Minimum Distance from Independence Estimation of Nonseparable Instrumental Variables Models

Alexander Torgovitsky*

April 9, 2016

Abstract

I develop a semiparametric minimum distance from independence estimator for a nonseparable instrumental variables model. An independence condition identifies the model for many types of discrete and continuous instruments. The estimator is taken as the parameter value that most closely satisfies this independence condition. Implementing the estimator requires a quantile regression of the endogenous variables on the instrument, so the procedure is two-step, with a finite or infinite-dimensional nuisance parameter in the first step. I prove consistency and establish asymptotic normality for a parametric, but flexibly nonlinear outcome equation. The consistency of the nonparametric bootstrap is also shown. I illustrate the use of the estimator by estimating the returns to schooling using data from the 1979 National Longitudinal Survey.

*Department of Economics, Northwestern University, a-torgovitsky@northwestern.edu. I thank Donald Andrews, Xiaohong Chen and Edward Vytlacil, who were gracious with their advice, support and feedback. I have benefited from comments by Joachim Freyberger, Bo Honoré, Joel Horowitz, Whitney Newey, Andres Santos, Elie Tamer. Three anonymous referees also provided helpful comments. I thank attendees of the 2011 Review of Economics Studies May Meetings at London Business School, IIES and CEMFI, as well as seminar participants at UT Austin, Penn. St., Brown, Stanford, UC Berkeley, Princeton, UC San Diego, Harvard/MIT, NYU, IUPUI and Chicago Booth.

1 Introduction

Econometric models with nonseparable (not additively separable) errors are fundamentally implied by most economic theory. If economic agents make decisions according to marginal incentives, and if these marginal incentives vary due to unobserved heterogeneity, then equations relating the levels of decision and outcome variables will be inherently nonseparable in a latent error. Ignoring such heterogeneity is a source of misspecification with important empirical consequences (Heckman, 2001).

Well-established inferential approaches are available for situations where a nonseparable error term is independent of the decision, treatment or explanatory variable of interest (perhaps after conditioning on observed covariates), see for example Matzkin (2003), Imbens (2004) and Koenker (2005). However, for a large number of economically interesting questions, such exogeneity or conditional exogeneity assumptions are untenable. This is especially true of situations in which the latent terms reflect factors that affect the decision process, such as preferences, expectations or private information. Fewer results exist for such cases, owing primarily to the difficulty of establishing nonparametric identification under both endogeneity and nonseparability.

Much of the recent literature on nonseparable models with endogeneity, including this paper, focuses on semi– or nonparametric instrumental variable (IV) models. Imbens and Newey (2009) establish a nonparametric point identification result by assuming that the instrument is continuously distributed with full support. Chesher (2003) does not impose a large support assumption, but still requires a continuously distributed instrument to nonparametrically identify local marginal effects. This is also true of Florens et al. (2008), who identify an outcome equation with a flexible polynomial form. However, continuity rules out binary and discrete instruments, such as the intent-to-treat, which are commonly used in empirical work. Torgovitsky (2015) shows that such instruments actually can be used for nonparametric identification in nonseparable models with endogeneity, as long as one restricts the dimension of heterogeneity

in the outcome and first-stage equations.[1]

This paper contains the development of an estimator that implements the identification result in Torgovitsky (2015). In the next section, I briefly review the model and assumptions for which the result is established. In Sections 3 and 4, I develop a semiparametric estimator of the parameters in the outcome equation. The estimator is based on an independence condition that holds only under the true data generating process. This independence condition depends on the distribution of the endogenous explanatory variables, conditional on the instruments, so the procedure is two-step. In the first step, the endogenous variables are quantile-regressed on the instruments. In the second step, an estimate of the outcome equation is found by attempting to satisfy a sample analog of the independence condition. Although the identification result is nonparametric, I assume for the purposes of estimation that the admissible collection of outcome equations can be indexed by a finite-dimensional parameter. This facilitates deriving the asymptotic distribution of the estimator of this parameter, while still accommodating functional forms that are flexibly nonlinear. The asymptotic variance turns out to be quite complicated, so I also verify that the nonparametric bootstrap is a valid inferential procedure. Section 5 contains the results of several Monte Carlo experiments. I illustrate the use of the estimator by estimating the returns to schooling in Section 6.

2 Model and Motivation

Estimating the causal effect of schooling on wages is a classic and long-standing problem in economics. It is usually thought that dependence between wages (Y) and schooling investment (X) is confounded by the effect of unobserved factors (ε), typically referred to loosely as ability, that predispose agents to obtain higher levels of both. As a result,

[1] This result was first shown in Torgovitsky (2010). D'Haultfœuille and Février (2015) showed that, under additional restrictions, some of these results can be interpreted using group theory.

the relationship between Y and X tends to overstate the causal effect of schooling on wages (the returns to schooling), which is typically the policy-relevant quantity of interest.

Despite the intuitive appeal of this argument, it has proven difficult to confirm empirically. Many studies have estimated a separable linear model using two-stage least squares in an attempt to obtain unbiased estimates of the returns to schooling. Often, these types of studies have shown the opposite of what is predicted by the ability bias argument. That is, IV estimates of the returns to schooling tend to be larger, not smaller, than their ordinary least squares (OLS) counterparts.[2]

As discussed by Card (1995, 2001), one potential explanation for this phenomenon is that the relationship between wages and schooling is nonseparable, as would result from heterogeneity in preferences, discount factors, or the ceteris paribus path of life-cycle earnings. If this is the case, then IV estimates will place more weight on the returns faced by agents whose schooling decisions are more heavily impacted by the particular instrument being used. Most of the instruments considered in the empirical literature are cost shifters that should have a larger impact on agents inclined to obtain lower levels of schooling. If there are diminishing returns to schooling, these agents will tend to have higher marginal returns, which will inflate the IV estimate. In other words, the ability bias explanation may be accurate, with the larger IV estimates resulting from a failure to account for heterogeneity in marginal returns.

The nonseparable model analyzed in Torgovitsky (2015), together with the estimator discussed in the rest of the paper, provide an empirical framework for evaluating this argument by obtaining IV estimates that allow for both heterogeneity and endogeneity. The model postulates that Y is determined as

$$Y = g_{\theta_0}(X, \varepsilon), \tag{1}$$

[2]See Table II in Card (2001) for a summary of several studies.

where X is a vector of included explanatory variables (or treatments) with support $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and ε is an unobservable with support $\mathcal{E} \subseteq \mathbb{R}$.[3] The outcome functions are parameterized by $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$, with $\theta_0 \in \Theta$ denoting the element that is assumed to generate the data.[4] In the returns to schooling problem, as in many other economic applications, X is endogenous, so it is assumed that there is an instrument Z with support $\mathcal{Z}$ that is excluded from (1). The following assumptions are maintained in conjunction with (1).[5]

Assumption I.

I1. ***(Continuity)*** *a) For every $\theta \in \Theta$, g_θ is everywhere continuous. b) $(X, \varepsilon)|Z = z$ is continuously distributed for each z.*

I2. ***(Scalar heterogeneity)*** *For every $\theta \in \Theta$, $g_\theta(x, \cdot)$ is strictly increasing for all x.*

I3. ***(Normalization)*** *If $\theta, \theta' \in \Theta$ and $\theta \neq \theta'$, then there does not exist a strictly increasing function ψ such that $g_\theta(x, e) = g_{\theta'}(x, \psi(e))$ for all $(x, e) \in \text{supp}(X, \varepsilon)$.*

I4. ***(First stage equation)*** *There exists an unobserved vector $\eta \in \mathbb{R}^{d_x}$ and functions h_k such that for each $k = 1, \ldots, d_x$, (i) $X_k = h_k(Z, \eta_k)$, (ii) $h_k(z, \cdot)$ is strictly increasing for every z and (iii) $(\eta, \varepsilon) \perp\!\!\!\perp Z$.*

Assumptions I are discussed in detail in Torgovitsky (2015). In that paper it is shown that Assumptions I, together with a relevance condition, imply that

$$(V, \varepsilon_\theta) \perp\!\!\!\perp Z \quad \Leftrightarrow \quad \theta = \theta_0, \tag{2}$$

where $\varepsilon_\theta \equiv g_\theta^{-1}(X, Y)$ for any $\theta \in \Theta$ and $V \equiv \vec{F}_{X|Z}(X \mid Z)$, where $\vec{F}_{X|Z}(x \mid z) \equiv (F_{X_1|Z}(x_1 \mid z), \ldots, F_{X_{d_x}|Z}(x_{d_x} \mid z))$ is the vector of marginal distributions of X, condi-

[3] For any random vector X, $\text{supp}(X)$ denotes the support of X, which is defined as the smallest closed set $\mathcal{S}$, such that $\mathbb{P}[X \in \mathcal{S}] = 1$.

[4] The assumption that Θ is finite-dimensional is made for the asymptotic analysis, but is not needed to establish identification.

[5] All of the assumptions and identification results extend easily to accommodate included covariates in (1). As I explain in Section 4.5, extending the estimation results to handle covariates is straightforward when they are discrete, but introduces some additional complications when they are continuous.

tional on $Z = z$. The equivalence in (2) shows that θ_0 is identified, because the random vector $(V, \varepsilon_\theta, Z)$ is comprised of identified or known functions of the observable random vector (X, Y, Z) for any given θ. The additional relevance condition needed for (2) is quite mild, and essentially just requires that the marginal distribution of X, conditional on $Z = z$, varies with z. See Torgovitsky (2015) for the exact statement. A remarkable feature about the relevance condition is that it can be satisfied regardless of whether Z has a continuous, discrete or even binary distribution.

Assumptions I do not require X to be independent of ε. Manski (1983) proposed a minimum distance from independence estimator for this case and established its consistency. Brown and Wegkamp (2002) derived the asymptotic distribution of a similar estimator, while Linton et al. (2008), Komunjer and Santos (2010) and Santos (2011) generalized the analysis to allow for various types of specifications with an infinite dimensional θ. These papers are all based on the identification condition that $\varepsilon_\theta \perp\!\!\!\perp X$ if and only if $\theta = \theta_0$. This condition reduces to (2) for $Z = X$, since V becomes degenerate in this case. Hence, the contribution of this paper can be viewed as extending the analysis of Brown and Wegkamp (2002) to allow for X to be endogenous. Since I assume that θ is finite dimensional, the model is not nested with those in Linton et al. (2008), Komunjer and Santos (2010) or Santos (2011).

3 Estimation

An implication of (2) is that the function

$$\begin{aligned} D_\theta(t) \equiv \mathbb{P}\left[\vec{F}_{X|Z}(X \mid Z) \leq t_v, g_\theta^{-1}(X,Y) \leq t_e, Z \leq t_z\right] \\ - \mathbb{P}\left[\vec{F}_{X|Z}(X \mid Z) \leq t_v, g_\theta^{-1}(X,Y) \leq t_e\right] \mathbb{P}\left[Z \leq t_z\right] \\ = \mathbb{P}\left[X \leq \vec{Q}_{X|Z}(t_v \mid Z), Y \leq g_\theta(X, t_e), Z \leq t_z\right] \\ - \mathbb{P}\left[X \leq \vec{Q}_{X|Z}(t_v \mid Z), Y \leq g_\theta(X, t_e)\right] \mathbb{P}\left[Z \leq t_z\right] \end{aligned} \tag{3}$$

is zero for every $t = (t_v, t_e, t_z) \in \mathcal{T} \equiv (0,1)^{d_x} \times \mathcal{E} \times \mathcal{Z}$ if and only if $\theta = \theta_0$, where $\vec{Q}_{X|Z}(v\,|\,z) \equiv (Q_{X_1|Z}(v_1\,|\,z), \ldots, Q_{X_{d_x}|Z}(v_{d_x}\,|\,z))$.[6,7] If $\|\cdot\|_\mu$ is the L_2–norm with respect to a probability measure μ with support containing $\mathcal{T}$, then $\|D_\theta\|_\mu \geq 0$ and $\|D_\theta\|_\mu = 0$ if and only if $\theta = \theta_0$. Given some appropriately consistent estimator $\widehat{D}_\theta$ of D_θ, it is natural to take the $\widehat{\theta} \in \Theta$ that minimizes $\|\widehat{D}_\theta\|_\mu$ as an estimator of θ_0.[8]

A first-step estimator of $Q_{X|Z}$ is needed to construct an estimator of $\widehat{D}_\theta$. For each $k = 1, \ldots, d_x$ let $q_{0,k} \equiv Q_{X_k|Z}$ and suppose that $q_{0,k} \in \mathcal{Q}_k$, a collection of functions from $(0,1) \times \mathcal{Z}$ into $\mathcal{X}_k$ that are weakly increasing in their first argument. Let $\mathcal{Q} = \mathcal{Q}_1 \times \cdots \times \mathcal{Q}_{d_x}$ and write $q(v|z) \equiv (q_1(v_1|z), \ldots, q_{d_x}(v_{d_x}|z))$ for any $q \in \mathcal{Q}$, $v \in (0,1)^{d_x}$ and $z \in \mathcal{Z}$. Then for any $q \in \mathcal{Q}$, define $D_{\theta,q}(t)$ just like $D_\theta(t)$ in (3), except with $\vec{Q}_{X|Z}(t_v \mid Z) \equiv q_0(t_v|Z)$ replaced by $q(t_v|Z)$.

The conditional quantile of X given Z, i.e. q_0, is identified from the observed data and can be estimated with a variety of quantile regression techniques. In Section 4.4, I describe a few such estimators and verify the regularity conditions required of them under the asymptotic framework provided in the next section. For now, just let $\widehat{q} \equiv (\widehat{q}_1, \ldots, \widehat{q}_{d_x})$ be some appropriate first-step estimator of q_0. Given $\widehat{q}$, a feasible estimator of $D_{\theta,\widehat{q}}(t)$ can be constructed as

$$\widehat{D}_{\theta,\widehat{q}}(t) \equiv \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[X_i \leq \widehat{q}(t_v|Z_i), Y_i \leq g_\theta(X_i, t_e), Z_i \leq t_z] - \Big(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[X_i \leq \widehat{q}(t_v|Z_i), Y_i \leq g_\theta(X_i, t_e)]\Big)\Big(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[Z_i \leq t_z]\Big)$$

[6] In (3) and throughout the paper, all inequalities involving vectors should be understood component-wise, i.e. $X \leq x$ if and only if $X_k \leq x_k$ for every $k = 1, \ldots, d_x$.

[7] I use the notation $Q_{X_k|Z}(v_k \mid z) \equiv \inf\{x_k : F_{X_k|Z}(x_k \mid z) \geq v_k\}$.

[8] Other norms of D_θ also yield potential objective functions. For similar problems, Manski (1983) proposed using the sup-norm and Brown and Matzkin (1998) suggested a type of bounded Lipschitz distance. My approach follows Brown and Wegkamp (2002) and Komunjer and Santos (2010) in using an L_2-norm because the projection geometry is useful for the distribution theory.

and $\widehat{\theta}$ can be defined as any $\theta \in \Theta$ that satisfies

$$\|\widehat{D}_{\widehat{\theta},\widehat{q}}\|_\mu \leq \inf_{\theta\in\Theta} \|\widehat{D}_{\theta,\widehat{q}}\|_\mu + o_{\mathbb{P}}(n^{-1/2}), \tag{4}$$

where the additional $o_{\mathbb{P}}(n^{-1/2})$ term allows for inaccuracies such as optimization or numerical error, as long as they are of sufficiently small magnitude.[9] Notice that if one takes $Z_i = X_i$ and $\widehat{q}(t_v|Z_i) = X_i$ for all $t_v \in (0,1)$, then $\widehat{\theta}$ is numerically equivalent to the estimator of Brown and Wegkamp (2002). Their estimator will be inconsistent if X is endogenous.

Implementing the estimator is straightforward in some important cases. For example, suppose that $d_x = 1$ and that $Z \in \{0,1\}$ is binary. This situation arises frequently in natural experiments, and in randomized experiments with partial compliance. A good choice for $\widehat{q}$ in this case is the empirical quantile function.[10] For a given θ, $\|\widehat{D}_{\theta,\widehat{q}}\|_\mu$ is computed by integrating $\widehat{D}_{\theta,\widehat{q}}(t)$ against μ over $\mathcal{T}$. This can be performed numerically in general, but for certain choices of μ, it can be implemented analytically. For example, if $\mu = \text{Unif}(0,1) \times \mu_\varepsilon \times \text{Unif}(\mathcal{Z})$ is a product measure, then it can be shown that

$$\|\widehat{D}_{\theta,\widehat{q}}\|_\mu^2 = \frac{1}{n^2\,|\mathcal{Z}|} \sum_{i=1}^n \sum_{j=1}^n \omega_{ij}(1 - \mu_\varepsilon[g_\theta^{-1}(X_i,Y_i) \vee g_\theta^{-1}(X_j,Y_j)]), \tag{5}$$

where the ω_{ij} terms are defined as

$$\omega_{ij} \equiv (1 - \widetilde{V}_i \vee \widetilde{V}_j) \left(\sum_{z\in\mathcal{Z}} [\mathbb{1}[Z_i \leq z] - \widehat{F}_Z(z)][\mathbb{1}[Z_j \leq z] - \widehat{F}_Z(z)] \right),$$

and where $\widetilde{V}_i \equiv \widehat{F}_{X|Z}(X_i \mid Z_i) - 1/N_{Z_i}$, N_z is the number of observations with $Z_i = z$, and $\widehat{F}_Z(z)$, $\widehat{F}_{X|Z}(x \mid z)$ are empirical and conditional empirical distribution functions. Computing $\|\widehat{D}_{\theta,\widehat{q}}\|_\mu$ does not require numerical integration in this case. The θ that

[9] For consistency this error only needs to be $o_{\mathbb{P}}(1)$. The $n^{-1/2}$ rate is used in the distribution theory.
[10] The definition of the empirical quantile function is provided in Example 1 in Section 4.4.

minimizes $\|\widehat{D}_{\theta,\widehat{q}}\|_\mu$ can be found by using a non-smooth optimization algorithm.[11] In practice, many of the computations, such as the ω_{ij} terms in the above case, do not need to be repeated for each θ. Also, the double summation can be replaced by a single summation after employing a sorting algorithm as in Abrevaya (1999).

4 Asymptotic Theory

Before starting the asymptotic analysis, it will be helpful to examine the structure of the objective function in more detail. First, collect the observable data together into a single vector, $W = (X, Y, Z)$. I refer to realizations of W as $w = (w_x, w_y, w_z)$. Define $A^t_{\theta,q}(w) \equiv \mathbb{1}[w_x \leq q(t_v|w_z), w_y \leq g_\theta(w_x, t_e)]$ and $B^t(w) \equiv \mathbb{1}[w_z \leq t_z]$. Then $D_{\theta,q}(t)$ can also be written as

$$D_{\theta,q}(t) \equiv \mathbb{E}\left[A^t_{\theta,q}B^t\right] - \mathbb{E}[A^t_{\theta,q}]\,\mathbb{E}[B^t] \equiv \mathbb{E}[A^t_{\theta,q}\overline{B}^t], \tag{6}$$

where $\overline{B}^t(w) \equiv B^t(w) - \mathbb{E}(B^t)$.[12] Letting $\mathbb{E}_n$ denote the expectation operator associated with the empirical measure, $\widehat{D}_{\theta,q}(t)$ can similarly be rewritten as

$$\widehat{D}_{\theta,q}(t) \equiv \mathbb{E}_n[A^t_{\theta,q}B^t] - \mathbb{E}_n[A^t_{\theta,q}]\,\mathbb{E}_n[B^t] \equiv \mathbb{E}_n[A^t_{\theta,q}\overline{B}^t_n], \tag{7}$$

where $\overline{B}^t_n(w) \equiv B^t(w) - \mathbb{E}_n(B^t)$.

Comparing (6) and (7) suggests that the behavior of $\widehat{D}$ as an estimator of D will be determined by the properties of the empirical process $\{\mathbb{Q}_n(A^t_{\theta,q}\overline{B}^t) : \theta \in \Theta, q \in \mathcal{Q}, t \in \mathcal{T}\}$, where $\mathbb{Q}_n \equiv \mathbb{E}_n - \mathbb{E}$. If $\widehat{D}$ approximates D well as a function on $\Theta \times \mathcal{Q} \times \mathcal{T}$, then it is reasonable to expect that $\widehat{\theta}$ is a consistent estimator of θ_0. From there, the asymptotic distribution of $\widehat{\theta}$ can be characterized by analyzing the impact on $\widehat{D}_{\theta,q}$ of

[11] The computational results reported in this paper used the `glbDirect` routine in TOMLAB (Holmström et al., 2010) for optimization, which implements the DIRECT algorithm proposed by Jones et al. (1993).

[12] Depending on the context, sometimes I write $B^t(z)$ and $\overline{B}^t(z)$ instead of $B^t(w)$ and $\overline{B}^t(w)$.

small movements in θ near θ_0. The consequences of estimating q_0 by $\widehat{q}$ can be understood by looking at the effect on $\widehat{D}_{\theta,q}$ of small perturbations in q near q_0.

For performing this analysis, another form for $D_{\theta,q}(t)$ is also useful. Using the law of iterated expectations, write

$$\begin{aligned} D_{\theta,q}(t) &= \mathbb{E}\left[\mathbb{P}\left(X \leq q(t_v \mid Z), g_\theta^{-1}(X,Y) \leq t_e \mid Z\right) \overline{B}^t(Z)\right] \\ &= \int_{\mathcal{Z}} F_{X\varepsilon_\theta|Z}(q(t_v \mid z), t_e \mid z) \overline{B}^t(z)\, dF_Z(z). \end{aligned} \tag{8}$$

The second equality here expresses expectation with respect to Z as an integral for notational clarity. Equation (8) is useful for analyzing the derivatives of $D_{\theta,q}(t)$ with respect to θ and q. It can also be written as

$$D_{\theta,q}(t) = \int_{\mathcal{Z}} C_\theta(\vec{F}_{X|Z}(q(t_v|z) \mid z), F_{\varepsilon_\theta|Z}(t_e \mid z); z) \overline{B}^t(z)\, dF_Z(z), \tag{9}$$

where $C_\theta(\cdot,\cdot;z)$ is the copula function of $(X, \varepsilon_\theta)|Z = z$. The following result can be used to simplify this expression further when evaluated at $(\theta, q) = (\theta_0, q_0)$.

Proposition 1. *Given I1, I4 holds if and only if both (i) $\varepsilon \perp\!\!\!\perp Z$ and (ii) the copula function of $(X,\varepsilon)|Z = z$ is equal to the copula function of $(X,\varepsilon)|Z = z'$ for all $z, z' \in \mathcal{Z}$.*

Proposition 1 shows that the copula of $(X,\varepsilon)|Z = z$, say $C_{\theta_0}(\cdot,\cdot)$, does not depend on the realization $Z = z$. This will be used later to simplify the asymptotic variance.

4.1 Consistency

Let $\|\cdot\|$ denote the Euclidean norm on $\mathbb{R}^{d_\theta}$ and let $\|q_k\|_\infty \equiv \sup_{v_k \in (0,1), z \in \mathcal{Z}} |q_k(v_k|z)|$ denote the sup-norm on $\mathcal{Q}_k$. With a slight abuse of notation, I also write $\|q\|_\infty \equiv \sum_{k=1}^{d_x} \|q_k\|_\infty$ for the product norm on $\mathcal{Q}$ and $\|(\theta, q)\| \equiv \|\theta\| + \|q\|_\infty$ for the product norm on $\Theta \times \mathcal{Q}$. Given the identification condition (2), the following assumptions are sufficient for $\widehat{\theta}$ to be consistent.

Assumption C.

C1. $\{W_i\}_{i=1}^n$ *is an independent and identically distributed sample.*

C2. Θ *is compact.*

C3. $\{A^t_{\theta,q} : \theta \in \Theta, q \in \mathcal{Q}\}$ *is Glivenko-Cantelli for every fixed* $t \in \mathcal{T}$.

C4. *For every* x, e *and all* θ', θ, $|g_{\theta'}(x,e) - g_\theta(x,e)| \leq g^{bd}_\Theta(e)\,\|\theta' - \theta\|$ *for some strictly positive* $g^{bd}_\Theta \in L_2(\mu)$.

C5. *There exists a* $f^{bd}_Y \in L_2(F_Z)$ *such that* $\sup_{y,x} f_{Y|XZ}(y \,|\, x, z) \leq f^{bd}_Y(z)$ *for every* z.

C6. *There exists a* $f^{bd}_X \in L_2(F_Z)$ *such that* $\sup_{x_k} f_{X_k|Z}(x_k \,|\, z) \leq f^{bd}_X(z)$ *for each* k *and every* z.

C7. $\|\widehat{q} - q_0\|_\infty \to_{\mathbb{P}} 0$ *and* $\|q_0\|_\infty < \infty$.

Theorem 1. *Under Assumptions I, C and condition* (2), $\widehat{\theta} \to_{\mathbb{P}} \theta_0$.

This theorem exhibits many of the features frequently encountered in extremum estimation, e.g. Newey and McFadden (1994). The key condition is C3 which, given C1, ensures that $\widehat{D}_{\theta,q}$ is consistent for $D_{\theta,q}$ in $L_2(\mu)$, uniformly over $\Theta \times \mathcal{Q}$. Low-level sufficient conditions for C3—actually for the stronger assumption that $\{A^t_{\theta,q} : \theta \in \Theta, q \in \mathcal{Q}\}$ is Donsker for each t—are provided in the next section. Assumptions C4-C6 are used to ensure that $\|D_{\theta,q}\|_\mu$ is appropriately continuous at (θ_0, q_0). Assumption C4 requires the outcome functions to be uniformly parameterized in a smooth way, while C5 and C6 essentially strengthen I1 and I2 to hold uniformly. Given this continuity, C2 is a standard way of ensuring that θ_0 is "uniquely" identified so that no other $\theta \in \Theta$ can come arbitrarily close to minimizing $\|D_{\theta,q_0}\|_\mu$. Assumption C7 reasonably requires the first-step estimator to itself be consistent. The requirement that $\|q_0\|_\infty < \infty$ is a by-product of using the sup-norm to measure consistency of the first-step estimator. While this can be restrictive, it will typically be implied by the low-level sufficient conditions

for asymptotic normality of $\widehat{\theta}$—see Section 4.4.[13]

4.2 Asymptotic Normality

Given the consistency of $\widehat{\theta}$, its asymptotic distribution can be derived by analyzing the behavior of $\|\widehat{D}_{\theta,q}\|_\mu$ near (θ_0, q_0). The approach I take follows that of Pakes and Pollard (1989) and Chen et al. (2003) for non-smooth objective functions and that of Andrews (1994), Newey (1994) and Chen et al. (2003) for two-step semiparametric M-estimators with an infinite-dimensional nuisance parameter (i.e., q) in the first step. The strategy in the non-smooth literature is to look at the effect of small deviations of the parameters on the smooth population objective function, $\|D_{\theta,q}\|_\mu$, rather than the non-smooth sample objective function, $\|\widehat{D}_{\theta,q}\|_\mu$. This is justified if the centered sample objective function, viewed as a stochastic process indexed by parameters, is stochastically equicontinuous.

The impact on $\|D_{\theta_0,q_0}\|_\mu$ of small perturbations in θ are described by the derivative of $D_{\theta,q_0}(t)$ with respect to θ at θ_0, denoted as $\Delta_0(t)$. Using the law of iterated expectations this can be calculated from (8) as

$$\begin{aligned}
\Delta_0(t) &= \nabla_\theta \, \mathbb{E}\left[\mathbb{1}[X \leq q_0(t_v|Z)]\mathbb{P}\left[\varepsilon \leq t_e \mid X, Z\right] \overline{B}^t(Z)\right] \\
&= \mathbb{E}\left[\mathbb{1}[X \leq q_0(t_v|Z)]\nabla_\theta F_{Y|XZ}(g_{\theta_0}(X, t_e) \mid X, Z)\overline{B}^t(Z)\right] \\
&= \mathbb{E}\left[\mathbb{1}[X \leq q_0(t_v|Z)] f_{Y|XZ}(g_{\theta_0}(X, t_e) \mid X, Z)\nabla_\theta g_{\theta_0}(X, t_e)\overline{B}^t(Z)\right]. \qquad (10)
\end{aligned}$$

Small deviations in q, which is an element of an infinite-dimensional space, $\mathcal{Q}$, are

[13] It is possible to relax this restriction by replacing the sup-norm on $\mathcal{Q}$ by the norm

$$\|q\|_\dagger \equiv \left(\int_{\mathcal{T}} \left(\sup_{z \in \mathcal{Z}} \|q(t_v|z)\|\right)^2 d\mu(t)\right)^{1/2}.$$

This norm is weaker than $\|\cdot\|_\infty$, and in particular it is possible to have $\|q_0\|_\dagger < \infty$ even if X has unbounded support. However, in the analysis ahead, it is important to be able to derive a Bahadur representation for $\widehat{q} - q$ under the norm being used, which would be complicated if this norm were $\|\cdot\|_\dagger$ rather than $\|\cdot\|_\infty$.

analyzed using a Fréchet derivative. In the Appendix, the function

$$\Pi_{\theta_0,q_0}^{[q-q_0]}(t) \equiv \int_{\mathcal{Z}} (q-q_0)(t_v|z)' \nabla_x F_{X\varepsilon_{\theta_0}|Z}(q_0(t_v|z), t_e \mid z) \overline{B}^t(z)\, dF_Z(z) \tag{11}$$

is shown to satisfy $\|D_{\theta_0,q_n} - D_{\theta_0,q_0} - \Pi_{\theta_0,q_0}^{[q_n-q_0]}\|_\mu = o(\|q_n - q_0\|_\infty)$ for sequences $q_n \to q_0$. Formally, $\Pi_{\theta_0,q_0}^{[q-q_0]}$ is the $L_2(\mu)$–Fréchet derivative of $D_{\theta_0,q}$ at q_0, in the direction $q - q_0$. The following conditions are sufficient for $\sqrt{n}(\widehat{\theta} - \theta_0)$ to be asymptotically normal and, in particular, ensure that the above calculations are actually valid.

Assumption D.

D1. *θ_0 is in the interior of Θ.*

D2. *$g_\theta(x,e)$ is differentiable with respect to θ for every x, e.*

D3. *$\{A_{\theta,q}^t : \theta \in \Theta, q \in \mathcal{Q}\}$ is Donsker for every fixed $t \in \mathcal{T}$.*

D4. *For some $\mathcal{T}' \subseteq \mathcal{T}$ with $\mu(\mathcal{T}') > 0$, $\{\Delta_0(t) : t \in \mathcal{T}'\}$ is not a proper linear subspace of $\mathbb{R}^{d_\theta}$.*

D5. *$\widehat{q} \in \mathcal{Q}$ with probability approaching 1 and either a) $\|\widehat{q} - q_0\|_\infty = O_{\mathbb{P}}(n^{-1/2})$ or b) $\|\widehat{q} - q_0\|_\infty = o_{\mathbb{P}}(n^{-1/4})$ and each of $\nabla_{x_k} g_\theta(x,e)$, $\nabla_{x_k} F_{Y|XZ}(y \mid x,z)$ and $\nabla_{x_k} f_{X|Z}(x|z)$ exist with $\sup_x |\nabla_{x_k} g_\theta(x,e)| \le \nabla_x g^{bd}(e)$, $\sup_{y,x} |\nabla_{x_k} F_{Y|XZ}(y|x,z)| \le \nabla_x F_Y^{bd}(z)$ and $\sup_x |\nabla_{x_k} f_{X|Z}(x \mid z)| \le \nabla_x f_X^{bd}(z)$ for some $\nabla_x g^{bd} \in L_2(\mu)$, $\nabla_x F_Y^{bd} \in L_2(F_Z)$ and $\nabla_x f_X^{bd} \in L_1(F_Z)$.*

D6. *$\sqrt{n}\Pi_{\theta_0,q_0}^{[\widehat{q}-q_0]}(t) = \sqrt{n}\,\mathbb{E}_n\,\psi(t) + o_{\mathbb{P}}(1)$ for some $\psi(t)$ with $\mathbb{E}\,\psi(t) = 0$ and $\Psi(t,\widetilde{t}) \equiv \mathbb{E}\,\psi(t)\psi(\widetilde{t}) < \infty$ uniformly over $t, \widetilde{t} \in \mathcal{T}$. The $o_{\mathbb{P}}(1)$ term is uniform over $t \in \mathcal{T}$.*

Theorem 2. *Under the assumptions of Theorem 2 together with Assumptions D,*

$\sqrt{n}(\widehat{\theta} - \theta_0) \rightsquigarrow N(0, \overline{\Delta}_0^{-1}\overline{\Sigma}_0\overline{\Delta}_0^{-1})$, *where* $\overline{\Delta}_0 \equiv \int_{\mathcal{T}} \Delta_0(t)\Delta_0(t)'\, d\mu(t)$ *and*

$$\begin{aligned}
\overline{\Sigma}_0 &\equiv \int_{\mathcal{T}\times\mathcal{T}} \Delta_0(t)\Delta_0(\widetilde{t})'[\sigma(t,\widetilde{t}) + \nu(t,\widetilde{t})]\, d\mu(t)\, d\mu(\widetilde{t}), \\
\textit{with} \quad \sigma(t,\widetilde{t}) &\equiv \Big[C_{\theta_0}(t_v \wedge \widetilde{t}_v, F_{\varepsilon_{\theta_0}}(t_e \wedge \widetilde{t}_e)) - C_{\theta_0}(t_v, F_{\varepsilon_{\theta_0}}(t_e))C_{\theta_0}(\widetilde{t}_v, F_{\varepsilon_{\theta_0}}(\widetilde{t}_e))\Big] \\
&\qquad \times \left[F_Z(t_z \wedge \widetilde{t}_z) - F_Z(t_z)F_Z(\widetilde{t}_z)\right] \qquad (12) \\
\textit{and} \quad \nu(t,\widetilde{t}) &\equiv \Psi(t,\widetilde{t}) + \mathbb{E}\left[A^t_{\theta_0,q_0}\overline{B}^t\psi(\widetilde{t}) + A^{\widetilde{t}}_{\theta_0,q_0}\overline{B}^{\widetilde{t}}\psi(t)\right].
\end{aligned}$$

Assumptions D1 and D2 are standard and needed for the existence of $\Delta_0(t)$ in (10). Assumption D4 is a weak regularity assumption that requires Δ_0, as a function on $\mathcal{T}$, to carry some information about each component of θ. Assumption D5 requires $\Pi^{[q-q_0]}_{\theta_0,q_0}$ to be an $O_{\mathbb{P}}(n^{-1/2})$ approximation of $\|D_{\theta_0,q}\|_\mu$. This will occur if $\widehat{q}$ converges to q_0 at the parametric rate, which will typically be the case if Z is discretely distributed or if a parametric model is used to estimate q_0. If the additional smoothness conditions of D5 b) hold then $\Pi^{[q-q_0]}_{\theta_0,q_0}$ is improved to an $O(\|q-q_0\|^2_\infty)$ approximation of $\|D_{\theta_0,q}\|_\mu$ near q_0, which allows the required rate of convergence of $\|q-q_0\|_\infty$ to be reduced to $n^{-1/4}$. This is the well-known semiparametric rate discussed by Newey (1994) and others. It can be attained by nonparametric smoothing estimators of q_0 when Z is continuous. Given this rate of convergence, D6 further requires that $\Pi^{[\widehat{q}-q_0]}_{\theta_0,q_0}(t)$ has an asymptotically linear (or Bahadur) representation. Since $\Pi^{[q-q_0]}_{\theta_0,q_0}(t)$ is linear in $q-q_0$, this property will follow from a Bahadur representation for $(\widehat{q}-q_0)(t_v|z)$—see (11). These are commonplace in the literature, e.g. Chapter 4 of Koenker (2005).

An important component of Theorem 2 is D3. Like C3, this is a high-level condition that ensures the empirical process indexed by $\{A^t_{\theta,q} : \theta \in \Theta, q \in \mathcal{Q}\}$ is weakly convergent for every t. Following Brown and Wegkamp (2002), sufficient conditions for D3 can be derived by exploiting the structure of $A^t_{\theta,q}$. Specifically, $A^t_{\theta,q}$ is an indicator function for the intersection of the subgraphs of g_θ and q, i.e. $A^t_{\theta,q}(w) = \mathbb{1}[w_y \leq g_\theta(w_x, t_e)]\mathbb{1}[w_x \leq q(t_v|w_z)]$. The entropy of collections of indicator functions like these can be controlled

by assuming that the boundary of the subgraph is sufficiently smooth with respect to the index.

Proposition 2. *Assumption D3 (hence C3) is satisfied if the collection* $\{g_\theta : \theta \in \Theta\}$ *satisfies either*

1a. *For every* $e \in \mathcal{E}$ *there exists an integer* J^e *and functions* $\{\beta_j^e\}_{j=1}^{J^e}$ *such that for every* $\theta \in \Theta$ *there is an* $\alpha_\theta^e \in \mathbb{R}^{J^e}$ *with* $g_\theta(x,e) = \sum_{j=1}^{J^e} \alpha_{\theta,j}^e \beta_j^e(x)$.

1b. $\mathcal{X}$ *is bounded,* f_{YX} *is uniformly bounded and for every* $e \in \mathcal{E}$, $\{g_\theta(\cdot,e) : \theta \in \Theta\}$ *is a subset of the Hölder ball of order* γ_Θ, *denoted* $\mathcal{C}^{\gamma_\Theta}(\mathcal{X})$, *with* $\gamma_\Theta > d_x$.

and if every one of the collections $\mathcal{Q}_k, k = 1, \ldots, d_x$ *satisfies either*

2a. *For every* $v \in (0,1)$ *there exists an integer* J^v *and functions* $\{\beta_j^v\}_{j=1}^{J^v}$ *such that for every* $q_k \in \mathcal{Q}_k$ *there is an* $\alpha_q^v \in \mathbb{R}^{J^v}$ *with* $q_k(v|z) = \sum_{j=1}^{J^v} \alpha_{q,j}^v \beta_j^v(z)$.

2b. $\mathcal{Z}$ *is bounded,* f_{XZ} *is uniformly bounded and for every* $v \in (0,1)$, $\{q_k(v|\cdot) : q_k \in \mathcal{Q}_k\} \subseteq \mathcal{C}^{\gamma_{\mathcal{Q}_k}}(\mathcal{Z})$ *with* $\gamma_{\mathcal{Q}_k} > d_z$.

Condition 1a means that $\{g_\theta(\cdot,e) : \theta \in \Theta\}$ is a subset of a finite-dimensional vector space of functions for each fixed e. This is a widely used sufficient condition in empirical process theory, see e.g. Pollard (1984) or van der Vaart and Wellner (1996). It ensures that the collection of subgraphs corresponding to $\{g_\theta(\cdot,e) : \theta \in \Theta\}$ is a Vapnik-Červonenkis–(VC–)class. The collection of indicator functions for a VC-class satisfies the Donsker property. Condition 1b takes a different approach and directly bounds the bracketing number of the collection of indicator functions. This comes from well-known bounds on the bracketing numbers for collections of functions satisfying the smoothness requirements in that assumption, as well as bounds on the density, which ensure that the probability mass is smoothly spread out.[14] That the conditions in Proposition 2 can depend on each e, for $\{g(\cdot,e) : \theta \in \Theta\}$, and on each v, for $\{q_k(v|\cdot) : q_k \in \mathcal{Q}_k\}$, adds some additional flexibility. It is possible because of the monotonicity of these functions

[14] These density bounds are nearly redundant given C5 and C6.

in e and v. Note that 2a is always satisfied when Z is discretely distributed with finite support.

The asymptotic variance given in (12) is very complicated. It would be difficult to construct a feasible estimator of this quantity. However, as I show in the next section, the bootstrap can be used to perform inference. It is also difficult to gain much intuition from the form of the asymptotic variance, although a few things can be said. First, the contribution of the first-step estimator is captured by $\nu(t, \widetilde{t})$. This term itself depends on a complicated interaction between the components of the criterion function, i.e. the functions $A^t_{\theta_0,q_0}$ and $\overline{B}^t$, and the influence function for the Fréchet derivative, $\psi(t)$. The latter depends on the model used for $q_0 \equiv Q_{X|Z}$, although for the examples considered in Section 4.4 I have not found that specifying the model provides any useful simplification.

If the distribution of $X|Z$ were known and did not need to be estimated then the limiting variance would be determined by Δ_0, σ and μ. As discussed, the first of these is the pointwise derivative of $D_{\theta,q}$ at (θ_0, q_0) and as such captures the amount of local information about θ in the underlying data generating process. The second is the covariance function for the limiting process of $\{\sqrt{n}\widehat{D}_{\theta_0,q_0}(t) : t \in \mathcal{T}\}$, which corresponds to the identification condition (2). The last ingredient, μ, is a measure chosen by the analyst. In principle, it should be possible to determine an optimal choice of μ and then construct a data-driven procedure for implementing it, as in GMM. However this appears to be quite difficult, and the work of Carrasco and Florens (2000) suggests that the resulting procedure would be an ill-posed inverse problem requiring careful regularization. It seems appropriate to leave this problem for future research.

4.3 Bootstrap

Let $\{W^\star_i\}^n_{i=1}$ denote a nonparametric bootstrap sample drawn with replacement from $\{W_i\}^n_{i=1}$. That is, $\{W^\star_i\}^n_{i=1}$ are independently and identically distributed according to

the empirical measure $\mathbb{P}_n$, conditional on the realizations $\{W_i\}_{i=1}^n$. Define

$$D^\star_{\theta,q}(t) \equiv \frac{1}{n}\sum_{i=1}^n \mathbb{1}[X_i^\star \leq q(t_v|Z_i^\star), Y_i^\star \leq g_\theta(X_i^\star, t_e), Z_i^\star \leq t_z] \\ - \left[\frac{1}{n}\sum_{i=1}^n \mathbb{1}[X_i^\star \leq q(t_v|Z_i^\star), Y_i^\star \leq g_\theta(X_i^\star, t_e)]\right]\left[\frac{1}{n}\sum_{i=1}^n \mathbb{1}[Z_i^\star \leq t_z]\right]$$

as the bootstrap counterpart to $\widehat{D}_{\theta,q}(t)$. Next, let $q^\star$ be an estimate of q_0 using $\{W_i^\star\}_{i=1}^n$ and take $\theta^\star$ to be any $\theta \in \Theta$ that satisfies

$$\|D^\star_{\theta^\star,q^\star}\|_\mu \leq \inf_{\theta\in\Theta} \|D^\star_{\theta,q^\star}\|_\mu + o_{\mathbb{P}^\star}(n^{-1/2}). \tag{13}$$

In (13), $\mathbb{P}^\star = \mathbb{P}[\cdot \mid \{W_i\}_{i=1}^n]$ denotes the population measure, conditional on the data, and the inequality is meant to hold for almost any realization of the data.[15] Note that the criterion in (13) is not re-centered around a quantity based on $\widehat{\theta}$, since this is not necessary for simply approximating the asymptotic distribution of $\widehat{\theta}$ through simulation (Hahn, 1996).

Giné and Zinn (1990) show that the weak convergence of empirical processes extends quite generally to conditional weak convergence of bootstrapped empirical processes. The next theorem leverages these results to show that $\sqrt{n}(\theta^\star - \widehat{\theta})$ converges weakly under $\mathbb{P}^\star$ to the limiting distribution of $\sqrt{n}(\widehat{\theta} - \theta_0)$ with $\mathbb{P}$–probability approaching 1. Let $\mathbb{P}_n^\star$ denote the empirical measure of the bootstrap sample and $\mathbb{E}_n^\star$ the expectation operator with respect to $\mathbb{P}_n^\star$.

Assumption D*. *For almost every realization of* $\{W_i\}_{i=1}^\infty$,

D5*. $q^\star \in \mathcal{Q}$ *with* $\mathbb{P}^\star$*–probability approaching* 1 *and either a)* $\|q^\star - \widehat{q}\|_\infty = O_{\mathbb{P}^\star}(n^{-1/2})$ *or b)* $\|q^\star - \widehat{q}\|_\infty = o_{\mathbb{P}^\star}(n^{-1/4})$ *and the smoothness conditions in D5 b) hold.*

D6*. $\sqrt{n}\Pi^{[q^\star - \widehat{q}]}_{\theta_0,q_0}(t) = \sqrt{n}(\mathbb{E}_n^\star - \mathbb{E}_n)\psi(t) + o_{\mathbb{P}^\star}(1)$ *where* $\psi(t)$ *is the same influence function as in D6 and the* $o_{\mathbb{P}^\star}(1)$ *term is uniform over* $\mathcal{T}$.

[15] As is common in the literature, the dependence of $\mathbb{P}^\star$ on n is suppressed in the notation. The symbol $\mathbb{P}_n^\star$, introduced ahead, denotes the empirical measure with respect to the bootstrap sample.

Theorem 3. *Under the assumptions of Theorem 2 together with Assumptions D*,* $\sqrt{n}(\theta^\star - \widehat{\theta}) \rightsquigarrow N(0, \overline{\Delta}_0^{-1}\overline{\Sigma}_0\overline{\Delta}_0^{-1})$ *with respect to* $\mathbb{P}^\star$*, with* $\mathbb{P}$*–probability approaching* 1.

Assumptions D5* and D6* are bootstrap counterparts to D5 and D6 and can be expected to hold in most circumstances. Giné and Zinn (1990) note that Theorem 3 can be used to approximate the asymptotic distribution of $\sqrt{n}(\widehat{\theta}-\theta_0)$ through Monte Carlo simulation by drawing bootstrap samples $\{\{W_{bi}^\star\}_{i=1}^n, b = 1, \ldots, B\}$ with replacement from $\{W_i\}_{i=1}^n$ for a large integer B. For each b, one uses $\{W_{bi}^\star\}_{i=1}^n$ to estimate $q_b^\star$ and compute $\theta_b^\star$ from (13). The sample quantiles of $\{\theta_b^\star\}_{b=1}^B$ can then be used to construct confidence intervals for θ_0 that have the correct size asymptotically.

4.4 First-Step Quantile Estimators

This section contains three examples which show that Assumptions D5 and D6 are broadly applicable.

Example 1 (**Empirical Conditional Quantile Function**)**.** When $\mathcal{Z}$ is a finite set, an attractive choice for $\widehat{q}_k(v_k|z)$ is the empirical quantile function, defined as the j^{th} order statistic of $\widehat{\mathcal{X}}_{k,z}$ for $v_k \in (\frac{j-1}{N_{k,z}}, \frac{j}{N_{k,z}}]$, where $\widehat{\mathcal{X}}_{k,z} = \{X_{i,k} : Z_i = z\}$ has $N_{k,z} > 0$ elements. It is well known that $\|\widehat{q}_k - q_{0,k}\|_\infty = O_{\mathbb{P}}(n^{-1/2})$ with $\sqrt{n}(\widehat{q}_k - q_{0,k})(v_k|z) = \sqrt{n}\,\mathbb{E}_n\,\phi_k(v_k|z) + o_{\mathbb{P}}(1)$ where

$$\phi_k(v_k|z)(w) \equiv \frac{v_k - \mathbb{1}[w_{x_k} \leq q_{0,k}(v_k|z), w_z = z]/\,\mathbb{P}[Z = z]}{f_{X_k|Z}(q_{0,k}(v_k|z) \mid z)}$$

has mean zero under $\mathbb{P}$. The remainder is uniform over $\mathcal{Z}$ due to its finiteness, and uniform over t_v when $X_k|Z = z$ is supported on a compact interval and has a density that is continuous and bounded away from 0.[16] Let $\phi(v|z) \equiv (\phi_1(v_1|z), \ldots, \phi_{d_x}(v_{d_x}|z))$.

[16] See, for example, Lemma 21.4 and Corollary 21.5 of van der Vaart (1998).

Then $\sqrt{n}\Pi_{\theta_0,q_0}^{[\widehat{q}-q_0]}(t) = \sqrt{n}\,\mathbb{E}_n\,\psi(t) + o_{\mathbb{P}}(1)$ with

$$\psi(t) = \int_{\mathcal{Z}} \phi(t_v|z)'\nabla_x F_{XU_{\theta_0}|Z}(q_0(t_v|z), t_e \mid z)\overline{B}^t(z)\, dF_Z(z).$$

An application of Fubini's Theorem shows that $\mathbb{E}\,\psi(t) = 0$ as required by D6. Conditions D5* and D6* can be verified using similar arguments and the results in Section 5 of Bickel and Freedman (1981).

Example 2 (**Linear Quantile Regression**)**.** Suppose that $q_{0,k}(v_k|Z) = Z'\beta_k(v_k)$ for each k, where Z is a random d_z–vector and $\beta_k(v_k) \in \mathbb{R}^{d_z}$ for each v_k, as in the celebrated linear quantile regression model of Koenker and Bassett (1978). Let $\widehat{\beta}_k(v_k)$ denote the linear quantile regression estimator. If the density of $X_k|Z$ is uniformly bounded away from 0 and the support of Z is bounded then $\|\widehat{q}_k - q_{0,k}\|_\infty = O(\sup_{v_k\in(0,1)} \|(\widehat{\beta}_k - \beta_k)(v_k)\|) = O_{\mathbb{P}}(n^{-1/2})$. In addition,

$$\sqrt{n}(\widehat{\beta}_k - \beta_k)(v_k) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \Xi_k^{-1} Z_i \left(v_k - \mathbb{1}[X_{i,k} \le Z_i'\beta_k(v_k)]\right) + o_{\mathbb{P}}(1),$$

where $\Xi_k \equiv \mathbb{E}\left[f_{X_k|Z}(Z'\beta_k(v_k) \mid Z)ZZ'\right]$ and the remainder is uniform over $v_k \in (0,1)$. Verification of D6 then proceeds as in Example 1. Conditions D5* and D6* can be verified using similar arguments and the results in Appendix F of Chernozhukov et al. (2009).

Example 3 (**Kernel Smoothing Quantile Regression**)**.** Nonparametric smoothing techniques can also be used to estimate the first stage quantile regression when Z is continuously distributed. For example, when $d_x = d_z = 1$, a local polynomial estimator $\widehat{q}(v|z)$ can be constructed by performing a polynomial quantile regression on the data $\{(X_i, Z_i) : Z_i \in [z-h, z+h]\}$, where $h > 0$ is a bandwidth parameter that tends to 0 as the sample size increases. Chaudhuri et al. (1997, Lemma 4.1) established Bahadur representations for general estimators of this type. These estimators exhibit rates of convergence that depend on d_z and the assumed smoothness of $q_0(v|z)$ as a function of

z. The rate is slower than $\sqrt{n}$, but can be faster than $n^{1/4}$ and thus satisfy D5. Despite this, $\Pi_{\theta_0,q_0}^{[\widehat{q}-q_0]}$ will typically converge at the $\sqrt{n}$–rate because it is an average over Z and hence depends on all n observations—see Newey (1994) or Chaudhuri et al. (1997).

4.5 Covariates

It is possible to include observed covariates, say $\widetilde{X}$, into outcome equation (1), so that $Y = g_{\theta_0}(X, \widetilde{X}, \varepsilon_{\theta_0})$. With appropriate modifications of Assumption I, the identification condition (2) becomes $(V, \varepsilon_\theta) \perp\!\!\!\perp Z|\widetilde{X}$ if and only if $\theta = \theta_0$, where now $V_k \equiv F_{X_k|\widetilde{X},Z}(X_k \mid \widetilde{X}, Z)$. The criterion function is adjusted by replacing (3) with

$$\begin{aligned} D_{\theta,q_0}(t) = \mathbb{P}\left[X \leq \vec{Q}_{X|\widetilde{X},Z}(t_v \mid t_{\widetilde{x}}, Z), Y \leq g_\theta(X, t_{\widetilde{x}}, t_e), Z \leq t_z \;\middle|\; \widetilde{X} = t_{\widetilde{x}}\right] \\ - \mathbb{P}\left[X \leq \vec{Q}_{X|\widetilde{X},Z}(t_v \mid t_{\widetilde{x}}, Z), Y \leq g_\theta(X, t_{\widetilde{x}}, t_e) \;\middle|\; \widetilde{X} = t_{\widetilde{x}}\right] \mathbb{P}\left[Z \leq t_z \;\middle|\; \widetilde{X} = t_{\widetilde{x}}\right], \end{aligned} \tag{14}$$

where $t \equiv (t_v, t_e, t_z, t_{\widetilde{x}}) \in \mathcal{T}$ and $\mathcal{T}$ now also covers the support of $\widetilde{X}$. If this support is finite, then it is straightforward to estimate $D_{\theta,q_0}(t)$ by conditioning on $\widetilde{X}$ in the first step and in the empirical measure $\mathbb{P}[\cdot \mid \widetilde{X}]$. The previous asymptotic analysis extends readily.

If some component of $\widetilde{X}$ is continuous then there are at least two reasonable approaches that could be taken. First, one could use a sieve to estimate $\mathbb{P}[\cdot \mid \widetilde{X}]$, as in Chen and Pouzo (2012) for conditional moment models. Alternatively, after rewriting (14) as

$$\begin{aligned} D_{\theta,q_0}(t) = \mathbb{E}\Big[\mathbb{1}[X \leq \vec{Q}_{X|\widetilde{X},Z}(t_v \mid \widetilde{X}, Z)]\mathbb{1}[Y \leq g_\theta(X, \widetilde{X}, t_e)] \\ \times \Big(\mathbb{1}[Z \leq t_z] - F_{Z|\widetilde{X}}(t_z \mid \widetilde{X})\Big) \mid \widetilde{X}\Big], \end{aligned}$$

one could use the observation of Dominguez and Lobato (2004) that this conditional

moment is equal to zero a.s. for a.e. $t = (t_v, t_e, t_z)$ if and only if

$$\widetilde{D}_{\theta,q_0}(t) \equiv \mathbb{E}\Big[\mathbb{1}[X \leq \vec{Q}_{X|\widetilde{X},Z}(t_v \mid \widetilde{X}, Z)]\mathbb{1}[Y \leq g_\theta(X, \widetilde{X}, t_e)] \\ \times \Big(\mathbb{1}[Z \leq t_z] - F_{Z|\widetilde{X}}(t_z \mid \widetilde{X})\Big)\mathbb{1}[\widetilde{X} \leq t_{\widetilde{x}}]\Big] = 0$$

for a.e. $t = (t_v, t_e, t_z, t_{\widetilde{x}})$. The asymptotic analysis in this approach would be similar to the case without covariates, except that now there is an additional infinite-dimensional nuisance parameter, $F_{Z|\widetilde{X}}$, that needs to be accounted for.

5 Monte Carlo

I conducted several Monte Carlo experiments to examine the finite sample performance of $\widehat{\theta}$ and of the bootstrap approximation to its limiting distribution. The data generating process (DGP) for these simulations was chosen to roughly approximate some summary statistics for the 1979 National Longitudinal Survey of Young Men (NLS) data used in the next section. The outcome equation is given by

$$g_\theta(x, e) = e + \theta_1(x - \overline{x})e + \theta_2(x - \overline{x}) + (\theta_3/10)(x^2 - \overline{x}^2), \tag{15}$$

where $\overline{x}$ is a known constant. This specification is like a Mincer equation that allows for both nonlinearity and unobserved heterogeneity in the returns to education. The centering around $\overline{x}$ implies I3 (Matzkin, 2003). I take $\overline{x} = 14$, which is roughly the unconditional median of X under the first-stage specification below. The marginal distribution of ε_{θ_0} is $N(6, .40^2)$ and $\theta_0 = (.05, .05, -.12)$.

The first stage equation is given by $X = 22L(\gamma Z + \eta)$, where $L(a) = (1 + e^{-a})^{-1}$. This specification restricts X to lie in $[0, 22]$ and ensures that the relevance condition required for (2) is satisfied for a variety of choices for the marginal distribution of Z. The strength of the instrument can be augmented by γ, with $\gamma = 0$ corresponding to a

completely irrelevant instrument. The marginal distribution of η is taken as $N(.40, .6^2)$. To add endogeneity to the outcome equation, I let the joint distribution of $(\varepsilon_{\theta_0}, \eta)$ be characterized by a Frank copula, i.e.

$$F_{\varepsilon_{\theta_0}\eta}(e,n) = -\frac{1}{\lambda}\log\left(1 + \frac{(\exp[-\lambda F_{\varepsilon_{\theta_0}}(e)] - 1)(\exp[-\lambda F_{\eta}(n)] - 1)}{\exp(-\lambda) - 1}\right), \qquad (16)$$

where $F_{\varepsilon_{\theta_0}}$ and F_η are the normal marginal distributions previously described and λ controls the degree of dependence. I let $\lambda = 2$, which creates a strong nonlinear dependence between ε_{θ_0} and η and hence between ε_{θ_0} and X. If instead of (16), $F_{\varepsilon_{\theta_0}\eta}(e,n) = F_{\varepsilon_{\theta_0}}(e)F_\eta(n)$, i.e. $\varepsilon_{\theta_0} \perp\!\!\!\perp \eta$, then the estimator in this paper would be unnecessary and g_{θ_0} could be consistently estimated using the simpler estimator of Brown and Wegkamp (2002).

Table 1 contains the results of the Monte Carlo for three different choices of the marginal distribution of Z, two choices of instrument strength, γ, and samples of size 400, 800 and 1600. The number of replications is set at 500 throughout. The integrating measure was taken to be $\mu = \text{Unif}(0,1) \times N(\overline{\varepsilon}, \sigma_\varepsilon^2) \times \text{Unif}(\mathcal{Z})$, where $\overline{\varepsilon}$ and σ_ε^2 are chosen conservatively such that the normal component places substantial mass on a data-determined approximation of the support of ε_{θ_0}.[17] The distributions of Z are chosen to place equal mass on $2, 4$ or 8 points of support and to satisfy $\mathbb{E}\, Z = 0$ with $\text{Var}\, Z = 1/4$.[18] As expected, mean-square errors decrease by roughly a factor of two as the sample size is doubled and decrease unambiguously when the instrument is augmented by a larger value of γ. An interesting finding is that the performance of the estimator deteriorates as the number of support points of Z increases. The likely explanation is that any potential benefits gained through additional identifying content are outweighed by higher variance in the first stage estimator.

I also verified that the bootstrap procedure in Section 4.3 works as intended. This

[17] In particular, $\{g_\theta^{-1}(\overline{x}, Y_i)\}_{i=1}^n$, which does not depend on θ, is used as a guide to the support of ε_{θ_0}. Then $\overline{\varepsilon}$ is taken as the midpoint of this support and σ_ε is taken equal to its length.

[18] Specifically, the supports are $\{\pm 1/2\}$, $\{\pm\sqrt{7}/4, \pm 1/4\}$ and $\{\pm\sqrt{29}/8, \pm 5/8, \pm 3/8, \pm 1/8\}$.

is a computationally demanding exercise, so I assume that only θ_2 is unknown to the analyst, which greatly speeds up the optimization. With this simplification, it becomes feasible to bootstrap 1000 times on each of 500 replications. Table 2 shows that the nominal and actual coverage probabilities for bootstrap confidence intervals of θ_2 are similar for sample sizes of $100, 200$ and 400.

6 Empirical Illustration

In this section, I illustrate the use of the estimator by estimating the returns to schooling using the extract of the 1979 NLS considered by Card (1995). In this setting, Y is log wage, X is years of schooling completed, and Z is an indicator variable for whether an individual grew up near an accredited four-year college.[19] The sample is restricted to the $N = 2{,}946$ men with at least 8 years of completed schooling. See Card (1995) for a complete description of the data.

I estimated four versions of the outcome equation (15) used in the Monte Carlo simulations with the same choice of μ used there. The estimated marginal effects of years of education on log wages that result from these specifications are shown in Table 3, along with 95% bootstrapped confidence regions constructed from $B = 500$ replications. For comparison, the column labeled "OLS" reports the coefficient on X_i in an ordinary linear regression, "OLS^2" reports the estimated marginal effects of schooling on log wages when X_i^2 is added to this regression, and "IV" reports the coefficient on X_i from the instrumental variables estimator that uses Z_i as an instrument for X_i. These comparison regressions exhibit the common and somewhat puzzling result that instrumenting for schooling actually increases the estimated marginal effect.

In specification (1) of (15), I restrict $\theta_1 = 0$ and $\theta_3 = 0$, so that the model is linear and separable as in the usual linear IV specification. The estimated marginal effects in

[19] To mitigate concerns about schooling being discrete (which would violate I1), I smooth it by adding an idiosyncratic Unif[-.1, .1] noise term to each X_i. The results are not very sensitive to the magnitude of this noise term.

the two specifications are roughly the same and the confidence intervals are of a similar length although slightly shifted. In specification (2), I restrict $\theta_1 = 0$, which renders the model separable, but allows for both a linear and quadratic endogenous term, even with only a binary instrument. The results for this specification suggest strongly decreasing marginal returns to schooling, although the estimates are still much larger than in the OLS^2 regression. Specification (3) restricts $\theta_3 = 0$, which allows for unobservable heterogeneity in marginal effects. The results here suggest that agents at the upper end of the log wage distribution have larger marginal returns to schooling, although the difference is not large. In specification (4), θ_1, θ_2 and θ_3 are estimated simultaneously, thereby allowing for both nonlinearity and heterogeneity in marginal effects. Patterns similar to those in specifications (2) and (3) are evident here as well. Specifically, the marginal returns to schooling are decreasing in schooling and increasing in quantiles.

7 Conclusion

In this paper, I proposed and analyzed a minimum distance from independence estimator for the instrumental variables model studied in Torgovitsky (2015). The model allows for both unobserved heterogeneity and endogeneity, and was shown in that paper to be identified under low-level conditions even if the available instruments are only binary. I established that the estimator is consistent and asymptotically normal under relatively weak regularity conditions, and I verified the consistency of the bootstrap. I reported the results of Monte Carlo simulations that support the predictions of the asymptotic theory. An empirical illustration demonstrated some notable features of the estimator in the context of estimating the returns to schooling with Card's (1995) geographic location instrument.

A Proofs

Proof of Proposition 1. Suppose that I4 holds. Then $\varepsilon \perp\!\!\!\perp Z$ and $F_{X_k|Z}(h_k(z,n_k)|z) = \mathbb{P}\left[h_k(z,\eta_k) \leq h_k(z,n_k) \mid Z=z\right] = \mathbb{P}\left[\eta_k \leq n_k \mid Z=z\right] = F_{\eta_k}(n_k)$ for any $n \in \mathbb{R}^{d_x}$ and each k. Let $\vec{F}_\eta(n) \equiv (F_{\eta_1}(n_1), \ldots, F_{\eta_{d_x}}(n_{d_x}))$. Then Sklar's Theorem implies that for any n, e and $z \in \mathcal{Z}$,

$$\begin{aligned} \mathbb{P}\left[\varepsilon \leq e, \eta \leq n \mid Z=z\right] &= \mathbb{P}\left[\varepsilon \leq e, X_k \leq h_k(z,n_k)\ \forall k \mid Z=z\right] \\ &= C(F_{\varepsilon|Z}(e \mid z), \vec{F}_\eta(n); z) = C(F_\varepsilon(e), \vec{F}_\eta(n); z). \end{aligned} \tag{17}$$

By hypothesis, the left-hand side of (17) does not depend on z, so $C(F_\varepsilon(e), \vec{F}_\eta(n); z) = C(F_\varepsilon(e), \vec{F}_\eta(n); z')$. Given I1, this implies $C(\cdot,\cdot;z) = C(\cdot,\cdot;z')$ for every z, z'.

Conversely, suppose that (i) and (ii) in the statement of the proposition are satisfied. Let $\eta_k \equiv F_{X_k|Z}(X_k \mid Z)$ and $h_k(z,\cdot) \equiv Q_{X_k|Z}(\cdot \mid z)$ for every $z \in \mathcal{Z}$. Then $h_k(z,\cdot)$ is strictly increasing, $h_k(Z,\eta_k) = Q_{X_k|Z}(F_{X_k|Z}(X_k \mid Z) \mid Z) = X_k$ and

$$\begin{aligned} \mathbb{P}\left[\varepsilon \leq e, \eta \leq n \mid Z=z\right] &= \mathbb{P}\left[\varepsilon \leq e, X_k \leq Q_{X_k|Z}(n_k \mid z)\ \forall k \,\middle|\, Z=z\right] \\ &= C(F_{\varepsilon|Z}(e \mid z), n; z) = C(F_\varepsilon(e), n; z). \end{aligned}$$

By hypothesis, the right-hand side does not depend on z, so $(\eta, \varepsilon) \perp\!\!\!\perp Z$. *Q.E.D.*

A.1 Consistency

Three lemmas are used in the proof of Theorem 1. Lemmas 1 and 2 establish continuity of the criterion function with respect to θ and q, respectively. Lemma 3 establishes the uniform consistency of $\widehat{D}_{\theta,q}$ for $D_{\theta,q}$.

Lemma 1. *Under C4 and C5, $\|D_{\theta,q}\|_\mu$ is continuous in θ for any q.*

Proof of Lemma 1. For any θ', θ and t,

$$\begin{aligned}|D_{\theta',q}(t) - D_{\theta,q}(t)| &\leq |\mathbb{E}\left(\mathbb{1}[Y \leq g_{\theta'}(X,t_e)] - \mathbb{1}[Y \leq g_{\theta}(X,t_e)]\right)| \\ &= \left|\mathbb{E}\left[F_{Y|X}(g_{\theta'}(X,t_e) \mid X) - F_{Y|X}(g_{\theta}(X,t_e) \mid X)\right]\right| \\ &\leq \overline{f}_Y^{bd}\mathbb{E}\left(|g_{\theta'}(X,t_e) - g_{\theta}(X,t_e)| \mid X\right) \leq \overline{f}_Y^{bd} g_\Theta^{bd}(t_e)\|\theta' - \theta\|,\end{aligned}$$

where the second line follows from the law of iterated expectations and bounds are obtained from C5 and C4 with $\overline{f}_Y^{bd} \equiv \mathbb{E}\, f_Y^{bd}(Z)$. Hence $|\|D_{\theta',q}\|_\mu - \|D_{\theta,q}\|_\mu| \leq \|D_{\theta',q} - D_{\theta,q}\|_\mu \leq \overline{f}_Y^{bd}\|g_\Theta^{bd}\|_\mu\|\theta' - \theta\|$, which implies continuity of $\|D_{\theta,q}\|_\mu$ in θ. *Q.E.D.*

Lemma 2. *Under C6,* $\|D_{\theta,q} - D_{\theta,q_0}\|_\mu = O(\|q - q_0\|_\infty)$ *uniformly in* θ.

Proof of Lemma 2. Theorem 2.10.7 of Nelsen (2006) shows that any copula, C, is Lipschitz with respect to the rectilinear distance with Lipschitz constant 1, i.e. $|C(v^a, s) - C(v^b, s)| \leq \sum_{k=1}^{d_x} \left|v_k^a - v_k^b\right|$ for any $v^a, v^b \in [0,1]^{d_x}$ and $s \in [0,1]$. Using this property, it follows from (9) that

$$\begin{aligned}|D_{\theta,q}(t) - D_{\theta,q_0}(t)| &\leq \int_{\mathcal{Z}} \Big|C_\theta(\vec{F}_{X|Z}(q(t_v|z) \mid z), F_{\varepsilon_\theta|Z}(t_e \mid z); z) \\ &\qquad - C_\theta(\vec{F}_{X|Z}(q_0(t_v|z) \mid z), F_{\varepsilon_\theta|Z}(t_e \mid z); z)\Big| \, dF_Z(z) \\ &\leq \int_{\mathcal{Z}} \sum_{k=1}^{d_x} \left|F_{X_k|Z}(q_k(t_{v_k}|z) \mid z) - F_{X_k|Z}(q_{0,k}(t_{v_k}|z) \mid z)\right| \, dF_Z(z) \\ &\leq \int_{\mathcal{Z}} f_X^{bd}(z) \sum_{k=1}^{d_x} |q_k(t_{v_k}|z) - q_{0,k}(t_{v_k}|z)| \, dF_Z(z) \leq \overline{f}_X^{bd} \left\|q - q_0\right\|_\infty,\end{aligned}$$

where the third inequality uses C6 with $\overline{f}_X^{bd} \equiv \mathbb{E}\, f_X^{bd}(Z)$. *Q.E.D.*

Lemma 3. *Under C1 and C3,* $\widehat{D}_{\theta,q}$ *converges almost surely to* $D_{\theta,q}$ *in* $L_2(\mu)$ *uniformly over* $\Theta \times \mathcal{Q}$, *i.e.* $\sup_{\theta\in\Theta, q\in\mathcal{Q}} \|\widehat{D}_{\theta,q} - D_{\theta,q}\|_\mu \to_{a.s.} 0$.

Proof of Lemma 3. Adding and subtracting $\mathbb{E}(A_{\theta,q}^t)\,\mathbb{E}_n(B^t)$ to $\widehat{D}_{\theta,q}(t) - D_{\theta,q}(t)$ in

(6) and (7), one has

$$\widehat{D}_{\theta,q}(t) - D_{\theta,q}(t) = \mathbb{Q}_n(A^t_{\theta,q}B^t) - \mathbb{Q}_n(A^t_{\theta,q})\,\mathbb{E}_n(B^t) - \mathbb{E}(A^t_{\theta,q})\mathbb{Q}_n(B^t), \tag{18}$$

where $\mathbb{Q}_n \equiv \mathbb{E}_n - \mathbb{E}$. From C1 and C3, both $\sup_{\theta,q}|\mathbb{Q}_n(A^t_{\theta,q}B^t)|$ and $\sup_{\theta,q}|\mathbb{Q}_n(A^t_{\theta,q})|$ are $o_{a.s.}(1)$, and by the strong law of large numbers $|\mathbb{Q}_n(B^t)| = o_{a.s.}(1)$ as well. Applying the triangle inequality to (18), one has that $\sup_{\theta,q}|\widehat{D}_{\theta,q}(t) - D_{\theta,q}(t)| = o_{a.s.}(1)$, for every $t \in \mathcal{T}$. The continuous mapping and dominated convergence theorems then imply that

$$\sup_{\theta\in\Theta, q\in\mathcal{Q}} \|\widehat{D}_{\theta,q} - D_{\theta,q}\|_\mu \leq \Big[\int_{\mathcal{T}} \Big(\sup_{\theta\in\Theta,q\in\mathcal{Q}} |\widehat{D}_{\theta,q}(t) - D_{\theta,q}(t)|\Big)^2 d\mu(t)\Big]^{1/2} = o_{a.s.}(1) \tag{19}$$

because $\widehat{D}_{\theta,q}(t)$ and $D_{\theta,q}(t)$ are each uniformly bounded by 2. *Q.E.D.*

Proof of Theorem 1. Let $\epsilon > 0$ be arbitrary. Lemma 1 combined with C2 and (2) imply that there exists a $\delta > 0$ such that $\inf_{\theta:\|\theta-\theta_0\|>\epsilon} \|D_{\theta,q_0}\|_\mu > \delta > 0 = \|D_{\theta_0,q_0}\|_\mu$. Hence $\mathbb{P}[\|\widehat{\theta} - \theta_0\| > \epsilon] \leq \mathbb{P}[\|D_{\widehat{\theta},q_0}\|_\mu > \delta]$. By the triangle inequality,

$$\|D_{\widehat{\theta},q_0}\|_\mu \leq \|D_{\widehat{\theta},q_0} - D_{\widehat{\theta},\widehat{q}}\|_\mu + \|D_{\widehat{\theta},\widehat{q}} - \widehat{D}_{\widehat{\theta},\widehat{q}}\|_\mu + \|\widehat{D}_{\widehat{\theta},\widehat{q}}\|_\mu. \tag{20}$$

The first term in (20) is $o_{\mathbb{P}}(1)$ by Lemma 2 and C7. The second term is $o_{\mathbb{P}}(1)$ by Lemma 3. Given the definition of $\widehat{\theta}$, i.e. (4), the final term of (20) satisfies

$$\|\widehat{D}_{\widehat{\theta},\widehat{q}}\|_\mu \leq \|\widehat{D}_{\theta_0,\widehat{q}}\|_\mu + o_{\mathbb{P}}(1) \leq \|\widehat{D}_{\theta_0,\widehat{q}} - D_{\theta_0,\widehat{q}}\|_\mu + \|D_{\theta_0,\widehat{q}} - D_{\theta_0,q_0}\|_\mu + o_{\mathbb{P}}(1),$$

which is $o_{\mathbb{P}}(1)$ by Lemma 3 applied to the first term and Lemma 2 with C7 applied to the second. It follows that $\mathbb{P}[\|\widehat{\theta} - \theta_0\| > \epsilon] \leq \mathbb{P}[\|D_{\widehat{\theta},q_0}\|_\mu > \delta] = \mathbb{P}[o_{\mathbb{P}}(1) > \delta] \to 0$, which shows that $\widehat{\theta} \to_{\mathbb{P}} \theta_0$, because $\epsilon > 0$ was arbitrary. *Q.E.D.*

A.2 Asymptotic Normality

Three additional lemmas are used in the proof of Theorem 2. Lemma 4 establishes the existence and properties of the $L_2(\mu)$–Fréchet derivative of $D_{\theta,q}$. Lemma 5 shows that the centered criterion is stochastically equicontinuous, which enables an approximation of the non-smooth sample objective function by the smooth population objective function. Lemma 6 establishes weak convergence for a sample average that appears in local approximations to the criterion function at (θ_0, q_0).

Lemma 4. *Suppose that* $\|q_n - q\|_\infty = o(1)$. *Then* $\|D_{\theta,q_n} - D_{\theta,q} - \Pi_{\theta,q}^{[q_n-q]}\|_\mu = o(\|q_n - q\|_\infty)$ *uniformly in* θ *for*

$$\Pi_{\theta,q}^{[q_n-q]}(t) \equiv \int_{\mathcal{Z}} (q_n - q)(t_v|z)' \nabla_x F_{X\varepsilon_\theta|Z}(q(t_v|z), t_e \mid z) \overline{B}^t(z)\, dF_Z(z). \tag{21}$$

Under the additional smoothness assumptions in D5 b), $\|D_{\theta,q_n} - D_{\theta,q} - \Pi_{\theta,q}^{[q_n-q]}\|_\mu = o(\|q_n - q\|_\infty^2)$ *uniformly over* θ. *In either case, C4, C5 and C6 imply that* $\Pi_{\theta,q}^{[q_n-q]}$ *is* $L_2(\mu)$*–Lipschitz in* θ, *so that for any* $\theta_n \to \theta \in \Theta$, $\|\Pi_{\theta_n,q}^{[q_n-q]} - \Pi_{\theta,q}^{[q_n-q]}\|_\mu = o(\|\theta_n - \theta\|)$.

Proof of Lemma 4. Consider the first-order Taylor series expansion of $F_{X\varepsilon_\theta|Z}(\cdot, t_e|z)$ at $q_n(t_v|z)$ around $q(t_v|z)$, i.e.

$$\begin{aligned} &F_{X\varepsilon_\theta|Z}(q_n(t_v|z), t_e \mid z) - F_{X\varepsilon_\theta|Z}(q(t_v|z), t_e \mid z) \\ &\quad = (q_n - q)(t_v|z)' \nabla_x F_{X\varepsilon_\theta|Z}(q(t_v|z), t_e \mid z) + o(\|(q_n - q)(t_v|z)\|). \end{aligned} \tag{22}$$

Using this expansion with (8) and (21), one has

$$\begin{aligned} &|D_{\theta,q_n}(t) - D_{\theta,q}(t) - \Pi_{\theta,q}^{[q_n-q]}(t)| \\ &\quad \le \int_{\mathcal{Z}} \Big| F_{X\varepsilon_\theta|Z}(q_n(t_v|z), t_e \mid z) - F_{X\varepsilon_\theta|Z}(q(t_v|z), t_e \mid z) \\ &\qquad\qquad - (q_n - q)(t_v|z)' \nabla_x F_{X\varepsilon_\theta|Z}(q(t_v|z), t_e \mid z) \Big| \, dF_Z(z) \\ &\quad = \int_{\mathcal{Z}} o(\|(q_n - q)(t_v|z)\|)\, dF_Z(z) \le o(\|q_n - q\|_\infty), \end{aligned}$$

which implies the first claim, i.e. $\|D_{\theta,q_n} - D_{\theta,q} - \Pi_{\theta,q}^{[q_n-q]}\|_\mu = o(\|q_n - q\|_\infty)$, because the bound is uniform in t. Taking the Taylor series expansion out to the second order replaces the $o(\|(q_n - q)(t_v|z)\|)$ term in (22) with

$$(q_n - q)(t_v|z)'\nabla_{xx}F_{X\varepsilon_\theta|Z}(q(t_v|z), t_e \mid z)(q_n - q)(t_v|z) + o(\|(q_n - q)(t_v|z)\|^2).$$

For any x, e and z, the Hessian term can be rewritten as

$$\begin{aligned}
\nabla_{xx}F_{X\varepsilon_\theta|Z}(x, e \mid z) &= \nabla_{xx}\mathbb{E}\left(\mathbb{1}[X \le x]\mathbb{P}\left[\varepsilon_\theta \le e \mid X, Z = z\right] \mid Z = z\right) \\
&= \nabla_{xx}\mathbb{E}\left(\mathbb{1}[X \le x]\mathbb{P}\left[Y \le g_\theta(X, e) \mid X, Z = z\right] \mid Z = z\right) \\
&= \nabla_{xx}\left[\int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_{d_x}} F_{Y|XZ}(g_\theta(\overline{x}, e) \mid \overline{x}, z) f_{X|Z}(\overline{x} \mid z)\, d\overline{x}\right],
\end{aligned}$$

Using the fundamental theorem of calculus, the off-diagonal terms of $\nabla_{xx}F_{X\varepsilon_\theta|Z}(x, e\,|\,z)$ are integrals of $F_{Y|XZ}(g_\theta(\overline{x}, e) \mid \overline{x}, z) f_{X|Z}(\overline{x} \mid z)$ over $\overline{x}$ with two components of $\overline{x}$ fixed at x. By C6, these terms are bounded by $f_X^{bd}(z) \in L_2(F_Z)$. The k^{th} diagonal term is an integral over $\overline{x}$ of

$$\begin{aligned}
&\left[f_{Y|XZ}(g_\theta(\overline{x}, e) \mid \overline{x}, z)\nabla_{x_k}g_\theta(\overline{x}, e) + \nabla_{x_k}F_{Y|XZ}(g_\theta(\overline{x}, e) \mid \overline{x}, z)\right] f_{X|Z}(\overline{x} \mid z) \\
&\quad + F_{Y|XZ}(g_\theta(\overline{x}, e) \mid \overline{x}, z)\nabla_{x_k}f_{X|Z}(\overline{x} \mid z),
\end{aligned}$$

with $\overline{x}_k$ fixed at x_k. Under C4-C6 and the additional smoothness assumptions in D5 b), the preceding display is bounded by the F_Z–integrable function $[f_Y^{bd}(z)\nabla_x g^{bd}(e) + \nabla_x F_Y^{bd}(z)]f_X^{bd}(z) + \nabla_x f_X^{bd}(z)$. It follows that

$$\begin{aligned}
&|D_{\theta,q_n}(t) - D_{\theta,q}(t) - \Pi_{\theta,q}^{[q_n-q]}(t)| \\
&\quad \le \int_{\mathcal{Z}} \Big|(q_n - q)(t_v|z)'\nabla_{xx}^2 F_{X\varepsilon_\theta|Z}(q(t_v|z), t_e \mid z)(q_n - q)(t_v|z) \\
&\qquad\qquad + o(\|(q_n - q)(t_v|z)\|^2)\Big|\, dF_Z(z) \le \nabla_x g^{bd}(t_e) O(\|q_n - q\|_\infty^2),
\end{aligned}$$

which provides the second claim for $\|q_n - q\|_\infty = o(1)$ because $\nabla_x g^{bd} \in L_2(\mu)$.

For the third claim, consider the first component of $\nabla_x F_{X\varepsilon_\theta|Z}(x, e \mid z)$ for any x, e and z. With the notation $X_{-1} \equiv (X_2, \ldots, X_{d_x})$ and similarly for x_{-1}, one has

$$\begin{aligned}\nabla_{x_1} F_{X\varepsilon_\theta|Z}(x, e \mid z) &= F_{X_{-1}\varepsilon_\theta|X_1 Z}(x_{-1}, e \mid x_1, z) f_{X_1|Z}(x_1 \mid z) \\ &= \mathbb{E}\left[\mathbb{1}[X_{-1} \le x_{-1}] F_{Y|XZ}(g_\theta(x, e) \mid x_1, X_{-1}, z) \,\middle|\, x_1, z\right] f_{X_1|Z}(x_1 \mid z),\end{aligned}$$

where the first equality uses the definition of conditional probability and the second is the law of iterated expectations. Given C4-C6, it follows that for any $\theta_n, \theta \in \Theta$,

$$\left|\nabla_{x_1} F_{X\varepsilon_{\theta_n}|Z}(x, e \mid z) - \nabla_{x_1} F_{X\varepsilon_\theta|Z}(x, e \mid z)\right| \le f_X^{bd}(z) f_Y^{bd}(z) g_\Theta^{bd}(e) \|\theta_n - \theta\|.$$

The same bound holds for each of the d_x components of $\nabla_x F_{X\varepsilon_\theta|Z}(x, e \mid z)$, so

$$\left\|\nabla_x F_{X\varepsilon_{\theta_n}|Z}(x, e \mid z) - \nabla_x F_{X\varepsilon_\theta|Z}(x, e \mid z)\right\| \le d_x^{1/2} f_X^{bd}(z) f_Y^{bd}(z) g_\Theta^{bd}(e) \|\theta_n - \theta\|$$

as well. Thus from the Cauchy-Schwartz inequality one has

$$\begin{aligned}&\left|\Pi_{\theta_n, q}^{[q_n - q]}(t) - \Pi_{\theta, q}^{[q_n - q]}(t)\right| \\ &\quad = \Big| \int_{\mathcal{Z}} (q_n - q)(t_v|z)' [\nabla_x F_{X\varepsilon_{\theta_n}|Z}(q(t_v|z), t_e \mid z) \\ &\qquad\qquad - \nabla_x F_{X\varepsilon_\theta|Z}(q(t_v|z), t_e \mid z)] \overline{B}^t(z) \, dF_Z(z) \Big| \\ &\quad \le 2 \|q_n - q\|_\infty \int_{\mathcal{Z}} d_x^{1/2} f_X^{bd}(z) f_Y^{bd}(z) g_\Theta^{bd}(e) \|\theta_n - \theta\| \, dF_Z(z) \\ &\quad = g_\Theta^{bd}(t_e) O(\|q_n - q\|_\infty) O(\|\theta_n - \theta\|),\end{aligned}$$

where the last equality uses $f_Y^{bd}, f_X^{bd} \in L_2(F_Z)$ from C5 and C6. This bounds implies the third claim when $\|\theta_n - \theta\| = o(1)$ and $\|q_n - q\|_\infty = O(1)$, because $g_\Theta^{bd} \in L_2(\mu)$. *Q.E.D.*

Lemma 5. *If C1, C4 and D3 hold then* $\{\widehat{D}_{\theta,q} - D_{\theta,q} : \theta \in \Theta, q \in \mathcal{Q}\}$ *is* $\sqrt{n}$*–stochastically equicontinuous in* $L_2(\mu)$ *at* (θ_0, q_0)*. That is, if* $\|(\theta_n, q_n) - (\theta_0, q_0)\| \to_{\mathbb{P}} 0$

then $\sqrt{n}\|\widehat{D}_{\theta_n,q_n} - D_{\theta_n,q_n} - (\widehat{D}_{\theta_0,q_0} - D_{\theta_0,q_0})\|_\mu \to_{\mathbb{P}} 0$.

Proof of Lemma 5. By C1 and D3, the sequence of stochastic processes $\mathbb{A}_n^t \equiv \{\mathbb{G}_n A_{\theta,q}^t : (\theta,q) \in \Theta \times \mathcal{Q}\}$ converges weakly to a mean-zero Gaussian process in $l^\infty(\Theta \times \mathcal{Q})$ for any fixed t, where $\mathbb{G}_n \equiv \sqrt{n}\mathbb{Q}_n \equiv \sqrt{n}(\mathbb{E}_n - \mathbb{E})$ and $l^\infty(\mathcal{I})$ denotes the space of bounded real-valued functions with domain $\mathcal{I}$. This implies that $\mathbb{A}_n^t$ is stochastically equicontinuous with respect to the $L_2(\mathbb{P})$ metric—see Example 1.5.10 of van der Vaart and Wellner (1996). By definition, this means that for any $\epsilon_1, \epsilon_2 > 0$ there exists a $\delta_1 > 0$ such that

$$\limsup_n \mathbb{P}\left[\sup_{(\theta_1,q_1),(\theta_2,q_2):\mathbb{P}(A_{\theta_1,q_1}^t - A_{\theta_2,q_2}^t)^2 < \delta_1} \left|\mathbb{G}_n(A_{\theta_1,q_1}^t) - \mathbb{G}_n(A_{\theta_2,q_2}^t)\right| > \epsilon_1\right] < \epsilon_2.$$

Below, I will establish that $\|(\theta_n,q_n) - (\theta_0,q_0)\| \to_{\mathbb{P}} 0$ implies $\mathbb{P}(A_{\theta_n,q_n}^t - A_{\theta_0,q_0}^t)^2 \to 0$. This in turn implies that there also exists a $\delta_2 > 0$ such that

$$\limsup_n \mathbb{P}\left[\sup_{(\theta,q):\|(\theta,q)-(\theta_0,q_0)\| < \delta_2} \left|\mathbb{G}_n(A_{\theta,q}^t) - \mathbb{G}_n(A_{\theta_0,q_0}^t)\right| > \epsilon_1\right] < \epsilon_2.$$

This is equivalent to the statement that $\|(\theta_n,q_n) - (\theta_0,q_0)\| \to_{\mathbb{P}} 0$ implies

$$\mathbb{G}_n(A_{\theta_n,q_n}^t) - \mathbb{G}_n(A_{\theta_0,q_0}^t) \xrightarrow{\mathbb{P}} 0 \tag{23}$$

for any fixed t.[20] The claimed $\sqrt{n}$–stochastic equicontinuity of $\{\widehat{D}_{\theta,q} - D_{\theta,q} : (\theta,q) \in \Theta \times \mathcal{Q}\}$ in $L_2(\mu)$ will then follow after some algebraic manipulations and an appeal to the continuous mapping and dominated convergence theorems.

By C4, $g_{\theta_n}(x,t_e) \leq g_{\theta_0}(x,t_e) + g_\Theta^{bd}(t_e)\,\|\theta_n - \theta_0\|$ and by the definition of the sup-

[20] See e.g. pp. 139-140 of Pollard (1984).

norm, $q_n(t_v \mid z) \leq q_{0,k}(t_v \mid z) + \|q_n - q_0\|_\infty$. Hence for $\|(\theta_n, q_n) - (\theta_0, q_0)\| \to_{\mathbb{P}} 0$,

$$\begin{aligned}
\mathbb{E}(A^t_{\theta_n,q_n}) &= \mathbb{P}\left[X \leq q_n(t_z \mid Z), Y \leq g_{\theta_n}(X, t_e)\right] \\
&\leq \mathbb{P}\left[X \leq q_0(t_z \mid Z) + \|q_n - q_0\|_\infty, Y \leq g_{\theta_0}(X, t_e) + g^{bd}_\Theta(t_e)\,\|\theta_n - \theta_0\|\right] \\
&\to \mathbb{P}\left[X \leq q_0(t_z \mid Z), Y \leq g_{\theta_0}(X, t_e)\right] = \mathbb{E}(A^t_{\theta_0,q_0}),
\end{aligned}$$

where the convergence follows because (X, Y) is continuously distributed conditional on Z, given I1 and I2. By similar reasoning,

$$\begin{aligned}
\mathbb{E}(A^t_{\theta_n,q_n} A^t_{\theta_0,q_0}) &= \mathbb{P}\left[X \leq q_n(t_z \mid Z) \wedge q_0(t_z \mid Z), Y \leq g_{\theta_n}(X, t_e) \wedge g_{\theta_0}(X, t_e)\right] \\
&\geq \mathbb{P}\left[X \leq q_0(t_z \mid Z) - \|q_n - q_0\|_\infty, \right. \\
&\qquad \left. Y \leq g_{\theta_0}(X, t_e) - g^{bd}_\Theta(t_e)\,\|\theta_n - \theta_0\|\right] \to \mathbb{E}(A^t_{\theta_0,q_0}). \qquad (24)
\end{aligned}$$

Since $\mathbb{E}(A^t_{\theta_n,q_n}) \geq \mathbb{E}(A^t_{\theta_n,q_n} A^t_{\theta_0,q_0})$ and $\mathbb{E}(A^t_{\theta_n,q_n} A^t_{\theta_0,q_0}) \leq \mathbb{E}(A^t_{\theta_0,q_0})$, it follows that

$$\mathbb{E}(A^t_{\theta_n,q_n} - A^t_{\theta_0,q_0})^2 = \mathbb{E}\, A^t_{\theta_n,q_n} - 2\,\mathbb{E}(A^t_{\theta_n,q_n} A^t_{\theta_0,q_0}) + \mathbb{E}\, A^t_{\theta_0,q_0} \to 0.$$

As argued above, this implies (23).

The stochastic equicontinuity of $\{\sqrt{n}(\widehat{D}_{\theta,q}(t) - D_{\theta,q}(t)) : \theta \in \Theta, q \in \mathcal{Q}\}$ at (θ_0, q_0) in $\|\cdot\|$ for fixed t now follows from decomposition (18) in Lemma 3:

$$\begin{aligned}
\sqrt{n}(\widehat{D}_{\theta_n,q_n}(t) - D_{\theta_n,q_n}(t)) &= \mathbb{G}_n(A^t_{\theta_n,q_n} B^t) + \mathbb{E}_n(B^t)\mathbb{G}_n(A^t_{\theta_n,q_n}) + \mathbb{E}(A^t_{\theta_n,q_n})\mathbb{G}_n(B^t) \\
&= \mathbb{G}_n(A^t_{\theta_0,q_0} B^t) + o_{\mathbb{P}}(1) + \mathbb{E}_n(B^t)\mathbb{G}_n(A^t_{\theta_0,q_0}) + O_{\mathbb{P}}(1)o_{\mathbb{P}}(1) \\
&\quad + \mathbb{E}(A^t_{\theta_0,q_0})\mathbb{G}_n(B^t) + o(1)O_{\mathbb{P}}(1) \\
&= \sqrt{n}(\widehat{D}_{\theta_0,q_0}(t) - D_{\theta_0,q_0}(t)) + o_{\mathbb{P}}(1), \qquad (25)
\end{aligned}$$

where the second equality uses (23), $\mathbb{E}(A^t_{\theta_n,q_n}) = \mathbb{E}(A^t_{\theta_0,q_0}) + o_{\mathbb{P}}(1)$ and $\mathbb{G}_n(B^t) = O_{\mathbb{P}}(1)$. The desired $\sqrt{n}$–stochastic equicontinuity in $L_2(\mu)$, i.e. $\sqrt{n}\|\widehat{D}_{\theta_n,q_n} - D_{\theta_n,q_n} -$

$(\widehat{D}_{\theta_0,q_0} - D_{\theta_0,q_0})\|_\mu = o_{\mathbb{P}}(1)$, then follows after applications of the continuous mapping and dominated convergence theorems similar to (19).[21] *Q.E.D.*

Lemma 6. *Given Assumptions I, C1 and D3,* $\{\sqrt{n}\widehat{D}_{\theta_0,q_0}(t) : t \in \mathcal{T}\}$ *converges weakly in* $l^\infty(\mathcal{T})$ *to a mean-zero Gaussian process with covariance function given by* σ *in* (12). *Moreover, for every* t, $\sqrt{n}\widehat{D}_{\theta_0,q_0}(t)$ *has the Bahadur representation* $\sqrt{n}\widehat{D}_{\theta_0,q_0}(t) = \sqrt{n}\,\mathbb{E}_n\,\chi(t) + o_{\mathbb{P}}(1)$ *where* $\chi(t) \equiv [A^t_{\theta_0,q_0} - \mathbb{E}(A^t_{\theta_0,q_0})][B^t - \mathbb{E}(B^t)]$ *has population mean zero and the remainder term is uniform over* $t \in \mathcal{T}$.

Proof of Lemma 6. Some algebra together with $\mathbb{E}\,\chi(t) = D_{\theta_0,q_0}(t) = 0$ shows that

$$\sqrt{n}\widehat{D}_{\theta_0,q_0}(t) = \mathbb{G}_n(\chi(t)) - \mathbb{Q}_n(A^t_{\theta_0,q_0})\mathbb{G}_n(B^t), \tag{26}$$

for all t. I claim that $\{\chi(t) : t \in \mathcal{T}\}$ is a Donsker class. By Example 2.10.8 of van der Vaart and Wellner (1996), this follows if both $\{A^t_{\theta_0,q_0} : t \in \mathcal{T}\}$ and $\{B^t : t \in \mathcal{T}\}$ are Donsker classes, because if so then $\{A^t_{\theta_0,q_0} - \mathbb{E}(A^t_{\theta_0,q_0}) : t \in \mathcal{T}\}$ and $\{B^t - \mathbb{E}(B^t) : t \in \mathcal{T}\}$ are uniformly bounded Donsker classes and hence their pairwise product, which contains $\{\chi(t) : t \in \mathcal{T}\}$, is also Donsker. The collection $\{B^t : t \in \mathcal{T}\}$ is the canonical example of a Donsker class. For $\{A^t_{\theta_0,q_0} : t \in \mathcal{T}\}$, note that for any t, $A^t_{\theta_0,q_0} = A^{t_e}_{\theta_0} \prod_{k=1}^{d_x} A^{t_{v_k}}_{q_{0,k}}$, where $A^{t_e}_{\theta_0}(w) \equiv \mathbb{1}[w_y \leq g_{\theta_0}(w_x, t_e)]$ and $A^{t_{v_k}}_{q_{0,k}}(w) \equiv \mathbb{1}[w_{x_k} \leq q_{0,k}(t_{v_k}|w_z)]$. The collection $\{A^{t_e}_{\theta_0} : t_e \in \mathcal{E} \subseteq \mathbb{R}\}$ is increasing in the index, t_e, because if $t_e \leq t'_e$ then $g_{\theta_0}(x, t_e) \leq g_{\theta_0}(x, t'_e)$ for all $x \in \mathcal{X}$ by I2 and hence

$$A^{t_e}_{\theta_0}(w) \equiv \mathbb{1}[w_y \leq g_{\theta_0}(w_x, t_e)] \leq \mathbb{1}[w_y \leq g_{\theta_0}(w_x, t'_e)] \equiv A^{t'_e}_{\theta_0}(w)$$

for every w. Lemma 9.10 of Kosorok (2008) establishes that collections with this property are VC-subgraph and hence Donsker. The same is true of the collection $\{A^{t_{v_k}}_{q_{0,k}} : t_{v_k} \in (0,1)\}$ because $q_{0,k} \in \mathcal{Q}_k$ is increasing in t_{v_k}. Appealing again to Ex-

[21] Note that this is the dominated convergence theorem for convergence in probability, which is an extension of the standard dominated convergence theorem, see e.g. Corollary 6.3.2 of Resnick (1999).

ample 2.10.8 of van der Vaart and Wellner (1996), it follows that $\{A^t_{\theta_0,q_0} : t \in \mathcal{T}\}$ is Donsker and thus that $\{\chi(t) : t \in \mathcal{T}\}$ is also Donsker.

Returning to (26), the Donsker property of $\{\chi(t) : t \in \mathcal{T}\}$ together with C1 imply that the first term converges weakly in $l^\infty(\mathcal{T})$ to a mean-zero Gaussian process. The Donsker properties of $\{A^t_{\theta_0,q_0} : t \in \mathcal{T}\}$ and $\{B^t : t \in \mathcal{T}\}$ with C1 imply that the second term is $o_{\mathbb{P}}(1)O_{\mathbb{P}}(1) = o_{\mathbb{P}}(1)$ uniformly in t. An application of Slutsky's Theorem establishes the weak convergence of $\{\sqrt{n}\widehat{D}_{\theta_0,q_0}(t) : t \in \mathcal{T}\}$. This also shows that $\sqrt{n}\widehat{D}_{\theta_0,q_0}(t) = \mathbb{G}_n\chi(t) + o_{\mathbb{P}}(1) = \sqrt{n}\,\mathbb{E}_n\,\chi(t) + o_{\mathbb{P}}(1)$ uniformly in t.

To compute the covariance function of this limiting process, first observe that

$$\begin{aligned}\mathbb{E}[A^t_{\theta_0,q_0}A^{\widetilde{t}}_{\theta_0,q_0}] &= \mathbb{E}\left[\mathbb{P}\left[X \le q_0(t_v \wedge \widetilde{t}_v|Z), Y \le g_{\theta_0}(X, t_e \wedge \widetilde{t}_e) \,\middle|\, Z\right]\right] \\ &= \mathbb{E}\left[\mathbb{P}\left[X \le q_0(t_v \wedge \widetilde{t}_v|Z), \varepsilon_{\theta_0} \le t_e \wedge \widetilde{t}_e \,\middle|\, Z\right]\right] = C_{\theta_0}(t_v \wedge \widetilde{t}_v, F_{\varepsilon_{\theta_0}}(t_e \wedge \widetilde{t}_e)),\end{aligned}$$

where the second equality uses $Y = g_{\theta_0}(X, \varepsilon_{\theta_0})$ and the last uses Sklar's Theorem and Proposition 1. Similarly, $\mathbb{E}(A^t_{\theta_0,q_0}) = C_{\theta_0}(t_v, F_{\varepsilon_{\theta_0}}(t_e))$ and hence

$$\begin{aligned}\mathbb{E}(\chi(t)\chi(\widetilde{t})) = \mathbb{E}\Big(&\mathbb{E}\left[(A^t_{\theta_0,q_0} - \mathbb{E}(A^t_{\theta_0,q_0}))(A^{\widetilde{t}}_{\theta_0,q_0} - \mathbb{E}(A^{\widetilde{t}}_{\theta_0,q_0})) \,\middle|\, Z\right] \\ &\times (B^t - \mathbb{E}(B^t))(B^{\widetilde{t}} - \mathbb{E}(B^{\widetilde{t}}))\Big) \\ = &\left[C_{\theta_0}(t_v \wedge \widetilde{t}_v, F_{\varepsilon_{\theta_0}}(t_e \wedge \widetilde{t}_e)) - C_{\theta_0}(t_v, F_{\varepsilon_{\theta_0}}(t_e))C_{\theta_0}(\widetilde{t}_v, F_{\varepsilon_{\theta_0}}(\widetilde{t}_e))\right] \\ &\times \left[F_Z(t_z \wedge \widetilde{t}_z) - F_Z(t_z)F_Z(\widetilde{t}_z)\right],\end{aligned}$$

which is equal to $\sigma(t, \widetilde{t})$ as given in (12). *Q.E.D.*

Proof of Theorem 2. The argument follows the same strategy as Theorem 3.3 of Pakes and Pollard (1989) and Theorem 2 of Chen et al. (2003). First, I establish that

$\widehat{\theta} = \theta_0 + O_{\mathbb{P}}(n^{-1/2})$. By the triangle inequality and Lemma 4 with D5 and D6,[22]

$$\begin{aligned}\|D_{\widehat{\theta},q_0}\|_\mu &\leq \|D_{\widehat{\theta},\widehat{q}}\|_\mu + \|D_{\widehat{\theta},\widehat{q}} - D_{\widehat{\theta},q_0} - \Pi^{[\widehat{q}-q_0]}_{\widehat{\theta},q_0}\|_\mu + \|\Pi^{[\widehat{q}-q_0]}_{\widehat{\theta},q_0} - \Pi^{[\widehat{q}-q_0]}_{\theta_0,q_0}\|_\mu + \|\Pi^{[\widehat{q}-q_0]}_{\theta_0,q_0}\|_\mu \\ &= \|D_{\widehat{\theta},\widehat{q}}\|_\mu + o(\|\widehat{\theta} - \theta_0\|) + O_{\mathbb{P}}(n^{-1/2}). \end{aligned} \tag{27}$$

From Lemma 5 followed by the triangle inequality and the implication of Lemma 6 that $\|\widehat{D}_{\theta_0,q_0}\|_\mu = O_{\mathbb{P}}(n^{-1/2})$, one has

$$\|D_{\widehat{\theta},\widehat{q}}\|_\mu = \|\widehat{D}_{\widehat{\theta},\widehat{q}} - \widehat{D}_{\theta_0,q_0} + D_{\theta_0,q_0}\|_\mu + o_{\mathbb{P}}(n^{-1/2}) \leq \|\widehat{D}_{\widehat{\theta},\widehat{q}}\|_\mu + O_{\mathbb{P}}(n^{-1/2}), \tag{28}$$

where $D_{\theta_0,q_0} = 0$ by (2). Using the definition of $\widehat{\theta}$ and Lemma 5,

$$\begin{aligned}\|\widehat{D}_{\widehat{\theta},\widehat{q}}\|_\mu &\leq \|\widehat{D}_{\theta_0,\widehat{q}}\|_\mu + o_{\mathbb{P}}(n^{-1/2}) \\ &= \|\widehat{D}_{\theta_0,q_0} - D_{\theta_0,\widehat{q}} + D_{\theta_0,q_0}\|_\mu + o_{\mathbb{P}}(n^{-1/2}) \\ &\leq \|D_{\theta_0,\widehat{q}} - D_{\theta_0,q_0} - \Pi^{[\widehat{q}-q_0]}_{\theta_0,q_0}\|_\mu + \|\Pi^{[\widehat{q}-q_0]}_{\theta_0,q_0}\|_\mu + O_{\mathbb{P}}(n^{-1/2}), \end{aligned} \tag{29}$$

where the second inequality uses the triangle inequality with $\|\widehat{D}_{\theta_0,q_0}\|_\mu = O_{\mathbb{P}}(n^{-1/2})$. Applying Lemma 4 with D5 and D6 in (29), one has $\|\widehat{D}_{\widehat{\theta},\widehat{q}}\|_\mu \leq O_{\mathbb{P}}(n^{-1/2})$. Combined with (27) and (28), this yields

$$\|D_{\widehat{\theta},q_0}\|_\mu \leq o(\|\widehat{\theta} - \theta_0\|) + O_{\mathbb{P}}(n^{-1/2}). \tag{30}$$

For each t, a Taylor expansion of $D_{\widehat{\theta},q_0}(t)$ around $D_{\theta_0,q_0}(t) = 0$ gives $D_{\widehat{\theta},q_0}(t) = (\widehat{\theta} - \theta_0)'\Delta_0(t) + o(\|\widehat{\theta} - \theta_0\|)$, where the differentiability of $D_{\theta,q_0}(t)$ at θ_0 follows from D1 and D2, given the bounds on $\nabla_\theta g_{\theta_0}$ and $f_{Y|XZ}$ provided by C4 and C5—see (10). Assumption D4 implies that $(\widehat{\theta} - \theta_0)'\Delta_0(t) \neq 0$ on a non-negligible subset of $\mathcal{T}$ which

[22] A more precise proof would multiply all quantities by indicators for the event that $\widehat{q} \in \mathcal{Q}$ and then appeal to the assumption in D5 that this event happens with probability approaching one. It is common in the literature to ignore this distinction and I will do so as well.

means there is a constant $\kappa_\Delta > 0$ such that

$$\|D_{\widehat{\theta},q_0}\|_\mu = \|(\widehat{\theta}-\theta_0)'\Delta_0 + o(\|\widehat{\theta}-\theta_0\|)\|_\mu \geq \|\widehat{\theta}-\theta_0\|\kappa_\Delta + o(\|\widehat{\theta}-\theta_0\|). \tag{31}$$

Together with (30), one has $O_{\mathbb{P}}(1) \geq \sqrt{n}\|\widehat{\theta}-\theta_0\|\kappa_\Delta + o(\sqrt{n}\|\widehat{\theta}-\theta_0\|)$, which shows that $\widehat{\theta} = \theta_0 + O_{\mathbb{P}}(n^{-1/2})$, as claimed.

Next, I show that $\sqrt{n}(\widehat{\theta}-\theta_0)$ is asymptotically normal. Define

$$\widehat{L}_\theta(t) \equiv \widehat{D}_{\theta_0,q_0}(t) + (\theta-\theta_0)'\Delta_0(t) + \Pi_{\theta_0,q_0}^{[\widehat{q}-q_0]}(t)$$

as a linear approximation of $\widehat{D}_{\theta,\widehat{q}}(t)$ for θ near θ_0. For any sequence $\theta_n \to_{\mathbb{P}} \theta_0$, one has

$$\|\widehat{D}_{\theta_n,\widehat{q}} - \widehat{L}_{\theta_n}\|_\mu \leq \|\widehat{D}_{\theta_n,\widehat{q}} - \widehat{D}_{\theta_0,q_0} - D_{\theta_n,\widehat{q}}\|_\mu + \|D_{\theta_n,\widehat{q}} - (\theta_n-\theta_0)'\Delta_0 - \Pi_{\theta_0,q_0}^{[\widehat{q}-q_0]}\|_\mu.$$

The first term is $o_{\mathbb{P}}(n^{-1/2})$ by Lemma 5 and the second term is bounded above by

$$\begin{aligned}\|D_{\theta_n,\widehat{q}} - D_{\theta_n,q_0} - \Pi_{\theta_n,q_0}^{[\widehat{q}-q_0]}\|_\mu + \|D_{\theta_n,q_0} - (\theta_n-\theta_0)'\Delta_0\|_\mu + \|\Pi_{\theta_n,q_0}^{[\widehat{q}-q_0]} - \Pi_{\theta_0,q_0}^{[\widehat{q}-q_0]}\|_\mu \\ = o_{\mathbb{P}}(n^{-1/2}) + o(\|\theta_n-\theta_0\|),\end{aligned}$$

where the rates for the first and third terms are due to Lemma 4 with D5, and that for the second term is from the definition of a derivative. The previous display is $o_{\mathbb{P}}(n^{-1/2})$ if $\theta_n = \theta_0 + O_{\mathbb{P}}(n^{-1/2})$, which implies that $\widehat{L}_{\theta_n}$ sufficiently well approximates $\widehat{D}_{\theta_n,\widehat{q}}$ for such sequences, i.e. $\|\widehat{D}_{\theta_n,\widehat{q}} - \widehat{L}_{\theta_n}\|_\mu = o_{\mathbb{P}}(n^{-1/2})$.

The vector that minimizes $\|\widehat{L}_\theta\|_\mu$ is the $\widetilde{\theta}$ such that $(\widetilde{\theta}-\theta_0)'\Delta_0$ is the $L_2(\mu)$-projection of $-\widehat{D}_{\theta_0,q_0} - \Pi_{\theta_0,q_0}^{[\widehat{q}-q_0]}$ onto the subspace of $L_2(\mu)$ spanned by the components of Δ_0—see e.g. pg. 51 of Luenberger (1968). Solving the normal equations for this projection and scaling by $\sqrt{n}$ provides

$$\sqrt{n}(\widetilde{\theta}-\theta_0) = -\overline{\Delta}_0^{-1}\int_{\mathcal{T}} \Delta_0(t)\sqrt{n}\left[\widehat{D}_{\theta_0,q_0}(t) + \Pi_{\theta_0,q_0}^{[\widehat{q}-q_0]}(t)\right] d\mu(t), \tag{32}$$

where $\overline{\Delta}_0 \equiv \int_{\mathcal{T}} \Delta_0(t)\Delta_0(t)'\, d\mu(t)$ is invertible by D4.[23] From Lemma 6 and D6,

$$\begin{aligned}
&\int_{\mathcal{T}} \Delta_0(t)\sqrt{n}\left[\widehat{D}_{\theta_0,q_0}(t) + \Pi^{[\widehat{q}-q_0]}_{\theta_0,q_0}(t)\right] d\mu(t) \\
&\quad = \int_{\mathcal{T}} \Delta_0(t)\left[\sqrt{n}\,\mathbb{E}_n\left(\chi(t)+\psi(t)\right) + o_{\mathbb{P}}(1)\right] d\mu(t) = \sqrt{n}\,\mathbb{E}_n\,\xi + o_{\mathbb{P}}(1),
\end{aligned}$$

for $\xi \equiv \int_{\mathcal{T}} \Delta_0(t)[\chi(t)+\psi(t)]\, d\mu(t)$. The remainder term is $o_{\mathbb{P}}(1)$ because the Bahadur representations are uniform over $t \in \mathcal{T}$ and $\Delta_0 \in L_2(\mu)$.[24] By Fubini's Theorem, Lemma 6 and D6, $\mathbb{E}\,\xi = \int_{\mathcal{T}} \Delta_0(t)\,\mathbb{E}\left(\chi(t)+\psi(t)\right)\, d\mu(t) = 0$ and

$$\begin{aligned}
\mathbb{E}\,\xi\xi' &= \int_{\mathcal{T}\times\mathcal{T}} \Delta_0(t)\Delta_0(\widetilde{t})'\,\mathbb{E}\left[(\chi(t)+\psi(t))(\chi(\widetilde{t})+\psi(\widetilde{t}))\right] d\mu(t)\, d\mu(\widetilde{t}) \\
&= \int_{\mathcal{T}\times\mathcal{T}} \Delta_0(t)\Delta_0(\widetilde{t})'\left[\sigma(t,\widetilde{t}) + \nu(t,\widetilde{t})\right]\, d\mu(t)\, d\mu(\widetilde{t}) \equiv \overline{\Sigma}_0,
\end{aligned}$$

which is finite because both $\sigma(t,\widetilde{t})$ and $\nu(t,\widetilde{t})$ are uniformly bounded, and $\Delta_0 \in L_2(\mu)$. Hence by the central limit theorem, $\sqrt{n}\,\mathbb{E}_n\,\xi \rightsquigarrow N(0,\overline{\Sigma}_0)$. It follows from (32) that $\sqrt{n}(\widetilde{\theta}-\theta_0) \rightsquigarrow N(0,\overline{\Delta}_0^{-1}\overline{\Sigma}_0\overline{\Delta}_0^{-1})$, where I note that $\overline{\Delta}_0$ is symmetric.

The remainder of the proof shows that $\sqrt{n}(\widehat{\theta}-\theta_0) = \sqrt{n}(\widetilde{\theta}-\theta_0) + o_{\mathbb{P}}(1)$. I continue to follow the approach of Pakes and Pollard (1989). Like those authors, I assume for simplicity that $\widetilde{\theta}$ is always in a small neighborhood of θ_0 that is strictly contained in Θ. Due to D1 and the already established result that $\widetilde{\theta} \to_{\mathbb{P}} \theta_0$, this will be true with probability approaching one, i.e. the event that $\widetilde{\theta}$ is not in such a neighborhood is asymptotically negligible.

As previously shown, $\widehat{L}_{\widehat{\theta}}$ and $\widehat{L}_{\widetilde{\theta}}$ are $o_{\mathbb{P}}(n^{-1/2})$ approximations in $L_2(\mu)$ of $\widehat{D}_{\widehat{\theta},\widehat{q}}$ and $\widehat{D}_{\widetilde{\theta},\widehat{q}}$, because both $\widehat{\theta} = \theta_0 + O_{\mathbb{P}}(n^{-1/2})$ and $\widetilde{\theta} = \theta_0 + O_{\mathbb{P}}(n^{-1/2})$. Hence, by the triangle inequality, $\|\widehat{L}_{\widehat{\theta}}\|_\mu - \|\widehat{D}_{\widehat{\theta},\widehat{q}}\|_\mu \leq \|\widehat{L}_{\widehat{\theta}} - \widehat{D}_{\widehat{\theta},\widehat{q}}\|_\mu = o_{\mathbb{P}}(n^{-1/2})$ and $\|\widehat{D}_{\widetilde{\theta},\widehat{q}}\|_\mu - \|\widehat{L}_{\widetilde{\theta}}\|_\mu \leq$

[23] For otherwise, there would exist a non-zero vector $a \in \mathbb{R}^{d_\theta}$ such that $a'\overline{\Delta}_0 a = \int_{\mathcal{T}} (a'\Delta_0(t))(a'\Delta_0(t))'\, d\mu(t) = 0$. This implies that $a'\Delta_0(t) = 0$ for μ–a.e. $t \in \mathcal{T}$ and a non-zero a, in contradiction with D4.

[24] The latter follows from C4 and C5, as can be seen from (10).

$\|\widehat{D}_{\widetilde{\theta},\widehat{q}} - \widehat{L}_{\widetilde{\theta}}\|_\mu = o_{\mathbb{P}}(n^{-1/2})$. Using the definition of $\widehat{\theta}$, it follows that

$$0 \leq \|\widehat{L}_{\widehat{\theta}}\|_\mu \leq \|\widehat{D}_{\widehat{\theta},\widehat{q}}\|_\mu + o_{\mathbb{P}}(n^{-1/2}) \leq \|\widehat{D}_{\widetilde{\theta},\widehat{q}}\|_\mu + o_{\mathbb{P}}(n^{-1/2}) \leq \|\widehat{L}_{\widetilde{\theta}}\|_\mu + o_{\mathbb{P}}(n^{-1/2}),$$

i.e. that $\|\widehat{L}_{\widehat{\theta}}\|_\mu = \|\widehat{L}_{\widetilde{\theta}}\|_\mu + o_{\mathbb{P}}(n^{-1/2})$. By the triangle inequality with Lemma 6, D6 and $\widetilde{\theta} = \theta_0 + O_{\mathbb{P}}(n^{-1/2})$, one has $\|\widehat{L}_{\widetilde{\theta}}\|_\mu = O_{\mathbb{P}}(n^{-1/2})$ and so $\|\widehat{L}_{\widehat{\theta}}\|_\mu^2 = (\|\widehat{L}_{\widetilde{\theta}}\|_\mu + o_{\mathbb{P}}(n^{-1/2}))^2 = \|\widehat{L}_{\widetilde{\theta}}\|_\mu^2 + o_{\mathbb{P}}(n^{-1})$.

Next, add and subtract $(\widetilde{\theta} - \theta_0)'\Delta_0$ to $\widehat{L}_{\widehat{\theta}}$, obtaining

$$\begin{aligned}\|\widehat{L}_{\widehat{\theta}}\|_\mu^2 &= \|\widehat{D}_{\theta_0,q_0} + (\widetilde{\theta} - \theta_0)'\Delta_0 + \Pi_{\theta_0,q_0}^{[\widehat{q}-q_0]} + (\widehat{\theta} - \widetilde{\theta})'\Delta_0\|_\mu^2 \\ &= \|\widehat{L}_{\widetilde{\theta}}\|_\mu^2 + 2(\widehat{\theta} - \widetilde{\theta})' \int_{\mathcal{T}} \widehat{L}_{\widetilde{\theta}}(t)\Delta_0(t)\, d\mu(t) + \|(\widehat{\theta} - \widetilde{\theta})'\Delta_0\|_\mu^2.\end{aligned}$$

The inner product term in this expression is equal to 0 because $\widehat{L}_{\widetilde{\theta}}$, i.e. the residual from projecting $-\widehat{D}_{\theta_0,q_0} - \Pi_{\theta_0,q_0}^{[\widehat{q}-q_0]}$ onto the subspace spanned by Δ_0, must be orthogonal to Δ_0 in $L_2(\mu)$ (Luenberger (1968), pg. 51). Hence, $\|\widehat{L}_{\widetilde{\theta}}\|_\mu^2 + \|(\widehat{\theta} - \widetilde{\theta})'\Delta_0\|_\mu^2 = \|\widehat{L}_{\widehat{\theta}}\|_\mu^2 = \|\widehat{L}_{\widetilde{\theta}}\|_\mu^2 + o_{\mathbb{P}}(n^{-1})$, i.e. $\|(\widehat{\theta} - \widetilde{\theta})'\Delta_0\|_\mu^2 = o_{\mathbb{P}}(n^{-1})$. By the same argument as in (31), $o_{\mathbb{P}}(n^{-1/2}) = \|(\widehat{\theta} - \widetilde{\theta})'\Delta_0\|_\mu \geq \|\widehat{\theta} - \widetilde{\theta}\|\kappa_\Delta$ for $\kappa_\Delta > 0$, and so $\|\widehat{\theta} - \widetilde{\theta}\| = o_{\mathbb{P}}(n^{-1/2})$. Given the previously derived limiting distribution of $\sqrt{n}(\widetilde{\theta} - \theta_0)$, one has by Slutsky's Theorem that $\sqrt{n}(\widehat{\theta} - \theta_0) = \sqrt{n}(\widetilde{\theta} - \theta_0) + o_{\mathbb{P}}(1) \rightsquigarrow N(0, \overline{\Delta}_0^{-1}\overline{\Sigma}_0\overline{\Delta}_0^{-1})$. *Q.E.D.*

Proof of Proposition 2. For any θ, q and t, write $A_{\theta,q}^t = A_\theta^{t_e} \prod_{k=1}^{d_x} A_{q_k}^{t_{v_k}}$, where $A_\theta^{t_e}(w) \equiv \mathbb{1}[w_y \leq g_\theta(w_x, t_e)]$ and $A_{q_k}^{t_{v_k}}(w) \equiv \mathbb{1}[w_{x_k} \leq q_k(t_{v_k}|w_z)]$ are indicator functions defined on $\mathcal{W}$. Note that $\{w \in \mathcal{W} : w_y \leq g_\theta(w_x, t_e)\} = \{(y, x, z) \in \mathbb{R} \times \mathcal{X} \times \mathcal{Z} : y \leq g_\theta(x, t_e)\} \cap \mathcal{W}$ is the subgraph of $g_\theta(\cdot, t_e)$ intersected with $\mathcal{W}$.[25] Under 1a or 1b, the collection of indicator functions for subgraphs of $\{g_\theta(\cdot, t_e) : \theta \in \Theta\}$ is Donsker by, respectively, Lemma 2.6.15 or Corollary 2.7.3 of van der Vaart and Wellner (1996), together with the uniform central limit theorem results in their Section 2.5. A similar comment applies to

[25] Subgraphs are often defined with a strict inequality. That the following analysis also applies to subgraphs defined with a weak inequality can be seen by Theorem 9.30 of Kosorok (2008), which shows that the pointwise closure of a Donsker class is also Donsker.

the collection $\{A_{q_k}^{t_{v_k}} : q_k \in \mathcal{Q}_k\}$ for every k, given 2a or 2b. It follows from Example 2.10.8 of van der Vaart and Wellner (1996) that $\{A_{\theta,q}^t : \theta \in \Theta, q \in \mathcal{Q}\}$, as the product of these uniformly bounded Donsker classes, is also Donsker, which is Assumption D3. Donsker classes are Glivenko-Cantelli, so C3 is also satisfied. *Q.E.D.*

A.3 Bootstrap

The proof of Theorem 3 requires Lemmas $5^\star$ and $6^\star$, which are bootstrap counterparts to Lemmas 5 and 6. Throughout this appendix statements of $\mathbb{P}^\star$–convergence can be understood to hold a.s.–$\mathbb{P}$ except where otherwise noted.

Lemma $5^\star$. *Assumptions C1, C4 and D3 imply that* $\sqrt{n}\|D^\star_{\theta_n,q_n} - \widehat{D}_{\theta_n,q_n} - (D^\star_{\theta_0,q_0} - \widehat{D}_{\theta_0,q_0})\|_\mu \to_{\mathbb{P}^\star} 0$, *for any sequence* $(\theta_n, q_n) \to_{\mathbb{P}^\star} (\theta_0, q_0)$.

Proof of Lemma $5^\star$. Let $\mathbb{G}^\star_n \equiv \sqrt{n}(\mathbb{E}^\star_n - \mathbb{E}_n)$. Giné and Zinn (1990) show that for almost every realization of $\{W_i\}_{i=1}^n$, the bootstrapped empirical process $\{\mathbb{G}^\star_n A^t_{\theta,q} : (\theta, q) \in \Theta \times \mathcal{Q}\}$ converges weakly under $\mathbb{P}^\star$ to the same limiting process that $\mathbb{A}^t_n$ from Lemma 5 does.[26] The same argument as in Lemma 5 shows that under C4 this implies $\mathbb{G}^\star_n(A^t_{\theta_n,q_n}) - \mathbb{G}^\star_n(A^t_{\theta_0,q_0}) \to_{\mathbb{P}^\star} 0$ for any sequence $(\theta_n, q_n) \to_{\mathbb{P}^\star} (\theta_0, q_0)$. The claim then follows from an argument analogous to that in Lemma 5. *Q.E.D.*

Lemma $6^\star$. *Suppose Assumptions I, C1 and D3 hold. Then* $\{\sqrt{n}(D^\star_{\theta_0,q_0}(t) - \widehat{D}_{\theta_0,q_0}(t)) : t \in \mathcal{T}\}$ *converges weakly with respect to* $\mathbb{P}^\star$ *in* $l^\infty(\mathcal{T})$ *to a mean-zero Gaussian process with covariance function* σ. *Moreover, for every* t, $\sqrt{n}(D^\star_{\theta_0,q_0}(t) - \widehat{D}_{\theta_0,q_0}(t)) = \mathbb{G}^\star_n\chi(t) + o_{\mathbb{P}^\star}(1) + o_{a.s.}(1)$ *where the remainder terms are uniform over* $\mathcal{T}$.

Proof of Lemma $6^\star$. Using algebra analogous to that in the derivation of (26),

$$D^\star_{\theta_0,q_0}(t) = \mathbb{E}^\star_n(\chi(t)) - [\mathbb{Q}^\star_n(A^t_{\theta_0,q_0}) + \mathbb{Q}_n(A^t_{\theta_0,q_0})][\mathbb{Q}^\star_n(B^t) + \mathbb{Q}_n(B^t)], \tag{33}$$

[26] The measurability and square-integrable envelope requirements in Giné and Zinn (1990) are satisfied easily in this setting.

where $\mathbb{Q}_n^\star \equiv \mathbb{E}_n^\star - \mathbb{E}_n$. Combining (26) and (33), one has

$$\sqrt{n}\left(D^\star_{\theta_0,q_0}(t) - \widehat{D}_{\theta_0,q_0}(t)\right) = \mathbb{G}_n^\star(\chi(t)) + \mathbb{Q}_n(A^t_{\theta_0,q_0})\mathbb{G}_n(B^t) \tag{34}$$
$$- \sqrt{n}[\mathbb{Q}_n^\star(A^t_{\theta_0,q_0}) + \mathbb{Q}_n(A^t_{\theta_0,q_0})][\mathbb{Q}_n^\star(B^t) + \mathbb{Q}_n(B^t)].$$

Given the entropy analysis for $\{\chi(t) : t \in \mathcal{T}\}$ from Lemma 6 and the results of Giné and Zinn (1990), $\{\mathbb{G}_n^\star\chi(t) : t \in \mathcal{T}\}$ converges weakly under $\mathbb{P}^\star$ to a mean-zero Gaussian process with covariance function σ. Their results also imply that $\mathbb{Q}_n^\star(A^t_{\theta_0,q_0}) = O_{\mathbb{P}^\star}(n^{-1/2})$ and $\mathbb{Q}_n^\star(B^t) = O_{\mathbb{P}^\star}(n^{-1/2})$. Since $\mathbb{Q}_n(A^t_{\theta_0,q_0}) = o_{a.s.}(1)$ by C3 and $\mathbb{Q}_n(B^t) = o_{a.s.}(1)$ by the strong law of large numbers, the third term in (34) is $\sqrt{n}O_{\mathbb{P}^\star}(n^{-1}) = o_{\mathbb{P}^\star}(1)$. From Lemma 6 and the strong law of large numbers, the second term in (34) is $O_{\mathbb{P}}(1)o_{a.s.}(1) = o_{a.s.}(1)$. The result now follows from (34) and Slutsky's Theorem. The uniformity of the remainders over $t \in \mathcal{T}$ follows from the same arguments as in Lemma 6. *Q.E.D.*

Proof of Theorem 3. First, it needs to be verified that $\theta^\star \to_{\mathbb{P}^\star} \theta_0$. The argument for this is the same as in Theorem 1, except that it requires $\sup_{\theta\in\Theta,q\in\mathcal{Q}} \|D^\star_{\theta,q} - D_{\theta,q}\|_\mu \to_{\mathbb{P}^\star} 0$, i.e. a bootstrap counterpart of Lemma 3. Using Lemma 3 and the triangle inequality, this is implied by $\sup_{\theta\in\Theta,q\in\mathcal{Q}} \|D^\star_{\theta,q} - \widehat{D}_{\theta,q}\|_\mu \to_{\mathbb{P}^\star} 0$. The latter follows from a decomposition analogous to (24) in Lemma 3, using the result of Giné and Zinn (1990) that $\sup_{\theta,q} |\mathbb{G}_n^\star(A^t_{\theta_0,q_0}B^t)| \to_{\mathbb{P}^\star} 0$ under C1 and C3. The other details of the consistency argument are the same and so are omitted.

The rest of the proof is analogous to that for Theorem 2. In a supplementary document (available from the author on request), I show that $\theta^\star = \theta_0 + O_{\mathbb{P}^\star}(n^{-1/2}) + O_{\mathbb{P}}(n^{-1/2})$ by repeatedly applying Lemmas 4, 5, 5*, 6 and 6*, together with Assumptions D*. In this document it is also shown that the linear approximation

$$L^\star_\theta(t) = D^\star_{\theta_0,q_0}(t) + (\theta - \theta_0)'\Delta_0(t) + \Pi^{[q^\star - q_0]}_{\theta_0,q_0}(t)$$

provides an $o_{\mathbb{P}^\star}(n^{-1/2}) + o_{\mathbb{P}}(n^{-1/2})$ approximation to $D^\star_{\theta_n,q^\star}$ for sequences θ_n within $O_{\mathbb{P}^\star}(n^{-1/2}) + O_{\mathbb{P}}(n^{-1/2})$ of θ_0. An argument in this document further establishes that the minimizer of $\|L^\star_\theta\|_\mu$, call it $\widetilde{\theta}^\star$, satisfies

$$\sqrt{n}(\widetilde{\theta}^\star - \widehat{\theta}) = \sqrt{n}(\mathbb{E}^\star_n - \mathbb{E}_n)\xi + o_{\mathbb{P}^\star}(1) + o_{\mathbb{P}}(1),$$

with ξ defined as in the proof of Theorem 2.

Theorem 2.2 of Bickel and Freedman (1981) shows that $\sqrt{n}(\mathbb{E}^\star_n - \mathbb{E}_n)\xi \rightsquigarrow N(0, \overline{\Sigma}_0)$ with respect to $\mathbb{P}^\star$ a.s. ($\mathbb{P}$). Theorem 20.5(ii) of Billingsley (1995) can then be used to show that this implies that $\sqrt{n}(\widetilde{\theta}^\star - \widehat{\theta}) \rightsquigarrow N(0, \overline{\Delta}_0^{-1}\overline{\Sigma}_0\overline{\Delta}_0^{-1})$ in $\mathbb{P}^\star$–probability with $\mathbb{P}$–probability approaching 1. See Proposition O(xiii) of Hahn (1993) for a detailed statement and justification. The same argument as in Theorem 2 shows that $\sqrt{n}\|\theta^\star - \widetilde{\theta}^\star\| = o_{\mathbb{P}^\star}(1) + o_{\mathbb{P}}(1)$, so I omit the details. It follows that $\sqrt{n}(\theta^\star - \widehat{\theta}) = \sqrt{n}(\widetilde{\theta}^\star - \widehat{\theta}) + o_{\mathbb{P}^\star}(1) + o_{\mathbb{P}}(1)$, which establishes the claim of the theorem after appealing again to Hahn's (1993) Proposition O(xiii).

Q.E.D.

References

Abrevaya, J. (1999): "Computation of the maximum rank correlation estimator," *Economics Letters*, 62, 279–285. 9

Andrews, D. W. K. (1994): "Asymptotics for Semiparametric Econometric Models Via Stochastic Equicontinuity," *Econometrica*, 62, 43–72. 12

Bickel, P. J. and D. A. Freedman (1981): "Some Asymptotic Theory for the Bootstrap," *The Annals of Statistics*, 9, 1196–1217. 19, 41

Billingsley, P. (1995): *Probability and measure*, New York [u.a.]: Wiley. 41

Brown, D. J. and R. Matzkin (1998): "Estimation of Nonparametric Functions in Simultaneous Equations Models, With an Application to Consumer Demand," *Cowles Foundation Discussion Paper 1175*. 7

Brown, D. J. and M. H. Wegkamp (2002): "Weighted Minimum Mean-Square Distance from Independence Estimation," *Econometrica*, 70, 2035–2051. 6, 7, 8, 14, 22

Card, D. (1995): "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, ed. by L. N. Christofides, K. E. Grant, and R. Swidinsky, Toronto: University of Toronto Press, 201–222. 4, 23, 24

——— (2001): "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 69, 1127–1160. 4

Carrasco, M. and J.-P. Florens (2000): "Generalization of GMM to a Continuum of Moment Conditions," *Econometric Theory*, 16, 797–834. 16

Chaudhuri, P., K. Doksum, and A. Samarov (1997): "On Average Derivative Quantile Regression," *The Annals of Statistics*, 25, 715–744. 19, 20

Chen, X., O. Linton, and I. van Keilegom (2003): "Estimation of Semiparametric Models When the Criterion Function Is Not Smooth," *Econometrica*, 71, 1591–1608. 12, 34

Chen, X. and D. Pouzo (2012): "Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals," *Econometrica*, 80, 277–321. 20

Chernozhukov, V., I. Fernández-Val, and B. Melly (2009): "Inference on counterfactual distributions," *Cemmap working paper CWP09/09*. 19

Chesher, A. (2003): "Identification in Nonseparable Models," *Econometrica*, 71, 1405–1441. 2

D'Haultfœuille, X. and P. Février (2015): "Identification of Nonseparable Triangular Models With Discrete Instruments," *Econometrica*, 3, 1199–1210. 3

Dominguez, M. A. and I. N. Lobato (2004): "Consistent Estimation of Models Defined by Conditional Moment Restrictions," *Econometrica*, 72, 1601–1615. 20

Florens, J. P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of Treatment Effects Using Control Functions in Models With Continuous, Endogenous Treatment and Heterogeneous Effects," *Econometrica*, 76, 1191–1206. 2

Giné, E. and J. Zinn (1990): "Bootstrapping General Empirical Measures," *The Annals of Probability*, 18, 851–869. 17, 18, 39, 40

Hahn, J. (1993): "Three Essays in Econometrics," Ph.D. thesis, Harvard University. 41

——— (1996): "A Note on Bootstrapping Generalized Method of Moments Estimators," *Econometric Theory*, 12, 187–197. 17

Heckman, J. J. (2001): "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *The Journal of Political Economy*, 109, 673–748. 2

Holmström, K., A. O. Göran, and M. M. Edvall (2010): *User's Guide for TOMLAB 7*. 9

Imbens, G. W. (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *The Review of Economics and Statistics*, 86, 4–29. 2

Imbens, G. W. and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77, 1481–1512. 2

Jones, D. R., C. D. Perttunen, and B. E. Stuckman (1993): "Lipschitzian optimization without the Lipschitz constant," *Journal of Optimization Theory and Applications*, 79, 157–181. 9

Koenker, R. (2005): *Quantile Regression*, Cambridge University Press. 2, 14

Koenker, R. and G. Bassett (1978): "Regression Quantiles," *Econometrica*, 46, 33–50. 19

Komunjer, I. and A. Santos (2010): "Semi-parametric estimation of non-separable models: a minimum distance from independence approach," *Econometrics Journal*, 13, S28–S55. 6, 7

Kosorok, M. R. (2008): *Introduction to empirical processes and semiparametric inference*, New York: Springer. 33, 38

Linton, O., S. Sperlich, and I. van Keilegom (2008): "Estimation of a Semiparametric Transformation Model," *The Annals of Statistics*, 36, 686–718. 6

Luenberger, D. G. (1968): *Optimization by vector space methods*, New York: Wiley. 36, 38

Manski, C. F. (1983): "Closest Empirical Distribution Estimation," *Econometrica*, 51, 305–319. 6, 7

Matzkin, R. L. (2003): "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 71, 1339–1375. 2, 21

Nelsen, R. (2006): *An introduction to copulas*, New York: Springer. 26

Newey, W. K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382. 12, 14, 20

Newey, W. K. and D. McFadden (1994): "Chapter 36 Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, ed. by R. F. Engle and D. L. McFadden, Elsevier, vol. Volume 4, 2111–2245. 11

Pakes, A. and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027–1057. 12, 34, 37

Pollard, D. (1984): *Convergence of stochastic processes*, New York: Springer-Verlag. 15, 31

Resnick, S. I. (1999): *A Probability Path*, Birkhäuser Boston. 33

Santos, A. (2011): "Semiparametric Estimation of Invertible Models," *Working paper*. 6

Torgovitsky, A. (2010): "Identification and Estimation of Nonparametric Quantile Regressions with Endogeneity," *Job market paper*. 3

——— (2015): "Identification of Nonseparable Models Using Instruments With Small Support," *Econometrica*, 83, 1185–1197. 2, 3, 4, 5, 6, 24

van der Vaart, A. W. (1998): *Asymptotic statistics*, New York: Cambridge University Press. 18

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes : With Applications to Statistics.*, New York: Springer. 15, 31, 33, 34, 38, 39

			$N = 400$			$N = 800$			$N = 1600$		
	γ	$\lvert\mathcal{Z}\rvert$	bias	(std)	mse	bias	(std)	mse	bias	(std)	mse
θ_1	.25	2	-.0372	(.0608)	.0051	-.0180	(.0416)	.0021	-.0082	(.0272)	.0008
		4	-.0577	(.0649)	.0075	-.0303	(.0446)	.0029	-.0156	(.0294)	.0011
		8	-.0628	(.0653)	.0082	-.0345	(.0457)	.0033	-.0176	(.0288)	.0011
	.35	2	-.0185	(.0418)	.0021	-.0086	(.0285)	.0009	-.0041	(.0190)	.0004
		4	-.0314	(.0468)	.0032	-.0147	(.0300)	.0011	-.0078	(.0195)	.0004
		8	-.0348	(.0480)	.0035	-.0165	(.0306)	.0012	-.0087	(.0195)	.0005
θ_2	.25	2	.1851	(.3661)	.1683	.0801	(.2418)	.0649	.0423	(.1556)	.0260
		4	.2870	(.3783)	.2255	.1358	(.2506)	.0812	.0727	(.1659)	.0328
		8	.3148	(.3748)	.2396	.1553	(.2532)	.0882	.0782	(.1653)	.0334
	.35	2	.0922	(.2430)	.0676	.0388	(.1623)	.0278	.0217	(.1082)	.0122
		4	.1517	(.2651)	.0933	.0669	(.1692)	.0331	.0362	(.1111)	.0137
		8	.1681	(.2605)	.0961	.0732	(.1706)	.0344	.0383	(.1132)	.0143
θ_3	.25	2	.0132	(.0858)	.0075	.0108	(.0499)	.0026	.0021	(.0338)	.0011
		4	.0223	(.0810)	.0071	.0185	(.0553)	.0034	.0081	(.0358)	.0014
		8	.0236	(.0848)	.0077	.0209	(.0558)	.0036	.0108	(.0351)	.0014
	.35	2	.0065	(.0556)	.0031	.0050	(.0342)	.0012	.0007	(.0238)	.0006
		4	.0140	(.0569)	.0034	.0086	(.0381)	.0015	.0042	(.0241)	.0006
		8	.0157	(.0575)	.0036	.0106	(.0379)	.0015	.0056	(.0230)	.0006

Table 1: Monte Carlo results for the performance of $\widehat{\theta}$. The strength of the instrument is controlled by γ. The number of points of support for the instrument is denoted as $|\mathcal{Z}|$. There were 500 replications for each experiment.

	nominal level		
N	.990	.950	.900
100	.991	.947	.913
200	.992	.948	.895
400	.985	.945	.893

Table 2: Actual coverage probabilities for bootstrap confidence intervals of θ_2. These experiments are the result of 1000 replications with 500 bootstrap samples for each replication.

					MDIV			
x	t	OLS	OLS^2	IV	(1)	(2)	(3)	(4)
		.0437	.0448	.1497	.1700	.1732	.1325	.1984
12	0.10	.0501	.0546	.2167	.2258	.2308	.2042	.2281
		.0565	.0644	.2837	.3066	.3119	.3285	.3815
							.1525	.1880
	0.25	—"—	—"—	—"—	—"—	—"—	.2137	.2297
							.3116	.3439
							.1681	.1767
	0.50	—"—	—"—	—"—	—"—	—"—	.2242	.2314
							.3027	.3211
							.1713	.1673
	0.75	—"—	—"—	—"—	—"—	—"—	.2350	.2331
							.3204	.3086
							.1658	.1530
	0.90	—"—	—"—	—"—	—"—	—"—	.2433	.2347
							.3496	.2948
			.0425			.1325		.1294
14	0.10	—"—	.0491	—"—	—"—	.1963	—"—	.1853
			.0557			.2948		.3417
								.1319
	0.25	—"—	—"—	—"—	—"—	—"—	—"—	.1869
								.3083
								.1217
	0.50	—"—	—"—	—"—	—"—	—"—	—"—	.1886
								.2738
								.1122
	0.75	—"—	—"—	—"—	—"—	—"—	—"—	.1903
								.2539
								.1024
	0.90	—"—	—"—	—"—	—"—	—"—	—"—	.1919
								.2417
			.0311			.0618		.0666
16	0.10	—"—	.0437	—"—	—"—	.1617	—"—	.1425
			.0562			.3030		.3414
								.0554
	0.25	—"—	—"—	—"—	—"—	—"—	—"—	.1441
								.3034
								.0435
	0.50	—"—	—"—	—"—	—"—	—"—	—"—	.1458
								.2625
								.0322
	0.75	—"—	—"—	—"—	—"—	—"—	—"—	.1475
								.2383
								.0221
	0.90	—"—	—"—	—"—	—"—	—"—	—"—	.1491
								.2329

Table 3: Estimated marginal effects for the empirical illustration evaluated at various combinations of x and $Q_\varepsilon(t)$, where the latter is estimated from the quantiles of $g_{\widehat{\theta}}^{-1}(X_i, Y_i)$. The large number in each cell is the point estimate and the small numbers are lower and upper bounds of 95% confidence regions. The ditto marks (—"—) indicate that by assumption the model would predict the same marginal effects as in the cell above it.