - \theta_{Fn}(t)]^2 \\
&= 2Var(p_{Fn}(W_i, W_j; t)) \leq 2\mathbb{E}p_{Fn}(W_i, W_j; t)^2
\end{aligned}$$

where the inequality on the third line holds by a property of U-statistic, given by (Serfling, 1980, p.182). Thus,

$$Var(\sqrt{n}R_{Fn}(t)) \leq n \cdot \binom{n}{2}^{-2} O(n^2) \cdot \mathbb{E}p_{Fn}(W_i, W_j; t)^2 = O(n^{-1}) \mathbb{E}p_{Fn}(W_i, W_j; t)^2$$

and it will suffice to show that $\mathbb{E}p_{Fn}(W_i, W_j; t)^2 = o(n)$ to prove that $R_{Fn}(t) = o_p(n^{-1/2})$ for given $t \in \mathcal{I}$. Note that

$$\mathbb{E}p_{Fn}(W_i, W_j; t)^2 = \frac{1}{4}\mathbb{E}[p^\dagger_{Fn}(W_i, W_j; t) + p^\dagger_{Fn}(W_j, W_i; t)]^2 \leq \mathbb{E}p^\dagger_{Fn}(W_i, W_j; t)^2$$

and that

$$
\begin{aligned}
&\mathbb{E}p^{\dagger}_{Fn}(W_i, W_j; t)^2 \\
=&\mathbb{E}[D_i\bar{\psi}(X_{1i}, t)\frac{1}{h^{d_1}_{Fn}}\tilde{B}_{h_{Fn}}(X_{1j}; X_{1i})\xi(Y_j, D_j, q_{1i}, X_{1i})]^2 \\
\lesssim&\frac{1}{h^{d_1}_{Fn}}\mathbb{E}[\frac{1}{h^{d_1}_{Fn}}\{e_1^{'}\Omega_1^{-1}\mu(X_{1j} - X_{1i}; h_{Fn})K_F(\frac{X_{1j} - X_{1i}}{h_{Fn}})\frac{1}{f_{X_1}(X_{1i})}\xi(Y_j, D_j, q_{1i}, X_{1i})\}^2] \\
\leq&\frac{1}{h^{d_1}_{Fn}}\mathbb{E}[\frac{1}{h^{d_1}_{Fn}}||e_1||^2||\Omega_1^{-1}\mu(X_{1j} - X_{1i}; h_{Fn})||^2K_F^2(\frac{X_{1j} - X_{1i}}{h_{Fn}})\{\frac{1}{f_{X_1}(X_{1i})}\xi(Y_j, D_j, q_{1i}, X_{1i})\}^2] \\
\lesssim&\frac{1}{h^{d_1}_{Fn}}\mathbb{E}[\frac{1}{h^{d_1}_{Fn}}K_F(\frac{X_{1j} - X_{1i}}{h_{Fn}})\xi(Y_j, D_j, q_{1i}, X_{1i})^2] \\
=&\frac{1}{h^{d_1}_{Fn}}\mathbb{E}[\frac{1}{h^{d_1}_{Fn}}K_F(\frac{X_{1j} - X_{1i}}{h_{Fn}})\mathbb{E}[\xi(Y_j, D_j, q_{1i}, X_{1i})^2|X_{1i}, X_{1j}]] \\
\leq&O(h^{-d_1}_{Fn})
\end{aligned}
$$

since $\mathbb{E}[\xi(Y_j, D_j, q_{1i}, X_{1i})^2] < \infty$. Hence, $\mathbb{E}p^{\dagger}_{Fn}(W_i, W_j; t)^2 = o(n)$ as $nh^{d_1}_{Fn} \to \infty$ and this implies that $Var(\sqrt{n}R_{Fn}(t)) = o(1)$.

To obtain the uniform convergence, I adopt the approach used in the proof of Lemma 3 in Huang et al. (2016). Recall that the weighting function is of the form of $\psi(X_i, t) = \mathbf{w}(X_i^{'}t)$ where $\mathbf{w}(\cdot)$ is analytical, and hence Ψ is Vapnik-Červonenkis (VC)-type. Let $\mathcal{P} \equiv \{p(W_i, W_j; t) : t \in \mathcal{I}\}$, then $\mathcal{P} = \Psi(t)\cdot\{\mathcal{K}(W_i, W_j) + \mathcal{K}(W_j, W_i)\}$, where $\mathcal{K}(W_i, W_j) \equiv \frac{1}{h^{d_1}_{Fn}}D_i\tilde{B}_{h_{Fn}}(X_{1j}; X_{1i})\xi(Y_j, D_j, q_{1i}, X_{1i})$. Since $\mathcal{K}$ is a bounded function, $\mathcal{P}$ is again VC-type with a square-integrable envelope by Lemma 2.6.18 in van der Vaart and Wellner (1996). Note that, by the standard arguments in the local polynomial regression, $\mathbb{E}\mathcal{K}^2(W_i, W_j) = O(h^{-d_1}_{Fn})$ under the conditions imposed

in this lemma and thus one obtains

$$\mathbb{E}\sup_{t\in\mathcal{J}}|\frac{n(n-1)}{n}R_n(t)|^2 \lesssim \mathbb{E}\mathcal{K}^2(W_i,W_j) = O(h_{Fn}^{-d_1})$$

by Proposition 4 in Delgado and Manteiga (2001). Thus, $\mathbb{E}\sup_{t\in\mathcal{J}}|\sqrt{n}R_{Fn}(t)|^2 = O((nh_{Fn}^{d_1})^{-1}) = o(1)$ and this leads to that $\sup_{t\in\mathcal{J}}|\sqrt{n}R_{Fn}(t)| = o_p(1)$.

Now, I consider the projected term $\theta_{Fn}(t) + \frac{2}{n}\sum_i r_{Fn}(W_i;t)$. Recall that

$$r_{Fn}(W_i;t) = \mathbb{E}[p_{Fn}(W_i,W_j;t)|W_i],$$

then one can show that

$$\begin{aligned}
2r_{Fn}(W_i;y,t) &= \mathbb{E}[p_{Fn}^{\dagger}(W_j,W_i;y,t)|W_i] \\
&= \mathbb{E}[D_j\bar{\psi}(X_{1j},t)\frac{1}{h_{Fn}^{d_1}}\tilde{B}_{h_{Fn}}(X_{1j};X_{1i})\xi(Y_i,D_i,y,X_{1j})|W_i] \\
&= \mathbb{E}[\mathbb{E}[D_j\bar{\psi}(X_{1j},t)\frac{1}{h_{Fn}^{d_1}}\tilde{B}_{h_{Fn}}(X_{1j};X_{1i})\xi(Y_i,D_i,y,X_{1j})|X_j,W_i]|W_i] \\
&= \mathbb{E}[p_0(X_j)\bar{\psi}(X_{1j},t)\frac{1}{h_{Fn}^{d_1}}\tilde{B}_{h_{Fn}}(X_{1j};X_{1i})\xi(Y_i,D_i,y,X_{1j})|W_i] \\
&= \mathbb{E}[\tilde{\psi}(X_{1j},t)\frac{1}{h_{Fn}^{d_1}}\tilde{B}_{h_{Fn}}(X_{1j};X_{1i})\xi(Y_i,D_i,y,X_{1j})|W_i],
\end{aligned}$$

where $\tilde{\psi}(X_{1j},t) = \mathbb{E}[p_0(X_j)\bar{\psi}(X_{1j},t)|X_{1j}]$. Since $\xi(Y_i,D_i,y,X_{1j})$ is continuously differentiable with respect to X_{1j}, applying change of variables and the standard arguments of the local polynomial regression (e.g. (Fan and Gijbels, 1996, p.64))

yields that

$$
\begin{aligned}
2r_{Fn}(W_i;y,t) &= \mathbb{E}[\tilde{\psi}(X_{1j},t)\frac{1}{h_{Fn}^{d_1}}\tilde{B}_{h_{Fn}}(X_{1j};X_{1i})\xi(Y_i,D_i,y,X_{1j})|W_i] \\
&= \int \tilde{\psi}(x_1,t)\xi(Y_i,D_i,y,x_1)\frac{1}{h_{Fn}^{d_1}}e_1'\Omega_1^{-1}\mu(\frac{x_1-X_{1i}}{h_{Fn}})K_F(\frac{x_1-X_{1i}}{h_{Fn}})dx_1 \\
&= \tilde{\psi}(X_{1i},t)\xi(Y_i,D_i,y,X_{1i}) + o_p(n^{-1/2}), \qquad \text{(B.4.11)}
\end{aligned}
$$

where $o_p(n^{-1/2})$ holds uniformly in y and t. This also implies that $\theta_{Fn}(t) = o_p(n^{-1/2})$ uniformly in t. In all, one can obtain

$$
\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_i D_i\psi(X_i,t)\{\hat{F}_{1n}^S(q_{1i}|X_{1i}) - F_1(q_{1i}|X_{1i})\} \\
=&\frac{1}{\sqrt{n}}\sum_i(1-\tau)\tilde{\psi}(X_{1i},t)\xi(Y_i,D_i,q_{1i},X_{1i}) + o_p(1)
\end{aligned}
$$

as $(1 - F_{T|X_1}(q_{1i}|X_{1i})) = 1 - \tau$, and this completes the proof. □

Proof of theorem 2.3.8

Proof. Recall that

$$
\hat{J}_n(t;\tau) = J_n(t;\tau) + \nu_n^p(t,\hat{q}_1;\tau) - \nu_n^p(t;q_1;\tau) + \hat{J}_{sn}(t;\tau).
$$

By lemma B.4.2 and the fact that $||\hat{Q}_{T|X_1}(\tau|\cdot) - Q_{T|X_1}(\tau|\cdot)||_\infty = o_p(1)$, one has that $\nu_n^p(t,\hat{q}_1;\tau) - \nu_n^p(t;q_1;\tau) = o_p(1)$. On the other hand, (B.4.9) and lemma B.4.3 together imply that

$$
\hat{J}_{sn}(t;\tau) = -\frac{1}{\sqrt{n}}\sum_i(1-\tau)\tilde{\psi}(X_{1i},t)\xi(Y_i,D_i,q_{1i},X_{1i}) + o_p(1).
$$

Therefore,

$$\hat{J}_n(t;\tau) = \frac{1}{\sqrt{n}}\sum_i [\psi(X_i,t)D_i\{\mathbf{1}(Y_i \leq Q_{T|X_1}(\tau|X_{1i})) - \tau\} \\ - (1-\tau)\tilde{\psi}(X_{1i},t)\xi(Y_i,D_i,q_{1i},X_{1i})] + o_p(1).$$

To finalize the proof, one needs to show that the class of functions

$$\mathcal{M} \equiv \{\psi(X_i,t)D_i\{\mathbf{1}(Y_i \leq Q_{T|X_1}(\tau|X_{1i}))-\tau\}-(1-\tau)\tilde{\psi}(X_{1i},t)\xi(Y_i,D_i,q_{1i},X_{1i}) : t \in \mathcal{I}\}$$

is Donsker. Let $m(W_i,t)$ be a generic element of $\mathcal{M}$. Since ξ is square-integrable and the weighting function, one can easily show that under Assumption 2.3.6 and for any $t_1, t_2 \in \mathcal{I}$,

$$|m(W_i,t_1) - m(W_i,t_2)| \leq B(W_i)||t_1 - t_2||_E$$

for some square-integrable function $B(\cdot)$. Thus, the class of functions $\mathcal{M}$ is a *type-IV* class and thus it satisfies Ossiander's L^2-entropy condition. In all, $\mathcal{M}$ is Donsker by Theorem 3.1 in Ossiander (1987), and thus the process $\hat{J}_n(\cdot;\tau)$ converges weakly to a tight Gaussian process $\hat{\mathbb{G}}(\cdot)$ in $l^\infty(\mathcal{I})$, where $\hat{\mathbb{G}}(\cdot)$ is a Gaussian process with zero mean and covariance kernel $\hat{\Sigma}(t_1,t_2) = \mathbb{E}[m(W_i,t_1)m(W_i,t_2)]$. □

B.5 Proof of Corollary 2.3.9

Proof. Since the test statistics are continuous functionals of the process $\hat{J}_n(\cdot;\tau)$, it can be proven by applying the continuous mapping theorem (e.g. Theorem 1.11.1 in van der Vaart and Wellner (1996)). □

B.6 Proof of Theorem 2.3.11

To prove theorem 2.3.11, I first examine the asymptotic behavior of the feasible process $J_n(\cdot;\tau)$ under the local alternative specification. Under the local alternative (2.3.3), observe that the infeasible process can be written as following:

$$\begin{aligned}
J_n(t;\tau) &= \frac{1}{\sqrt{n}}\sum_i \psi(X_i,t)D_i\{\mathbf{1}(Y_i \le Q_{T|X_1}(\tau|X_{1i})) - \tau\} \\
&= \frac{1}{\sqrt{n}}\sum_i \psi(X_i,t)D_i\{\mathbf{1}(Y_i \le Q_{T|X_1}(\tau|X_{1i})) - \mathbf{1}(Y_i \le Q_{T|X}(\tau|X_i)) \\
&+ \mathbf{1}(Y_i \le Q_{T|X}(\tau|X_i)) - \tau\} \\
&= \frac{1}{\sqrt{n}}\sum_i \psi(X_i,t)D_i\{\mathbf{1}(Y_i \le Q_{T|X}(\tau|X_i)) - \tau\} \\
&+ \nu_n^A(t;Q_{T|X_1},\tau) - \nu_n^A(t;Q_{T|X},\tau) \\
&+ \frac{1}{\sqrt{n}}\sum_i \psi(X_i,t)D_i\{F(Q_{T|X_1}(\tau|X_{1i})|X_i) - F(Q_{T|X}(\tau|X)|X_i)\} \\
&= \frac{1}{\sqrt{n}}\sum_i \psi(X_i,t)D_i\{\mathbf{1}(Y_i \le Q_{T|X}(\tau|X_i)) - \tau\} \\
&- \frac{1}{n}\sum_i \psi(X_i,t)D_i f(Q_{T|X_1}(\tau|X_{1i})|X_i)\tilde{Q}(\tau|X_i) \\
&+ \nu_n^A(t;Q_{T|X_1},\tau) - \nu_n^A(t;Q_{T|X},\tau) + o_p(1),
\end{aligned} \tag{B.6.1}$$

where $\nu_n^A(t;q,\tau) \equiv \frac{1}{\sqrt{n}}\sum_i \psi(X_i,t)D_i\{\mathbf{1}(Y_i \le q_i) - F_{T|X}(q_i|X_i)\}$ is a stochastic process indexed by t and q. If one can show that show that $\nu_n^A(\cdot;\cdot,\tau)$ is stochastically equicontinuous with respect to an appropriately chosen norm, then it implies that $\nu_n^A(t;Q_{T|X_1},\tau) - \nu_n^A(t;Q_{T|X},\tau) = o_p(1)$ uniformly in $t \in \mathcal{T}$. One may think that the Bracketing central limit theorem (Theorem 2.11.9 in van der Vaart and Wellner (1996)) can be applied to prove the stochastic equicontinuity of the process $\nu_n^A(\cdot;\cdot,\tau)$ in a similar way of lemma B.4.2, but it is needed to find another way as the condi-

tional quantile function under the local alternative depends on n. Therefore, I use Theorem 2.11.23 in van der Vaart and Wellner (1996), which generalizes Theorem 2.11.9 in van der Vaart and Wellner (1996).

Lemma B.6.1. *Suppose that Assumptions 2.3.1, 2.3.3, 2.3.6, and 2.3.10 hold. Then under the local alternative given in (2.3.3),*

$$J_n(\cdot;\tau) \Rightarrow \mathbb{G}(\cdot) - R_a(\cdot) \text{ in } l^\infty(\mathcal{I}),$$

where $\mathbb{G}(\cdot)$ is the Gaussian process defined in theorem 2.3.7 and

$$R_a(t) \equiv \mathbb{E}[p_0(X_i)\psi(X_i,t)f(Q_{T|X_1}(\tau|X_{1i})|X_i)\tilde{Q}(\tau|X_i)].$$

Proof. Define $\Theta \equiv \mathcal{I} \times \Lambda_{\tilde{R}}^{p_A}(\mathcal{X})$ and let $\theta_n \equiv (t, q_n) \in \Theta_n$, where $\Theta_n = \Theta$ for all n. I first show that $\nu_n^A(\cdot;\cdot,\tau)$ is stochastically equicontinuous with respect to the norm $\rho(\theta_1,\theta_2) \equiv ||t_1 - t_2||_E + ||q_1 - q_2||_\infty$ by verifying the conditions of Theorem 2.11.23 in van der Vaart and Wellner (1996). Note that the space Θ is totally bounded with respect to the semi-norm ρ as before. Consider the class of functions

$$\mathcal{G}_n^A \equiv \{\psi(X,t)D\{\mathbf{1}(Y \leq q_n) - F_{T|X}(q_n|X)\} : t \in \mathcal{T},\ q_n \in \Lambda_{\tilde{R}}^{p_A}(\mathcal{X}) \text{ for all } n\},$$

which is indexed by t and q_n. Since one can choose the sequence of envelope functions for each n as a constant function, say C_g, one can show that $\mathbb{E}C_g = O(1)$ and that $\mathbb{E}[C_g^2\mathbf{1}(C_g > \eta/\sqrt{n})] = o(1)$ for any $\eta > 0$. Thus the first two conditions of (2.11.21) in van der Vaart and Wellner (1996) obviously hold. Consider the last condition of (2.11.21) in van der Vaart and Wellner (1996). Take any $\epsilon_n \downarrow 0$, then for any

$g_i^A(\theta_{1n}), g_i^A(\theta_{2n}) \in \mathcal{G}_n^A$,

$$
\begin{aligned}
&\sup_{\rho(\theta_{1n},\theta_{2n})\leq\epsilon_n} \mathbb{E}[g_i^A(\theta_{1n}) - g_i^A(\theta_{2n})]^2 \\
\lesssim &\sup_{\rho(\theta_1,\theta_2)\leq\epsilon_n} \mathbb{E}[\psi(X_i,t_1) - \psi(X_i,t_2)]^2 + \sup_{\rho(\theta_1,\theta_2)\leq\epsilon_n} \mathbb{E}[\mathbf{1}(T_i \leq q_{1ni}) - \mathbf{1}(T_i \leq q_{2ni})]^2 \\
&+ \sup_{\rho(\theta_1,\theta_2)\leq\epsilon_n} ||q_1 - q_2||_\infty^2 \\
=&o(1)
\end{aligned}
$$

by the same logic in the proof of lemma B.4.2. Thus, all conditions of (2.11.21) in van der Vaart and Wellner (1996) are met. Lastly, it is required to calculate the L^2-bracketing number of $\mathcal{G}_n^A$. Observe that $\mathcal{G}_n^A = \Psi \cdot \tilde{\mathcal{G}}_n^A$, where $\tilde{\mathcal{G}}_n^A \equiv \{D\{\mathbf{1}(Y \leq q_n) - F_{T|X}(q_n|X)\} : q_n \in \Lambda_{\tilde{R}}^{p_A}(\mathcal{X})\}$. Since both spaces Ψ and $\tilde{\mathcal{G}}_n^A$ are uniformly bounded, Lemma A.1 in Escanciano et al. (2014) implies that

$$N_{[]}(C_g\epsilon, \mathcal{G}_n^A, ||\cdot||_2) \leq N_{[]}(\tilde{C}\epsilon, \Psi, ||\cdot||_2) \cdot N_{[]}(\tilde{C}\epsilon, \tilde{\mathcal{G}}_n^A, ||\cdot||_2)$$

for some $\tilde{C} > 0$. Let $\tilde{\mathcal{G}}_{1n}^A \equiv \{D\mathbf{1}(Y \leq q_n) : q_n \in \Lambda_{\tilde{R}}^{p_A}(\mathcal{X})\}$ and $\tilde{\mathcal{G}}_{2n}^A \equiv \{-DF_{T|X}(q_n|X) : q_n \in \Lambda_{\tilde{R}}^{p_A}(\mathcal{X})\}$. Since $\tilde{\mathcal{G}}_n^A = \tilde{\mathcal{G}}_{1n}^A + \tilde{\mathcal{G}}_{2n}^A$, it is straightforward to see that

$$N_{[]}(\tilde{C}\epsilon, \tilde{\mathcal{G}}_n^A, ||\cdot||_2) \leq N_{[]}(C\epsilon, \tilde{\mathcal{G}}_{1n}^A, ||\cdot||_2) \cdot N_{[]}(C\epsilon, \tilde{\mathcal{G}}_{2n}^A, ||\cdot||_2).$$

Since one can show that

$$
\begin{aligned}
\mathbb{E}[D_i\{\mathbf{1}(Y_i \leq q_{1n}) - \mathbf{1}(Y_i \leq q_{2n})\}]^2 &= \mathbb{E}[p_0(X_i) \cdot |\mathbf{1}(T_i \leq q_{1n}) - \mathbf{1}(T_i \leq q_{2n})|^2] \\
&\leq \mathbb{E}[\mathbf{1}(|T_i - q_{1n}| \leq 2||q_{1n} - q_{2n}||_\infty)] \\
&\lesssim ||q_{1n} - q_{2n}||_\infty,
\end{aligned}
$$

Theorems 2.7.11 and 2.7.1 in van der Vaart and Wellner (1996) together yield that

$$\log N_{[]}(C\epsilon, \tilde{\mathcal{G}}_{1n}^{A}, ||\cdot||_2) \leq \log N(\hat{C}\epsilon, \Lambda_{\tilde{R}}^{p_A}(\mathcal{X}), ||\cdot||) \leq \epsilon^{-\frac{d_x}{p_A}}.$$

On the other hand, the class of functions $\tilde{\mathcal{G}}_{2n}^{A}$ is also Lipschitz in the index q_n under Assumption 2.3.3. By the same way above, one obtains that

$$\log N_{[]}(C\epsilon, \tilde{\mathcal{G}}_{2n}^{A}, ||\cdot||_2) \leq \log N(\hat{C}\epsilon, \Lambda_{\tilde{R}}^{p_A}(\mathcal{X}), ||\cdot||) \leq \epsilon^{-\frac{d_x}{p_A}}.$$

Therefore, it follows that

$$\log N_{[]}(C_g\epsilon, \mathcal{G}_n^{A}, ||\cdot||_2) \lesssim \epsilon^{-\frac{d_x}{p_A}}$$

and that for any $\alpha_n \downarrow 0$,

$$\int_0^{\alpha_n} \sqrt{\log N_{[]}(C_g\epsilon, \mathcal{G}_n^{A}, ||\cdot||_2)}d\epsilon \lesssim \int_0^{\alpha_n} \epsilon^{-\frac{d_x}{2p_A}} d\epsilon \leq C \cdot \alpha_n^{1-\frac{d_x}{2p_A}} = o(1)$$

under Assumption 2.3.10. In all, $\nu_n^A(\cdot;\cdot,\tau)$ is stochastically equicontinuous with respect to the norm ρ. Since $||Q_{T|X}(\tau|\cdot) - Q_{T|X_1}(\tau|\cdot)||_\infty = \frac{1}{\sqrt{n}} \sup_{x\in\mathcal{X}} |\tilde{Q}(\tau|x)| = o(1)$ and $\nu_n^A(\cdot;\cdot,\tau)$ is stochastically equicontinuous, it follows that $\nu_n^A(t; Q_{T|X_1}, \tau) - \nu_n^A(t; Q_{T|X}, \tau) = o_p(1)$ uniformly in $t \in \mathcal{I}$.

Next, one can show that the leading term $\frac{1}{\sqrt{n}} \sum_i \psi(X_i, t) D_i\{\mathbf{1}(Y_i \leq Q_{T|X}(\tau|X_i)) - \tau\}$ converges weakly to the Gaussian process $\mathbb{G}(\cdot)$ defined in theorem 2.3.7 by verifying conditions of Theorem 2.11.23 in van der Vaart and Wellner (1996) and the verification can be accomplished by the same way above. Note that in this case the only indexing variable is t and thus the condition on the bracketing number in that

theorem is easily satisfied. Lastly, note that

$$\begin{aligned}
&\mathbb{E}\frac{1}{n}\sum_i \psi(X_i,t)D_i f(Q_{T|X_1}(\tau|X_{1i})|X_i)\tilde{Q}(\tau|X_i)\\
=&\mathbb{E}[p_0(X_i)\psi(X_i,t)f(Q_{T|X_1}(\tau|X_{1i})|X_i)\tilde{Q}(\tau|X_i)]\\
=&R_a(t)
\end{aligned}$$

and that the class of functions $\{\psi(X_i,t)D_i f(Q_{T|X_1}(\tau|X_{1i})|X_i)\tilde{Q}(\tau|X_i) : t \in \mathcal{I}\}$ is Donsker, so it is Glivenko-Cantelli. In all,

$$\sup_{t\in\mathcal{I}} |\frac{1}{n}\sum_i \psi(X_i,t)D_i f(Q_{T|X_1}(\tau|X_{1i})|X_i)\tilde{Q}(\tau|X_i) - R_a(t)| = o_p(1).$$

Finally, applying the continuous mapping theorem yields that

$$J_n(\cdot;\tau) \Rightarrow \mathbb{G}(\cdot) - R_a(\cdot) \text{ in } l^\infty(\mathcal{I})$$

and this completes the proof. □

Proof of theorem 2.3.11

Proof. Recall that

$$\hat{J}_n(t;\tau) = J_n(t;\tau) + \nu_n^p(t,\hat{q}_1;\tau) - \nu_n^p(t;q_1;\tau) + \hat{J}_{sn}(t;\tau),$$

which is the same to (2.3.2). By lemma B.4.2, $\nu_n^p(t,\hat{q}_1;\tau) - \nu_n^p(t;q_1;\tau) = o_p(1)$ uniformly in t. Likewise, the asymptotic behavior of the smoothed term $\hat{J}_{sn}(t;\tau)$ under the local alternative remains the same to the one under the null hypothesis.

From the proof of lemma B.6.1, one obtains that

$$\hat{J}_n(t;\tau) = \frac{1}{\sqrt{n}}\sum_i [\psi(X_i,t)D_i\{\mathbf{1}(Y_i \leq Q_{T|X}(\tau|X_i)) - \tau\} - \tilde{\psi}(X_{1i},t)\xi(Y_i,D_i,q_{1i},X_{1i})] \\ - \frac{1}{n}\sum_i \psi(X_i,t)D_i f(Q_{T|X_1}(\tau|X_{1i})|X_i)\tilde{Q}(\tau|X_i) + o_p(1).$$

Since the leading term converges weakly to the Gaussian process $\hat{\mathbb{G}}(\cdot)$ in $l^\infty(\mathcal{J})$ and latter term converges in probability to $R_a(t)$, uniformly in t, it follows that

$$\hat{J}_n(\cdot;\tau) \Rightarrow \hat{\mathbb{G}}(\cdot) - R_a(t)$$

in $l^\infty(\mathcal{J})$, and the theorem is established. □

B.7 Proof of Theorem 2.4.1

Proof. I follow the proof of Theorem 2 in Whang (2006a). To prove (i), recall that

$$\hat{KS}_{n,b,i} = \sup_{t\in\mathcal{J}} |\hat{J}_{n,b,i}(t;\tau)|;\ \hat{CM}_{n,b,i} = \int_{\mathcal{J}} |\hat{J}_{n,b,i}(t;\tau)|^2 d\mu(t).$$

Define

$$F_b^{KS}(z) \equiv \Pr(\hat{KS}_{n,b,i} \leq z);\ F_b^{CM}(z) \equiv \Pr(\hat{CM}_{n,b,i} \leq z).$$

Since $\hat{KS}_n$ and $\hat{CM}_n$ are functionals of a Gaussian process with a nonsingular covariance kernel, it suffices to show that for all $z \in \mathbb{R}$,

$$\hat{F}_{n,b}^{KS}(z) - F_b^{KS}(z) \xrightarrow{p} 0;\ \hat{F}_{n,b}^{CM}(z) - F_b^{CM}(z) \xrightarrow{p} 0$$

as the proof of Theorem 2 in Whang (2006a). From now, I only consider the case of the CM statistic since the case of the KS statistic can be proven by a similar way.

Note that

$$\mathbb{E}\hat{F}_{n,b}^{CM}(z) = F_b^{CM}(z). \tag{B.7.1}$$

Let $I_i \equiv \mathbf{1}(\hat{CM}_{n,b,i} \leq z)$ for $i = 1, 2, ..., n-b+1$. Then one can show that

$$\begin{aligned}
Var(\hat{F}_{n,b,}^{CM}(z)) &= Var(\frac{1}{n-b+1}\sum_{i=1}^{n-b+1} I_i) \\
&= (\frac{1}{n-b+1})^2\mathbb{E}[\{\sum_{i=1}^{n-b+1}(I_i - F_b^{CM}(z))\}\{\sum_{j=1}^{n-b+1}(I_j - F_b^{CM}(z))\}] \\
&= (\frac{1}{n-b+1})^2[\sum_{i}^{n-b+1} Var(I_i) + \sum_{i}^{n-b+1}\sum_{j\neq i, |i-j|\leq b} Cov(I_i, I_j)] \\
&\leq O(\frac{1}{n}) + (\frac{1}{n-b+1})^2 \sum_{i}^{n-b+1}\sum_{j\neq i, |i-j|\leq b}\sqrt{Var(I_i)}\sqrt{Var(I_j)} \\
&= O(\frac{1}{n}) + O(\frac{b}{n-b+1}) = O(\frac{1}{n}) + O(\frac{b}{n}) = o(1).
\end{aligned} \tag{B.7.2}$$

Combining (B.7.1) and (B.7.2) yields that $\hat{F}_{n,b}^{CM}(z) - F_b^{CM}(z) \xrightarrow{p} 0$ for all $z \in \mathbb{R}$.

To prove (ii), one can refer to the proof of Corollary 5 in Whang (2006a) and the proof of (i) above. □

Appendix C

Chapter 3 Appendix

I introduce notation that will be used throughout this section. Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space . For a set S, $l^{\infty}(S)$ is the space of uniformly bounded functions defined on the set S. $\mathbb{E}$ denotes the expectation operator. For a sequence of random map $X_n : \Omega \to \mathbb{R}$ and a random variable X, $X_n \Rightarrow X$ ($X_n \xrightarrow{d} X$, resp.) indicates that X_n converges weakly[1] (in distribution, resp.) to X. For any real sequences (a_n) and (b_n), $a_n \lesssim b_n$ means that there is a constant C, not depending on n, such that $|a_n| \leq C \cdot |b_n|$ for all $n \in \mathbb{N}$.

C.1 Proof of Lemma 3.3.1

Proof. I only prove the identification result of $F_1(y)$. It is straightforward to see that

$$\begin{aligned} F_1(y) &= \Pr(Y_1 \leq y) \\ &= \Pr(Y_1 \leq y | D = 1) \Pr(D = 1) + \Pr(Y_1 \leq y | D = 0) \Pr(D = 0) \\ &= \Pr(Y \leq y | D = 1) \Pr(D = 1) + \Pr(Y_1 \leq y | D = 0) \Pr(D = 0). \end{aligned}$$

[1] See Definition 1.3.3 in van der Vaart and Wellner (1996) for the precise definition of weak convergence.

Since $\Pr(Y_1 \leq y|D=0)$ lies in the unit interval $[0,1]$, one obtains the identification region of $F_1(y)$. □

C.2 Proof of Lemma 3.3.2

Proof. Since $F_1(y_1) \in [LB_1(y_1), UB_1(y_1)]$, one can show that $\tau \leq LB_1(Q_1^U(\tau)) \leq F_1(Q_1^U(\tau))$, which implies that $Q_1(\tau) \leq Q_1^U(\tau)$. In addition to this, since $F_1(y) \leq UB_1(y)$, it is straightforward to see that $\tau \leq F_1(Q_1(\tau)) \leq UB_1(Q_1(\tau))$. By the minimality of $Q_1^L(\tau)$, one has $Q_1^L(\tau) \leq Q_1(\tau)$. Thus, the τ-th quantile of Y_1, $Q_1(\tau)$, lies between $Q_1^L(\tau)$ and $Q_1^U(\tau)$. Equation (3.3.4) can be proven by the same way. The identification result of the τ-th QTE given by equation (3.3.5) is a direct consequence of equations (3.3.3) and (3.3.4). □

C.3 Proof of Theorem 3.3.4

Proof. Recall that $F_1(y) = \Pr(Y \leq y|D=1)\Pr(D=1) + \Pr(Y_1 \leq y|D=0)\Pr(D=0)$. Since $Y_1|D=1$ first-order stochastically dominates $Y_1|D=0$, it follows that $\Pr(Y_1 \leq y|D=0) = F_1(y|D=0) \geq F_1(y|D=1) = \Pr(Y_1 \leq y|D=1)$. Thus, $\Pr(Y_1 \leq y) \geq LB_1^{FSD1}(y)$. Note that $UB_1^{FSD1}(y)$ and $LB_0^{FSD1}(y)$ are identical to $UB_1(y)$ and $LB_0(y)$ in Lemma 3.3.1 and hence it holds. Since $F_0(y) = LB_0^{FSD1}(y) + (1-\Pr(D=0))\Pr(Y_0 \leq y|D=1) \leq LB_0^{FSD1}(y) + (1-\Pr(D=0))\Pr(Y_0 \leq y|D=0) = UB_0^{FSD1}(y)$, this results in equations (3.3.6) and (3.3.7). □

C.4 Proof of Theorem 3.3.7

Proof. Since $Y_1|D=j$ first-order stochastically dominates $Y_0|D=j$ for all $j \in \{0,1\}$, it follows that, for all $y \in \mathbb{R}$, $\Pr(Y_1 \leq y|D=1) = F_1(y|D=1) \leq F_0(y|D=1) = \Pr(Y_0 \leq y|D=1)$ and $\Pr(Y_1 \leq y|D=0) = F_1(y|D=0) \leq F_0(y|D=0) = \Pr(Y_0 \leq y|D=0)$. Following similar steps in the proof of Theorem 3.3.4, one can obtain the identified sets of $F_1(y)$ and $F_0(y)$, given by equations (3.3.8) and (3.3.9). □

C.5 Proof of Corollary 3.3.8

Proof. This is directly implied by Theorems 3.3.4 and 3.3.7. By stochastic dominance, it can be shown that

$$\begin{aligned} UB_1^{FSD2}(y) &= \Pr(Y_1 \leq y|D=1)\Pr(D=1) + \Pr(Y_0 \leq y|D=0)\Pr(D=0) \\ &\leq \Pr(Y_1 \leq y|D=1)\Pr(D=1) + \Pr(D=0) \\ &= UB_1^{FSD1}(y) = UB_1(y) \end{aligned}$$

and $LB_1^{FSD1}(y) = \Pr(Y_1 \leq y|D=1) \geq \Pr(Y_1 \leq y|D=1)\Pr(D=1) = LB_1^{FSD2}(y) = LB_1(y)$, and hence the bounds on $F_1(y)$ are narrower than the previous ones. Similarly, it is straightforward to see that the bounds on $F_0(y)$ are narrower than the previous results. □

C.6 Proof of Theorem 3.3.10

Proof. From the Lemma 2.1 in Fan and Park (2010), it is shown that

$$\sup_y\{\max[F_1(y) - F_0(y-\delta), 0]\} \leq F_\Delta(\delta) \leq \inf_y\{\min[F_1(y) - F_0(y-\delta), 0]\} + 1.$$

Since the marginal distribution functions are partially identified by the hypothesis and the functions $\max[\cdot,\cdot]$ and $\min[\cdot,\cdot]$ are non-decreasing, one obtains

$$LB_{\Delta}(\delta) \leq \sup_{y}\{\max[F_1(y) - F_0(y-\delta), 0]\}$$

and

$$UB_{\Delta}(\delta) \geq \inf_{y}\{\min[F_1(y) - F_0(y-\delta), 0]\} + 1,$$

and this ends the proof. □

C.7 Proof of Theorem 3.4.3

Proof. Note that

$$\begin{aligned}
&\Pr(\Theta_I(F_1(y)) \subseteq \mathcal{C}_n(\alpha; F_1(y))) \\
=&\Pr(LB_1(y) \geq \hat{LB}_{1n}(y) - z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{LB}_{1n}(y))}{\sqrt{n}} \\
&\text{and } UB_1(y) \leq \hat{UB}_{1n}(y) + z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{UB}_{1n}(y))}{\sqrt{n}}) \\
=&\Pr(p^*F_{11}(y) \geq \hat{LB}_{1n}(y) - z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{LB}_{1n}(y))}{\sqrt{n}} \\
&\text{and } UB_1(y) \leq \hat{UB}_{1n}(y) + z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{UB}_{1n}(y))}{\sqrt{n}}) \\
=&1 - \Pr(p^*F_{11}(y) < \hat{LB}_{1n}(y) - z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{LB}_{1n}(y))}{\sqrt{n}} \\
&\text{or } UB_1(y) > \hat{UB}_{1n}(y) + z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{UB}_{1n}(y))}{\sqrt{n}}) \\
\geq&1 - \Pr(p^*F_{11}(y) < \hat{LB}_{1n}(y) - z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{LB}_{1n}(y))}{\sqrt{n}}) \\
&+ \Pr(UB_1(y) > \hat{UB}_{1n}(y) + z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{UB}_{1n}(y))}{\sqrt{n}}), \qquad \text{(C.7.1)}
\end{aligned}$$

where the inequality in the last line comes from the Bonferroni's inequality. Applying the standard arguments of the large sample theory results in

$$\begin{aligned}
&\Pr(p^* F_{11}(y) < \hat{LB}_{1n}(y) - z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{LB}_{1n}(y))}{\sqrt{n}}) \\
=&\Pr(\frac{1}{n}\sum_i^n \{D_i \mathbf{1}(Y_i \le y) - p^* F_{11}(y)\} > z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{LB}_{1n}(y))}{\sqrt{n}}) \\
=&\Pr(\frac{1}{\sigma_F(\sqrt{n}\hat{LB}_{1n}(y))} \cdot \frac{1}{\sqrt{n}} \sum_i^n \{D_i \mathbf{1}(Y_i \le y) - p^* F_{11}(y)\} > z_{\frac{\alpha+1}{2}}) \\
\to&\frac{1-\alpha}{2}.
\end{aligned} \tag{C.7.2}$$

Similarly, it can be shown that

$$\begin{aligned}
&\Pr(UB_1(y) > \hat{UB}_{1n}(y) + \frac{\sigma_F(\sqrt{n}\hat{UB}_{1n}(y))}{\sqrt{n}}) \\
=&\Pr(\hat{UB}_{1n}(y) - UB_1(y) < -\frac{\sigma_F(\sqrt{n}\hat{UB}_{1n}(y))}{\sqrt{n}}) \\
=&\Pr(\frac{1}{\sigma_F(\sqrt{n}\hat{UB}_{1n}(y))} \sqrt{n}(\hat{UB}_{1n}(y) - UB_1(y)) < -z_{\frac{\alpha+1}{2}}) \\
\to&\frac{1-\alpha}{2}.
\end{aligned} \tag{C.7.3}$$

Therefore, combining equations (C.7.1) through (C.7.3) gives that

$$\liminf_{n\to\infty} \Pr(\Theta_I(F_1(y)) \subseteq \mathcal{C}_n(\alpha; F_1(y))) \ge \alpha.$$

□

C.8 Proof of Theorem 3.4.5

Proof. Since the inversion map is Hadamard differentiable (e.g. Lemma 3.9.20 in van der Vaart and Wellner (1996)), one can show that

$$\begin{aligned}\hat{Q}_{1n}^{U}(\tau) - Q_1^{U}(\tau) &= \hat{LB}_{1n}^{\leftarrow}(\tau) - LB_1^{\leftarrow}(\tau) \\ &= \frac{1}{p^* f_{11}(Q_1^{U}(\tau))}(\hat{LB}_{1n}(Q_1^{U}(\tau)) - LB_1(Q_1^{U}(\tau))) + o_p(n^{-1/2})\end{aligned}$$

and that

$$\begin{aligned}\hat{Q}_{1n}^{L}(\tau) - Q_1^{L}(\tau) &= \hat{UB}_{1n}^{\leftarrow}(\tau) - UB_1^{\leftarrow}(\tau) \\ &= \frac{1}{p^* f_{11}(Q_1^{L}(\tau))}(\hat{UB}_{1n}(Q_1^{L}(\tau)) - UB_1(Q_1^{L}(\tau))) + o_p(n^{-1/2}).\end{aligned}$$

Since the class of functions, $\{\mathbf{1}(Y \leq y) : y \in \mathbb{R}\}$, is Donsker, Corollary 9.32 in Kosorok (2008) leads to that for each $j \in \{0,1\}$, $\sqrt{n}(\hat{LB}_{jn}(\cdot) - LB_j(\cdot)) \Rightarrow \mathbb{G}_j^{LB}(\cdot)$ and $\sqrt{n}(\hat{UB}_{jn}(\cdot) - UB_j(\cdot)) \Rightarrow \mathbb{G}_j^{UB}(\cdot)$ for some Gaussian processes $\mathbb{G}_j^{LB}(\cdot)$ and $\mathbb{G}_j^{UB}(\cdot)$ in $l^\infty(\mathbb{R})$. Therefore,

$$\sqrt{n}(\hat{LB}_{1n}(Q_1^{U}(\tau)) - LB_1(Q_1^{U}(\tau))) \xrightarrow{d} N(0, \sigma_F^2(\sqrt{n}\hat{LB}_{1n}(Q_1^{U}(\tau))))$$

and

$$\sqrt{n}((\hat{UB}_{1n}(Q_1^{L}(\tau)) - UB_1(Q_1^{L}(\tau))) \xrightarrow{d} N(0, \sigma_F^2(\sqrt{n}\hat{UB}_{1n}(Q_1^{L}(\tau)))).$$

It follows that

$$
\begin{aligned}
&\Pr(\Theta_I(Q_1(y)) \subseteq \mathcal{C}_n(\alpha; Q_1(y))) \\
=&\Pr(Q_1^L(\tau) \geq Q_{1n}^L(\tau) - z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{LB}_{1n}(Q_1^U(\tau)))}{\sqrt{n}p^* f_{11}(Q_1^L(\tau))} \\
&\text{ and } Q_1^U(\tau) \leq Q_{1n}^U(\tau) + z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{UB}_{1n}(Q_1^L(\tau)))}{\sqrt{n}p^* f_{11}(Q_1^L(\tau))}) \\
=&1 - \Pr(Q_1^L(\tau) < Q_{1n}^L(\tau) - z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{LB}_{1n}(Q_1^U(\tau)))}{\sqrt{n}p^* f_{11}(Q_1^L(\tau))} \\
&\text{ or } Q_{1n}^U(\tau) + z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{UB}_{1n}(Q_1^L(\tau)))}{\sqrt{n}p^* f_{11}(Q_1^L(\tau))} < Q_1^U(\tau)) \\
\geq&1 - \{\Pr(Q_1^L(\tau) < Q_{1n}^L(\tau) - z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{LB}_{1n}(Q_1^U(\tau)))}{\sqrt{n}p^* f_{11}(Q_1^L(\tau))}) \\
&+ \Pr(Q_{1n}^U(\tau) + z_{\frac{\alpha+1}{2}} \cdot \frac{\sigma_F(\sqrt{n}\hat{UB}_{1n}(Q_1^L(\tau)))}{\sqrt{n}p^* f_{11}(Q_1^L(\tau))} < Q_1^U(\tau))\} \\
\to&1 - \{1 - \Phi(z_{\frac{\alpha+1}{2}}) + \Phi(-z_{\frac{\alpha+1}{2}})\} = \alpha,
\end{aligned}
$$

and thus this ends the proof of equation (3.4.7). It is straightforward to show that

$$
\begin{aligned}
\hat{Q}_{0n}^U(\tau) - Q_0^U(\tau) &= \hat{LB}_{0n}^{\leftarrow}(\tau) - LB_0^{\leftarrow}(\tau) \\
&= \frac{1}{(1-p^*)f_{00}(Q_0^U(\tau))}\{\hat{LB}_{0n}(Q_0^U(\tau)) - LB_0(Q_0^U(\tau))\} + o_p(n^{-1/2})
\end{aligned}
$$

and that

$$
\begin{aligned}
\hat{Q}_{0n}^L(\tau) - Q_0^L(\tau) &= \hat{UB}_{0n}^{\leftarrow}(\tau) - UB_0^{\leftarrow}(\tau) \\
&= \frac{1}{(1-p^*)f_{00}(Q_0^L(\tau))}\{\hat{UB}_{0n}(Q_0^L(\tau)) - UB_0(Q_0^L(\tau))\} + o_p(n^{-1/2}).
\end{aligned}
$$

The remaining part of the proof is the same as before, so it is omitted. □

C.9 Proof of Theorem 3.4.8

The proof strategy is the same to the one for Proposition 3.1 in Fan and Park (2010). To prove that the confidence region for $\Theta_I(F_\Delta(\delta))$, given in Theorem 3.4.8, is pointwise consistent in level α, I first provide several lemmas.

Lemma C.9.1. *Suppose that Assumptions 3.4.1 and 3.4.2 hold.. Then for given* $\delta \in \mathbb{R}$,

$$\sup_y\{\hat{LB}_{1n}(y) - \hat{UB}_{0n}(y-\delta)\} \xrightarrow{p} \sup_y\{LB_1(y) - UB_0(y-\delta)\},$$
$$\inf_y\{\hat{UB}_{1n}(y) - \hat{LB}_{0n}(y-\delta)\} \xrightarrow{p} \inf_y\{UB_1(y) - LB_0(y-\delta)\}.$$

Proof. It is enough to show that, for any $\delta \in \mathbb{R}$,

$$|\sup_y\{\hat{LB}_{1n}(y) - \hat{UB}_{0n}(y-\delta)\} - \sup_y\{LB_1(y) - UB_0(y-\delta)\}| \xrightarrow{p} 0$$

and

$$|\inf_y\{\hat{UB}_{1n}(y) - \hat{LB}_{0n}(y-\delta)\} - \inf_y\{UB_1(y) - LB_0(y-\delta)\}| \xrightarrow{p} 0.$$

Pick any $\delta \in \mathbb{R}$. Note that the supremum map is uniformly continuous; i.e.

$$\begin{aligned}
|\sup_t x(t) - \sup_t y(t)| &= |\sup_t\{x(t) - y(t) + y(t)\} - \sup_t y(t)| \\
&\leq |\sup_t\{x(t) - y(t)\} + \sup_t y(t) - \sup_t y(t)| \\
&\leq \sup_t |x(t) - y(t)|.
\end{aligned}$$

Therefore, one obtains that

$$
\begin{aligned}
&|\sup_y\{\hat{LB}_{1n}(y) - \hat{UB}_{0n}(y-\delta)\} - \sup_y\{LB_1(y) - UB_0(y-\delta)\}| \\
\leq &\sup_y |\hat{LB}_{1n}(y) - LB_1(y)| + \sup_y |\hat{UB}_{0n}(y-\delta) - UB_0(y-\delta)|.
\end{aligned}
$$

Now it suffices to show that the classes of functions, $\{\hat{LB}_{1n}(y) : y \in \mathbb{R}\}$, $\{\hat{LB}_{0n}(y) : y \in \mathbb{R}\}$, $\{\hat{UB}_{1n}(y) : y \in \mathbb{R}\}$, and $\{\hat{UB}_{0n}(y) : y \in \mathbb{R}\}$, are Glivenko-Cantelli. It is well-known that the class of functions, $\{\mathbf{1}(Y \leq y) : y \in \mathbb{R}\}$, is Donsker and thus Glivenko-Cantelli. Applying Corollary 9.32 in Kosorok (2008) results in that the four classes of functions are Donsker, and thus they are Glivenko-Cantelli. Therefore, for any given $\delta \in \mathbb{R}$,

$$
\begin{aligned}
&\sup_{y\in\mathbb{R}} |\hat{LB}_{1n}(y) - LB_1(y)| \xrightarrow{p} 0, \\
&\sup_{y\in\mathbb{R}} |\hat{UB}_{1n}(y) - UB_{1n}(y)| \xrightarrow{p} 0, \\
&\sup_{y\in\mathbb{R}} |\hat{LB}_{0n}(y-\delta) - LB_0(y-\delta)| \xrightarrow{p} 0, \\
&\sup_{y\in\mathbb{R}} |\hat{UB}_{0n}(y-\delta) - UB_0(y-\delta)| \xrightarrow{p} 0.
\end{aligned}
$$

The fact that $\inf_t x(t) = -\sup_t -x(t)$ ends the proof. □

Lemma C.9.2. *Suppose that Assumptions 3.4.1, (3.4.2), and 3.4.6 are satisfied. Then, for any given $\delta \in \mathbb{R}$, $\hat{y}_n^{sup}(\delta) \xrightarrow{p} y^{sup}(\delta)$ and $\hat{y}_n^{inf}(\delta) \xrightarrow{p} y^{inf}(\delta)$.*

Proof. Lemma C.9.1 and Assumption 3.4.6 together imply that the conditions of Theorem 5.7 in Van der Vaart (1998) are satisfied. Applying Theorem 5.7 in Van der Vaart (1998) follows that both estimators are consistent. □

For given $\delta \in \mathbb{R}$, define

$$M_n^L(y;\delta) \equiv \hat{LB}_{1n}(y) - \hat{UB}_{0n}(y-\delta) = \frac{1}{n}\sum_i m_i^L(y;\delta),$$
$$M_n^U(y;\delta) \equiv \hat{UB}_{1n}(y) - \hat{LB}_{0n}(y-\delta) = \frac{1}{n}\sum_i m_i^U(y;\delta),$$
$$M^L(y;\delta) \equiv LB_1(y) - UB_0(y-\delta) = \mathbb{E}m_i^L(y;\delta),$$
$$M^U(y;\delta) \equiv UB_1(y) - LB_0(y-\delta) = \mathbb{E}m_i^U(y;\delta).$$

Then it is clear that under Assumption 3.4.1 and for given y and δ, $\mathbb{E}M_n^L(y;\delta) = M^L(y;\delta)$ and $\mathbb{E}M_n^U(y;\delta) = M^U(y;\delta)$.

Lemma C.9.3. *Let $\delta \in \mathbb{R}$ be given. Suppose that Assumptions 3.4.1, 3.4.2, 3.4.4, and 3.4.6 hold. Then*

$$\hat{y}_n^{sup}(\delta) - y^{sup}(\delta) = O_p(n^{-1/3}), \tag{C.9.1}$$
$$\hat{y}_n^{inf}(\delta) - y^{inf}(\delta) = O_p(n^{-1/3}). \tag{C.9.2}$$

Proof. I verify the conditions for Theorem 3.2.5 in van der Vaart and Wellner (1996) and only prove equation (C.9.1). Then one can show that for any y in a neighborhood of $y^{sup}(\delta)$,

$$\begin{aligned}
&\mathbb{E}[m_i^L(y;\delta) - m_i^L(y^{sup}(\delta);\delta)] \\
=&\{LB_1(y) - UB_0(y-\delta)\} - \{LB_1(y^{sup}(\delta) - UB_0(y^{sup}(\delta)-\delta)\} \\
=&\{p^*F_{11}(y) - (1-p^*)\{F_{00}(y-\delta)\} - \{p^*F_{11}(y^{sup}(\delta)) - (1-p^*)F_{00}(y^{sup}(\delta)-\delta)\} \\
=&p^*(F_{11}(y) - F_{11}(y^{sup}(\delta))) - (1-p^*)(F_{00}(y-\delta) - F_{00}(y^{sup}(\delta)-\delta)).
\end{aligned}$$

By applying Taylor's expansion around $y^{sup}(\delta)$ and $y^{sup}(\delta) - \delta$ to each term, one obtains that

$$\begin{aligned}
&\mathbb{E}[m_i^L(y;\delta) - m_i^L(y^{sup}(\delta);\delta)] \\
=&\{p^* f_{11}(y^{sup}(\delta))(y - y^{sup}(\delta)) - (1-p^*) f_{00}(y^{sup}(\delta) - \delta)(y - y^{sup}(\delta))\} \\
&+ \{p^* f_{11}'(\tilde{y}(\delta))(y - y^{sup}(\delta))^2 - (1-p^*) f_{00}'(\tilde{\tilde{y}}(\delta) - \delta)(y - y^{sup}(\delta))^2\} \\
\stackrel{say}{=}& A_1(y,\delta) + A_2(y,\delta).
\end{aligned}$$

Observe that

$$A_1(y,\delta) = \{p^* f_{11}(y^{sup}(\delta)) - (1-p^*) f_{00}(y^{sup}(\delta) - \delta)\} \cdot (y - y^{sup}(\delta)) = 0 \quad \text{(C.9.3)}$$

by the first-order condition for $y^{sup}(\delta)$. Since f_{11}' and f_{00}' are continuous, the second-order condition in the theorem implies that

$$\mathbb{E}[m_i^L(y;\delta) - m_i^L(y^{sup}(\delta);\delta)] \leq C \cdot (y - y^{sup}(\delta))^2,$$

where $C = p^* f_{11}'(\tilde{y}(\delta)) + (1-p^*) f_{00}'(\tilde{\tilde{y}}(\delta) - \delta)) < 0$.

Second, consider a class of functions, $\mathcal{M}^L(\delta) \equiv \{M_n^L(y;\delta) - M^L(y;\delta) : y \in \mathbb{R}\}$. It is required to show that for all n and for any small $\eta > 0$,

$$\mathbb{E}^* \sup_{|y - y^{sup}(\delta)| < \eta} \sqrt{n} |(M_n^L(y;\delta) - M^L(y;\delta)) - (M_n^L(y^{sup}(\delta);\delta) - M^L(y^{sup}(\delta);\delta)| \lesssim \psi(\eta), \quad \text{(C.9.4)}$$

where $\psi(\eta)$ is a function such that $\psi(\eta)/\eta^\alpha$ is decreasing for some $\alpha < 2$ and $\mathbb{E}^*$ is an outer expectation. Take any small $\eta > 0$ and define a class of functions $\mathcal{M}_\eta^L(\delta) \equiv \{m_i^L(y;\delta) - m_i^L(y^{sup}(\delta);\delta) : |y - y^{sup}(\delta)| < \eta\}$. From Lemma 19.38 in

Van der Vaart (1998), it can be shown that the left-hand side of equation (C.9.4) is bounded by $J(1, \mathcal{M}_\eta^L(\delta), L_2) \cdot (\mathbb{E}^* \bar{M}_\eta^L(\delta)^2)^{1/2}$, where $J(1, \mathcal{M}_\eta^L(\delta), L_2)$ is the uniform entropy integral[2] and $\bar{M}_\eta^L(\delta)$ is an envelope function of the class $\mathcal{M}_\eta^L(\delta)$. If one can take $\bar{M}_\eta^L(\delta) \equiv \{\mathbf{1}(Y_i \leq y^{sup}(\delta) + \eta) - \mathbf{1}(Y_i \leq y^{sup}(\delta) - \eta)\} + \{\mathbf{1}(Y_i \leq y^{sup}(\delta) + \eta - \delta) - \mathbf{1}(Y_i \leq y^{sup}(\delta) - \eta - \delta)\}$ [3], then $(\mathbb{E}^* \bar{M}_\eta^L(\delta)^2)^{1/2} < \infty$, so it only requires to calculate the uniform entropy integral of $\mathcal{M}_\eta^L(\delta)$. Since the class of functions, $\{D_i \mathbf{1}(Y_i \leq y) : y \in \mathbb{R}\}$, is a Vapnik-Červonenkis (VC) class, applying Lemma 9.9 in Kosorok (2008) leads to that the class $\mathcal{M}_\eta^L(\delta)$ is a VC class and thus has bounded uniform entropy integral. Since $(\mathbb{E}^* \bar{M}_\eta^L(\delta)^2)^{1/2} \lesssim \eta^{1/2}$ under Assumptions 3.4.1 and 3.4.4, one can put $\psi(\eta) \equiv \eta^{1/2}$. Then $\psi(\eta)/\eta^\alpha$ is decreasing in η for any $\alpha > 1/2$.

Let $r_n = n^\beta$, then it is easy to see that $r_n^2 \psi(r_n^{-1}) = n^{2\beta - \frac{\beta}{2}} \lesssim \sqrt{n}$ holds if $\beta = 1/3$. By Theorem 3.2.5 in van der Vaart and Wellner (1996), one obtains that

$$\hat{y}_n^{sup}(\delta) - y^{sup}(\delta) = O_p(n^{-1/3}).$$

By the similar way, one can prove $\hat{y}_n^{inf}(\delta) - y^{inf}(\delta) = O_p(n^{-1/3})$. □

Lemma C.9.4. *Let $\delta \in \mathbb{R}$ be given. Suppose that the conditions in Theorem 3.4.8*

[2] See, for example, (Van der Vaart, 1998, p.274) for its definition.

[3] To see this, take any y such that $|y - y^{sup}(\delta)| < \eta \leq 1$. Then one obtains that $y \in (y^{sup}(\delta) - \eta, y^{sup}(\delta) + \eta)$, and hence $\mathbf{1}(Y_i \leq y^{sup}(\delta) - \eta) \leq \mathbf{1}(Y_i \leq y) \leq \mathbf{1}(Y_i \leq y^{sup}(\delta) + \eta)$. Since it is obvious that $\mathbf{1}(Y_i \leq y^{sup}(\delta) - \eta) \leq \mathbf{1}(Y_i \leq y^{sup}(\delta)) \leq \mathbf{1}(Y_i \leq y^{sup}(\delta) + \eta)$, one has that

$$|\mathbf{1}(Y_i \leq y) - \mathbf{1}(Y_i \leq y^{sup}(\delta))| \leq \mathbf{1}(Y_i \leq y^{sup}(\delta) + \eta) - \mathbf{1}(Y_i \leq y^{sup}(\delta) - \eta)$$

and that

$$|\mathbf{1}(Y_i \leq y - \delta) - \mathbf{1}(Y_i \leq y^{sup}(\delta) - \delta)| \leq \mathbf{1}(Y_i \leq y^{sup}(\delta) + \eta - \delta) - \mathbf{1}(Y_i \leq y^{sup}(\delta) - \eta - \delta).$$

are satisfied. Then

$$\sqrt{n}(M_n^L(\hat{y}_n^{sup}(\delta);\delta) - M^L(y^{sup}(\delta);\delta) \xrightarrow{d} N(0, Var(m_i^L(y^{sup}(\delta);\delta)), \tag{C.9.5}$$

$$\sqrt{n}(M_n^U(\hat{y}_n^{inf}(\delta);\delta) - M^U(y^{inf}(\delta);\delta) \xrightarrow{d} N(0, Var(m_i^U(y^{inf}(\delta);\delta)). \tag{C.9.6}$$

Proof. I only prove equation (C.9.5). Note that

$$\begin{aligned}
&\sqrt{n}(M_n^L(\hat{y}_n^{sup}(\delta);\delta) - M_n^L(\hat{y}_n^{sup}(\delta);\delta) + M_n^L(\hat{y}_n^{sup}(\delta);\delta) - M^L(y^{sup}(\delta);\delta)) \\
=&\sqrt{n}(M_n^L(\hat{y}_n^{sup}(\delta);\delta) - M_n^L(y^{sup}(\delta);\delta)) + \sqrt{n}(M_n^L(y^{sup}(\delta);\delta) - M^L(y^{sup}(\delta);\delta)).
\end{aligned}$$

I first show that $\sqrt{n}(M_n^L(\hat{y}_n^{sup}(\delta);\delta) - M_n^L(y^{sup}(\delta);\delta)) = o_p(1)$ and that the latter term $\sqrt{n}(M_n^L(y^{sup}(\delta);\delta) - M^L(y^{sup}(\delta);\delta))$ determines the asymptotic distribution. Recall that

$$\begin{aligned}
&\sqrt{n}(M_n^L(\hat{y}_n^{sup}(\delta);\delta) - M_n^L(y^{sup}(\delta);\delta)) \\
=&\frac{1}{\sqrt{n}}\sum_i \{m_i^L(\hat{y}_n^{sup}(\delta);\delta) - m_i^L(y^{sup}(\delta);\delta)\} \\
=&\frac{1}{\sqrt{n}}\sum_i [\{m_i^L(\hat{y}_n^{sup}(\delta);\delta) - M^L(\hat{y}_n^{sup}(\delta);\delta)\} - \{m_i^L(y^{sup}(\delta);\delta) - M^L(y^{sup}(\delta);\delta)\}] \\
&- \frac{1}{\sqrt{n}}\sum_i \{M^L(\hat{y}_n^{sup}(\delta);\delta) - M^L(y^{sup}(\delta);\delta)\} \\
\equiv&\tilde{M}_n(\hat{y}_n^{sup}(\delta);\delta) - \tilde{M}_n(y^{sup}(\delta);\delta) - \frac{1}{\sqrt{n}}\sum_i \{M^L(\hat{y}_n^{sup}(\delta);\delta) - M^L(y^{sup}(\delta);\delta)\},
\end{aligned}$$

where $\tilde{M}_n(y;\delta) \equiv \frac{1}{\sqrt{n}}\sum_i [\{m_i^L(y;\delta) - M^L(y;\delta)\}$ is an empirical process indexed by y. I use a stochastic equicontinuity argument to prove that the term $\tilde{M}_n(\hat{y}_n^{sup}(\delta);\delta) - \tilde{M}_n(y^{sup}(\delta);\delta)$ is $o_p(1)$. Since the class of functions, $\{m_i(y;\delta) : y \in Supp(Y)\}$, is Donsker and $(Supp(Y), |\cdot|)$ is a totally bounded metric space, $\tilde{M}_n(\cdot;\delta)$ is stochas-

tically equicontinuous. Lemma C.9.2, together with the stochastic equicontinuity, implies that $\tilde{M}_n(\hat{y}_n^{sup}(\delta);\delta) - \tilde{M}_n(y^{sup}(\delta);\delta) = o_p(1)$.

Expanding the term $\frac{1}{\sqrt{n}}\sum_i\{M^L(\hat{y}_n^{sup}(\delta);\delta) - M^L(y^{sup}(\delta);\delta)\}$ around $y^{sup}(\delta)$ results in that

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_i\{M^L(\hat{y}_n^{sup}(\delta);\delta) - M^L(y^{sup}(\delta);\delta)\} \\
=&\sqrt{n}[\{p^*F_{11}(\hat{y}_n^{sup}(\delta)) - (1-p^*)\{F_{00}(\hat{y}_n^{sup}(\delta)-\delta)\} \\
&- \{p^*F_{11}(y^{sup}(\delta)) - (1-p^*)F_{00}(y^{sup}(\delta)-\delta)\}] \\
=&\{p^*f_{11}^{'}(\tilde{y}^{sup}(\delta)) - (1-p^*)f_{00}^{'}(\tilde{y}^{sup}(\delta))\}\sqrt{n}(\hat{y}_n^{sup}(\delta) - y^{sup}(\delta))^2,
\end{aligned}$$

where $\tilde{y}^{sup}(\delta)$ is a value between $y^{sup}(\delta)$ and $\hat{y}_n^{sup}(\delta)$. Note that the first-order terms disappear by the first-order condition for $y^{sup}(\delta)$ (i.e. equation (C.9.3)). Lemma C.9.3 and Assumption 3.4.4 together imply that $\frac{1}{\sqrt{n}}\sum_i\{M^L(\hat{y}_n^{sup}(\delta);\delta) - M^L(y^{sup}(\delta);\delta)\} = o_p(1)$.

It remains to show that

$$\sqrt{n}(M_n^L(y^{sup}(\delta);\delta) - M^L(y^{sup}(\delta);\delta)) \xrightarrow{d} N(0, Var(m_i^L(y^{sup}(\delta);\delta)).$$

Recall that the class $\{m_i(y;\delta) : y \in Supp(Y)\}$ is Donsker, and thus

$$\sqrt{n}(M_n^L(\cdot;\delta) - M^L(\cdot;\delta)) \Rightarrow \mathbb{G}_{M^L}(\cdot) \text{ in } l^\infty(Supp(Y)),$$

where $\mathbb{G}_{M^L}(\cdot)$ is a Gaussian process with mean zero and covariance kernel $\Sigma_{M^L}(y_1, y_2) \equiv$

$Cov(m_i(y_1;\delta), m_i(y_2;\delta))$. Therefore,

$$\sqrt{n}(M_n^L(y^{sup}(\delta);\delta) - M^L(y^{sup}(\delta);\delta)) \xrightarrow{d} N(0, Var(m_i^L(y^{sup}(\delta);\delta)).$$

Similarly, one can establish that

$$\sqrt{n}(M_n^U(y^{inf}(\delta);\delta) - M^U(y^{inf}(\delta);\delta)) \xrightarrow{d} N(0, Var(m_i^U(y^{inf}(\delta);\delta)),$$

and this ends the proof. □

Proof of the theorem

Proof. Recall that $\max[\cdot,\cdot]$ and $\min[\cdot,\cdot]$ are continuous functions. By Lemma C.9.4 and the continuous mapping theorem, it can be shown that

$$\sqrt{n}(\hat{LB}_{\Delta n}(\delta) - LB_\Delta(\delta)) \xrightarrow{d} \max[N(0, Var(m_i^L(y^{sup}(\delta);\delta)), 0]$$

and that

$$\sqrt{n}(\hat{UB}_{\Delta n}(\delta) - UB_\Delta(\delta)) \xrightarrow{d} \min[N(0, Var(m_i^U(y^{inf}(\delta);\delta)), 0] + 1,$$

and this ends the proof. □

Bibliography

Abadie, A., J. Angrist, and G. Imbens (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica 70*(1), 91–117.

Ai, C. and X. Chen (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica 71*(6), 1795–1843.

Altonji, J. G., T. E. Elder, and C. R. Taber (2005). An evaluation of instrumental variable strategies for estimating the effects of catholic schooling. *Journal of Human Resources 40*(4), 791–821.

Andrews, D. W. (1994a). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica 62*(1), 43–72.

Andrews, D. W. (1994b). Empirical process methods in econometrics. *Handbook of Econometrics 4*, 2247–2294.

Andrews, D. W. and X. Shi (2013). Inference based on conditional moment inequalities. *Econometrica 81*(2), 609–666.

Andrews, D. W. and G. Soares (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica 78*(1), 119–157.

Andrews, D. W. and Y.-J. Whang (1990). Additive interactive regression models: circumvention of the curse of dimensionality. *Econometric Theory 6*(4), 466–479.

Bang, H. and A. A. Tsiatis (2000). Estimating medical costs with censored data. *Biometrika 87*(2), 329–343.

Belloni, A. and V. Chernozhukov (2011). l_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics 39*(1), 82–130.

Belloni, A., V. Chernozhukov, D. Chetverikov, and I. Fernández-Val (2016). Conditional quantile processes based on series or many regressors. *arXiv preprint arXiv:1105.6154* .

Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical report, University of California, Berkeley.

Berrington, A. and I. Diamond (2000). Marriage or cohabitation: A competing risks analysis of first-partnership formation among the 1958 british birth cohort. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 163*(2), 127–151.

Bhattacharya, J., D. Goldman, and D. McCaffrey (2006). Estimating probit models with self-selected treatments. *Statistics in Medicine 25*(3), 389–413.

Bhattacharya, J., A. M. Shaikh, and E. Vytlacil (2008). Treatment effect bounds under monotonicity assumptions: An application to swan-ganz catheterization. *The American Economic Review 98*(2), 351–356.

Bierens, H. J. (1982). Consistent model specification tests. *Journal of Econometrics 20*(1), 105–134.

Bierens, H. J. (1990). A consistent conditional moment test of functional form. *Econometrica 58*(6), 1443–1458.

Bierens, H. J. (2008). Semi-nonparametric interval-censored mixed proportional hazard models: Identification and consistency results. *Econometric Theory 24*(3), 749–794.

Bierens, H. J. (2014). Consistency and asymptotic normality of sieve ml estimators under low-level conditions. *Econometric Theory 30*(5), 1021–1076.

Bierens, H. J. and D. K. Ginther (2001). Integrated conditional moment testing of quantile regression models. *Empirical Economics 26*(1), 307–324.

Bierens, H. J. and W. Ploberger (1997). Asymptotic theory of integrated conditional moment tests. *Econometrica 65*(5), 1129–1151.

Bitler, M. P., J. B. Gelbach, and H. W. Hoynes (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *The American Economic Review 96*(4), 988–1012.

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir (2007). Changes in the distribution of male and female wages accounting for employment composition using bounds. *Econometrica 75*(2), 323–363.

Buchinsky, M. (1994). Changes in the us wage structure 1963-1987: Application of quantile regression. *Econometrica 62*(2), 405–458.

Buchinsky, M. and J. Hahn (1998). An alternative estimator for the censored quantile regression model. *Econometrica 66*(3), 653–671.

Bugni, F. A. (2010). Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set. *Econometrica 78*(2), 735–753.

Card, D. (1999). The causal effect of education on earnings. *Handbook of Labor Economics 3*, 1801–1863.

Chao, S.-K., S. Volgushev, and G. Cheng (2016). Quantile processes for semi and nonparametric regression. *arXiv preprint arXiv:1604.02130*.

Chaudhuri, P. (1991). Nonparametric estimates of regression quantiles and their local Bahadur representation. *The Annals of Statistics 19*(2), 760–777.

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

Chen, X. and Y. Fan (1999). Consistent hypothesis testing in semiparametric and nonparametric models for econometric time series. *Journal of Econometrics 91*(2), 373–401.

Chen, X. and Y. Fan (2006a). Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification. *Journal of Econometrics 135*(1), 125–154.

Chen, X. and Y. Fan (2006b). Estimation of copula-based semiparametric time series models. *Journal of Econometrics 130*(2), 307–335.

Chen, X., Y. Fan, and V. Tsyrennikov (2006). Efficient estimation of semiparametric multivariate copula models. *Journal of the American Statistical Association 101*(475), 1228–1240.

Chen, X., Y. Hu, and A. Lewbel (2009). Nonparametric identification and estimation of nonclassical errors-in-variables models without additional information. *Statistica Sinica*, 949–968.

Chen, X., Z. Liao, and Y. Sun (2014). Sieve inference on possibly misspecified semi-nonparametric time series models. *Journal of Econometrics 178*(3), 639–658.

Chen, X., O. Linton, and I. Van Keilegom (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica 71*(5), 1591–1608.

Chen, X. and D. Pouzo (2015). Sieve wald and qlr inferences on semi/nonparametric conditional moment models. *Econometrica 83*(3), 1013–1079.

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

Chernozhukov, V. and I. Fernández-Val (2005). Subsampling inference on quantile regression processes. *Sankhyā: The Indian Journal of Statistics 67*(2), 253–276.

Chernozhukov, V. and C. Hansen (2005). An iv model of quantile treatment effects. *Econometrica 73*(1), 245–261.

Chernozhukov, V. and H. Hong (2002). Three-step censored quantile regression and extramarital affairs. *Journal of the American Statistical Association 97*(459), 872–882.

Chernozhukov, V., H. Hong, and E. Tamer (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica 75*(5), 1243–1284.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013). Intersection bounds: estimation and inference. *Econometrica 81*(2), 667–737.

Chesher, A. (2005). Nonparametric identification under discrete variation. *Econometrica 73*(5), 1525–1550.

Chiburis, R. (2010). Semiparametric bounds on treatment effects. *Journal of Econometrics 159*(2), 267–275.

Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional kaplan-meier estimate. *The Annals of Statistics 17*(3), 1157–1167.

Dabrowska, D. M. (1992). Nonparametric quantile regression with censored data. *Sankhyā: The Indian Journal of Statistics, Series A 54*(2), 252–259.

Delgado, M. A. and W. G. Manteiga (2001). Significance testing in nonparametric regression based on the bootstrap. *The Annals of Statistics 29*(5), 1469–1507.

Donald, S. G. (1997). Inference concerning the number of factors in a multivariate nonparametric relationship. *Econometrica 65*(1), 103–131.

Donald, S. G. and Y.-C. Hsu (2014). Estimation and inference for distribution functions and quantile functions in treatment effect models. *Journal of Econometrics 178*(3), 383–397.

Escanciano, J. C. and S.-C. Goh (2014). Specification analysis of linear quantile models. *Journal of Econometrics 178*(3), 495–507.

Escanciano, J. C., D. T. Jacho-Chávez, and A. Lewbel (2014). Uniform convergence of weighted sums of non and semiparametric residuals for estimation and testing. *Journal of Econometrics 178*(3), 426–443.

Escanciano, J. C. and C. Velasco (2010). Specification tests of parametric dynamic conditional quantiles. *Journal of Econometrics 159*(1), 209–221.

Evans, W. N. and R. M. Schwab (1995). Finishing high school and starting college: Do catholic schools make a difference? *The Quarterly Journal of Economics 110*(4), 941–974.

Fan, J. and I. Gijbels (1996). *Local polynomial modelling and its applications*, Volume 66 of *Monographs on statistics and applied probability*. CRC Press.

Fan, Y., E. Guerre, and D. Zhu (2014). Partial identification and confidence sets for functionals of the joint distribution of "potential outcomes". Technical report, Working paper.

Fan, Y. and Q. Li (1996). Consistent model specification tests: omitted variables and semiparametric functional forms. *Econometrica 64*(4), 865–890.

Fan, Y. and R. Liu (2013). Partial identification and inference in censored quantile regression: A sensitivity analysis. Technical report, working paper.

Fan, Y. and S. S. Park (2010). Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory 26*(03), 931–951.

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

Freyberger, J. and M. Masten (2015). Compactness of infinite dimensional parameter spaces. Technical report, Centre for Microdata Methods and Practice.

Gallant, A. R. and D. W. Nychka (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica 55*(2), 363–390.

Geman, S. and C.-R. Hwang (1982). Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics 10*(2), 401–414.

Goldman, D., J. Bhattacharya, D. Mccaffrey, N. Duan, A. Leibowitz, G. Joyce, and S. Morton (2001). Effect of Insurance on Mortality in an HIV-Positive Population in Care. *Journal of the American Statistical Association 96*(455).

Gonzalez-Manteiga, W. and C. Cadarso-Suarez (1994). Asymptotic properties of a generalized kaplan-meier estimator with some applications. *Journal of Nonparametric Statistics 4*(1), 65–78.

Guerre, E. and C. Sabbah (2012). Uniform bias study and Bahadur representation for local polynomial estimators of the conditional quantile function. *Econometric Theory 28*(1), 87–129.

Han, A. and J. A. Hausman (1990). Flexible parametric estimation of duration and competing risk models. *Journal of Applied Econometrics 5*(1), 1–28.

Han, S. and E. Vytlacil (2017). Identification in a generalization of bivariate probit models with dummy endogenous regressors. *Journal of Econometrics 199*(1), 63–73.

Hardle, W. and E. Mammen (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics 21*(4), 1926–1947.

He, X. and L.-X. Zhu (2003). A lack-of-fit test for quantile regression. *Journal of the American Statistical Association 98*(464), 1013–1022.

Heckman, J. and B. Singer (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica 52*(2), 271–320.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica 47*(1), 153–162.

Heckman, J. J., J. Smith, and N. Clements (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies 64*(4), 487–535.

Heckman, J. J. and E. J. Vytlacil (2007). Econometric evaluation of social programs, part i: Causal models, structural models and econometric policy evaluation. *Handbook of Econometrics 6B*, 4779–4874.

Hendricks, L. and O. Leukhina (2014). The return to college: Selection and dropout risk.

Hong, Y. and H. White (1995). Consistent specification testing via nonparametric series regression. *Econometrica 63*(5), 1133–1159.

Honore, B., S. Khan, and J. L. Powell (2002). Quantile regression under random censoring. *Journal of Econometrics 109*(1), 67–105.

Horowitz, J. L. and S. Lee (2005). Nonparametric estimation of an additive quantile regression model. *Journal of the American Statistical Association 100*(472), 1238–1249.

Horowitz, J. L. and V. G. Spokoiny (2002). An adaptive, rate-optimal test of linearity for median regression models. *Journal of the American Statistical Association 97*(459), 822–835.

Hu, Y. and S. M. Schennach (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica 76*(1), 195–216.

Huang, M., Y. Sun, and H. White (2016). A flexible nonparametric test for conditional independence. *Econometric Theory 32*(6), 1434–1482.

Ieva, F., G. Marra, A. M. Paganoni, and R. Radice (2014). A semiparametric bivariate probit model for joint modeling of outcomes in stemi patients. *Computational and mathematical methods in medicine 2014*.

Imbens, G. and J. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Imbens, G. W. and C. F. Manski (2004). Confidence intervals for partially identified parameters. *Econometrica 72*(6), 1845–1857.

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

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Jun, S. J., J. Pinkse, and H. Xu (2011). Tighter bounds in triangular systems. *Journal of Econometrics 161*(2), 122–128.

Jun, S. J., J. Pinkse, and H. Xu (2012). Discrete endogenous variables in weakly separable models. *The Econometrics Journal 15*(2), 288–303.

Kaplan, E. L. and P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association 53*(282), 457–481.

Katz, L. F. and B. D. Meyer (1990). Unemployment insurance, recall expectations, and unemployment outcomes. *The Quarterly Journal of Economics 105*(4), 973–1002.

Khan, S., M. Ponomareva, and E. Tamer (2011). Sharpness in randomly censored linear models. *Economics Letters 113*(1), 23–25.

Khan, S. and E. Tamer (2009). Inference on endogenously censored regression models using conditional moment inequalities. *Journal of Econometrics 152*(2), 104–119.

Kim, G., M. J. Silvapulle, and P. Silvapulle (2007a). Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics & Data Analysis 51*(6), 2836–2850.

Kim, G., M. J. Silvapulle, and P. Silvapulle (2007b). Semiparametric estimation of the error distribution in multivariate regression using copulas. *Australian & New Zealand Journal of Statistics 49*(3), 321–336.

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

Koenker, R. and G. Bassett (1982). Tests of linear hypotheses and l" 1 estimation. *Econometrica 50*(6), 1577–1583.

Koenker, R. and K. Hallock (2001). Quantile regression. *Journal of Economic Perspectives 15*(4), 143–156.

Koenker, R. and J. A. Machado (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association 94*(448), 1296–1310.

Koenker, R. and Z. Xiao (2002). Inference on the quantile regression process. *Econometrica 70*(4), 1583–1612.

Kong, E., O. Linton, and Y. Xia (2010). Uniform Bahadur representation for local polynomial estimates of m-regression and its application to the additive model. *Econometric Theory 26*(5), 1529–1564.

Kong, E., O. Linton, and Y. Xia (2013). Global bahadur representation for nonparametric censored regression quantiles and its applications. *Econometric Theory 29*(5), 941–968.

Kong, E. and Y. Xia (2017). Uniform bahadur representation for nonparametric censored quantile regression: A redistribution-of-mass approach. *Econometric Theory 33*(1), 242–261.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Springer.

Koul, H. L. and W. Stute (1999). Nonparametric model checks for time series. *The Annals of Statistics 27*(1), 204–236.

Lavergne, P., S. Maistre, V. Patilea, et al. (2015). A significance test for covariates in nonparametric regression. *Electronic Journal of Statistics 9*(1), 643–678.

Lavergne, P. and V. Patilea (2008). Breaking the curse of dimensionality in nonparametric testing. *Journal of Econometrics 143*(1), 103–122.

Lavergne, P. and Q. Vuong (2000). Nonparametric significance testing. *Econometric Theory 16*(4), 576–601.

Lawless, J. F. and Y. E. Yilmaz (2011). Comparison of semiparametric maximum likelihood estimation and two-stage semiparametric estimation in copula models. *Computational Statistics & Data Analysis 55*(7), 2446–2455.

Lee, S. (2003). Efficient semiparametric estimation of a partially linear quantile regression model. *Econometric theory 19*(1), 1–31.

Lee, S., K. Song, and Y.-J. Whang (2015). Uniform asymptotics for nonparametric quantile regression with an application to testing monotonicity. *arXiv preprint arXiv:1506.05337*.

Li, Q., C. Hsiao, and J. Zinn (2003). Consistent specification tests for semiparametric/nonparametric models based on series estimation methods. *Journal of Econometrics 112*(2), 295–325.

Liao, Z. and X. Shi (2017). A uniform model selection test for semi/nonparametric models. *Working paper*.

Lorentz, G. (1966). *Approximation of functions*. Holt, Rinehart and Winston New York.

Manski, C. F. (1988). Identification of binary response models. *Journal of the American Statistical Association 83*(403), 729–738.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review 80*(2), 319–323.

Manski, C. F. (1994). The selection problem. In *Advances in Econometrics, Sixth World Congress*, Volume 1, pp. 143–70.

Manski, C. F. (1997). Monotone treatment response. *Econometrica 65*(6), 1311–1334.

Manski, C. F. and J. V. Pepper (2000). Monotone instrumental variables: With an application to the returns to schooling. *Econometrica 68*(4), 997–1010.

Marra, G. and R. Radice (2011). Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity. *Canadian Journal of Statistics 39*(2), 259–279.

Meyer, B. D. (1990). Unemployment insurance and unemployment spells. *Econometrica 58*(4), 757–782.

Mourifié, I. (2015). Sharp bounds on treatment effects in a binary triangular system. *Journal of Econometrics 187*(1), 74–81.

Mourifié, I. and R. Méango (2014). A note on the identification in two equations probit model with dummy endogenous regressor. *Economics Letters 125*(3), 360–363.

Neal, D. A. (1997). The effects of catholic secondary schooling on educational achievement. *Journal of Labor Economics 15*(1), 98–123.

Nelsen, R. B. (1999). *An introduction to copulas*. Springer Verlag.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica 62*(6), 1349–1382.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics 4*, 2111–2245.

Ossiander, M. (1987). A central limit theorem under metric entropy with L2 bracketing. *The Annals of Probability 15*(3), 897–919.

Politis, D. N. and J. P. Romano (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics 22*(4), 2031–2050.

Portnoy, S. (2003). Censored regression quantiles. *Journal of the American Statistical Association 98*(464), 1001–1012.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica 57*(6), 1403–1430.

Qu, Z. and J. Yoon (2015). Nonparametric estimation and inference on conditional quantile processes. *Journal of Econometrics 185*(1), 1–19.

Rhine, S. L., W. H. Greene, and M. Toussaint-Comeau (2006). The importance of check-cashing businesses to the unbanked: Racial/ethnic differences. *Review of Economics and Statistics 88*(1), 146–157.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica 56*(4), 931–954.

Romano, J. P. and A. M. Shaikh (2010). Inference for the identified set in partially identified econometric models. *Econometrica 78*(1), 169–211.

Ruggles, S., J. T. Alexander, K. Genadek, R. Goeken, M. B. Schroeder, and M. Sobek (2010). Integrated public use microdata series: Version 5.0 [machine-readable database]. Minneapolis, MN: University of Minnesota.

Sant'Anna, P. H. (2016). Nonparametric tests for treatment effect heterogeneity with duration outcomes. Technical report, Vanderbilt University.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. John Wiley & Sons.

Shaikh, A. M. and E. J. Vytlacil (2011). Partial identification in triangular systems of equations with binary dependent variables. *Econometrica 79*(3), 949–955.

Shen, X. and W. H. Wong (1994). Convergence rate of sieve estimates. *The Annals of Statistics 22*(2), 580–615.

Song, K. (2009). Testing conditional independence via Rosenblatt transforms. *The Annals of Statistics 37*(6B), 4011–4045.

Stinchcombe, M. B. and H. White (1998). Consistent specification testing with nuisance parameters present only under the alternative. *Econometric theory 14*(3), 295–325.

Stoye, J. (2009). More on confidence intervals for partially identified parameters. *Econometrica 77*(4), 1299–1315.

Stoye, J. (2010). Partial identification of spread parameters. *Quantitative Economics 1*(2), 323–357.

Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics 25*(2), 613–641.

Su, L. and H. White (2008). A nonparametric hellinger metric test for conditional independence. *Econometric Theory 24*(4), 829–864.

Su, L. and H. L. White (2012). Conditional independence specification testing for dependent processes with local polynomial quantile regression. In *Essays in Honor of Jerry Hausman*, pp. 355–434. Emerald Group Publishing Limited.

van de Geer, S. A. (2000). *Empirical Processes in M-estimation*, Volume 6. Cambridge university press.

van der Vaart, A. and J. Wellner (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer, New York.

Van der Vaart, A. W. (1998). *Asymptotic statistics*, Volume 3. Cambridge university press.

Vella, F. (1998). Models with sample selection bias: A survey. *Journal of Human Resources 33*(1), 127–169.

Volgushev, S., M. Birke, H. Dette, N. Neumeyer, et al. (2013). Significance testing in quantile regression. *Electronic Journal of Statistics 7*, 105–145.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica 57*(2), 307–333.

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

Wang, H. J. and L. Wang (2009). Locally weighted censored quantile regression. *Journal of the American Statistical Association 104*(487), 1117–1128.

Whang, Y.-J. (2006a). Consistent specification testing for quantile regression models. *Econometric Theory and Practice: Frontiers of Analysis and Applied Research*, 288–308.

Whang, Y.-J. (2006b). Smoothed empirical likelihood methods for quantile regression models. *Econometric Theory 22*(2), 173–205.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*(1), 1–25.

White, H. and J. Wooldridge (1991). Some results on sieve estimation with dependent observations. *Nonparametric and Semiparametric Methods in Economics*, 459–493.

White, N. E. and A. M. Wolaver (2003). Occupation choice, information, and migration. *The Review of Regional Studies 33*(2), 142.

Wilde, J. (2000). Identification of multiple equation probit models with endogenous dummy regressors. *Economics letters 69*(3), 309–312.

Williamson, R. C. and T. Downs (1990). Probabilistic arithmetic. I. numerical methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning 4*(2), 89–158.

Zheng, J. X. (1998). A consistent nonparametric test of parametric regression models under conditional quantile restrictions. *Econometric Theory 14*(1), 123–138.