Estimating a Nonparametric Triangular Model with Binary Endogenous Regressors[*†]

Sung Jae Jun,[‡] Joris Pinkse[§] and Haiqing Xu[¶]

Center for the Study of Auctions, Procurements and Competition Policy

Department of Economics

The Pennsylvania State University and the University of Texas

May 2, 2016

We consider identification and estimation in a nonparametric triangular system with a binary endogenous regressor and nonseparable errors. For identification we take a control function approach utilizing the Dynkin system idea developed in Jun, Pinkse, and Xu (2011, JPX11) and extended in Kédagni and Mourifié (2014, KM14). We articulate various tradeoffs, including continuity, monotonicity, and differentiability. For estimation, we use the idea of local instruments under smoothness assumptions, as in e.g. Carneiro and Lee (2009, CL09), Heckman and Vytlacil (1999) but we do not assume additive separability in latent variables. Our estimator uses nonparametric kernel regression techniques and its statistical properties are derived using the functional delta method. We establish that it is $n^{2/7}$–consistent and has a limiting normal distribution. We apply the method to estimate the returns on a college education. Unlike existing work, notably CL09 and Carneiro, Heckman, and Vytlacil (2011, CHV11), we find that returns on a college education are consistently positive. The returns curves we estimate are moreover inconsistent with the shape restrictions imposed in those papers.

Key Words: Triangular Models; Endogeneity; Nonparametric Estimation.

JEL Classification Codes: C14; C31; C35; C36.

---

[*]This paper is based on research supported by NSF grant SES–0922127. We thank the Human Capital Foundation for their support of the Penn State economics graduate program. We thank Ismael Mourifië and seminar participants at the 2014 Canadian Econometrics Study Group and departmental seminars at Georgetown, Indiana, LSE, Michigan, Stanford, and UCL.

[†]This research was conducted with restricted access to Bureau of Labor Statistics (BLS) data. The views expressed here do not reflect the views of the BLS.

[‡](corresponding author) 303 Kern Graduate Building, University Park 16802, sjun@psu.edu

[§]joris@psu.edu

[¶]h.xu@austin.utexas.edu

1. Introduction

We consider identification and estimation in a nonparametric triangular system with a binary endogenous regressor and nonseparable errors. For identification we take a control function approach utilizing the Dynkin system idea developed in Jun, Pinkse, and Xu (2011, JPX11) and extended in Kédagni and Mourifié (2014, KM14). We articulate various tradeoffs, including continuity, monotonicity, and differentiability. For estimation, we use the idea of local instruments under smoothness assumptions, as in e.g. Carneiro and Lee (2009, CL09), Heckman and Vytlacil (1999) but we do not assume additive separability in latent variables. Our estimator uses nonparametric kernel regression techniques and its statistical properties are derived using the functional delta method. We establish that it is $n^{2/7}$–consistent and has a limiting normal distribution. We apply the method to estimate the returns on a college education. Unlike existing work, notably CL09 and Carneiro, Heckman, and Vytlacil (2011, CHV11), we find that returns on a college education are consistently positive. The returns curves we estimate are moreover inconsistent with the shape restrictions imposed in those papers.

There is a related–but–different approach in the literature, namely the instrumental variable (IV) quantile model of e.g. Chernozhukov and Hansen (2005, 2006). Unlike the control function approach, the IV quantile model does not treat discrete endogenous regressors differently. Therefore, binary endogenous regressors do not cause any extra issues in identification and estimation. However, the control function approach is more suitable when it is of interest to explicitly control for potential heterogeneity from the *first* stage equation.

Most papers on nonparametric triangular models with nonseparable errors focus on nonparametric identification rather than estimation, especially when there is a discrete endogenous regressor. Hoderlein and Mammen (2007) consider a nonseparable model with exogenous regressors but without monotonicity. Chesher (2003, CH03) and Imbens and Newey (2009) study nonseparable models with continuous endogenous regressors under monotonicity. Based on the identification result of e.g. CH03, Ma and Koenker (2006) and Jun (2009) propose parametric and semiparametric estimation methods, respectively, but both require that endogenous regressors be continuous. Chesher (2005, CH05) establishes partial identification of the structural function at a given value in a triangular system with a discrete endogenous regressor, but CH05 contains little discussion on estimation and inference. JPX11 reconsider CH05's result and provide tighter bounds under a weaker rank condition using an independence assumption on instruments, but also stop short of estimation and inference.

A related but different literature uses an instrumental variable (IV) quantile approach (e.g. Chernozhukov and Hansen, 2005, 2006). Unlike the control function approach, the IV quantile model treats discrete endogenous regressors in the same way it treats continuous endogenous regressors. Therefore, binary endogenous regressors do not cause additional issues in identification and estimation. However, we use the control function approach, because it allows us to control for potential heterogeneity arising from the *first* stage equation explicitly.

JPX11 and CH05 serve as our starting point. The two papers study the same model, albeit that JPX11 use a global independence condition of instruments and errors to weaken CH05's rank condition and to tighten identification bounds. The difference between the two approaches is most profound in the presence of a binary endogenous regressor since CH05, unlike JPX11, does not allow for a binary endogenous regressor. JPX11 in fact establish conditions under which point identification obtains in the case of continuous instruments.

One of our objectives in this paper is to show how the Dynkin system idea can be used to obtain point identification in the triangular model with a binary endogenous regressor. Unlike the local instrument approach of Heckman and Vytlacil (1999, 2001) and CL09, we show that it is the continuous variation, rather than the differentiability, of the propensity score that is essential for identification, albeit that the latter is useful for estimation. A more formal discussion can be found in section 2.

Once we have articulated conditions for identification we turn our attention to nonparametric estimation. We propose a kernel–based nonparametric estimator, which allows for the full flexibility of the triangular model with nonseparable errors. We then develop limit results for the proposed estimator. These results are derived in section 3.

Finally, in section 5 we implement our estimator using the same NLSY–based data set that is used in CL09 and CHV11 and an index model described in section 4. Our specification is different from theirs and so are our conclusions. CL09 and CHV11 find that returns to a college education can be negative whereas we find them to be consistently positive and substantial. Further, the returns curves that we estimate do not satisfy the additive separability of errors assumption imposed by CL09 and CHV11, albeit that the specification in CL09 is not nested by ours. Nevertheless, the shape restrictions in the existing literature appear to be restrictive and should be studied in greater depth with a larger data set.

2. Identification

JPX11 show that the identified bounds provided in CH05 can be substantially tightened under a weaker rank condition when instruments are independent of the errors in a two equation triangular system. In an extreme case with independent and continuous instruments, the structural function evaluated at particular values of its arguments can even be point–identified. In this section we show that this general result is in fact closely related to existing results on the identification of treatment effects (e.g., CL09 and Heckman and Vytlacil (1999, 2001)) and also to the results for continuous endogenous regressors of CH03.

The approach taken in JPX11 and CH05 is general in that partial identification is discussed under the setup of a triangular system with discrete endogenous regressors. The source of the weaker rank condition and the tighter bounds of JPX11 is the independence between instrumental variables and unobserved errors, which makes it possible to combine multiple values of the instrumental variables to obtain identified bounds. The idea is best explained when an endogenous regressor is binary, which is the case we focus on in the current paper.

Consider the model

$$\begin{cases} \boldsymbol{y} &= g(\boldsymbol{x}, \boldsymbol{u}), \\ \boldsymbol{x} &= \phi(\boldsymbol{z}, \boldsymbol{v}), \end{cases} \tag{1}$$

where $\boldsymbol{x}$ is a binary regressor, $\boldsymbol{z} \in \mathcal{S}_z \subset \mathbb{R}^{d_z}$ is a vector of observed 'demographics,' and $\boldsymbol{y}$ is a scalar–valued outcome.[1] We omit covariates in the identification analysis but they will be introduced in the estimation section. We permit the errors to enter nonadditively, so $g(1, \boldsymbol{u}) - g(0, \boldsymbol{u})$ can vary with $\boldsymbol{u}$. Consequently there is a distinction between the difference between (conditional) quantiles of $g(1, \boldsymbol{u})$ and $g(0, \boldsymbol{u})$ and the (conditional) quantiles of the difference $g(1, \boldsymbol{u}) - g(0, \boldsymbol{u})$. We follow Doksum (1974) and many others and focus on the former.

Thus, for generic random variables $\boldsymbol{a}$ and $\boldsymbol{b}$ let $\mathbb{Q}_{\boldsymbol{a}|\boldsymbol{b}}(\tau|b)$ be the τ quantile of $\boldsymbol{a}$ given $\boldsymbol{b} = b$. The parameter of interest is

$$\psi^* = \psi^*(x^*, \tau^*|v^*) = g\{x^*, \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v^*)\} = \mathbb{Q}_{g(x^*,\boldsymbol{u})|\boldsymbol{v}}(\tau^*|v^*), \tag{2}$$

[1] Our analysis can be adapted to the potential outcome framework, where the two potential outcomes may have different errors. There is no loss of generality in assuming a scalar–valued error in a potential outcome variable by the Cantor–Schroeder–Bernstein theorem. But it becomes restrictive once monotonicity conditions are imposed, as we will do later. For a detailed discussion of issues in models with vector–valued errors, see Kasy (2011).

for given values of x^*, τ^*, v^*, where the last equality follows from assumptions D and E below.[2] The function ψ^* can be used to define causal parameters of interest. For instance, we will call

$$\text{QMTE}(\tau^*|v^*) = \psi^*(1, \tau^*|v^*) - \psi^*(0, \tau^*|v^*) \tag{3}$$

the *quantile marginal treatment effect*, which is the quantile version of the marginal treatment effect of e.g. Heckman and Vytlacil (2005). Integrating QMTE over τ^* yields the marginal treatment effect (MTE). Integrating the MTE over v^* with various weight functions is discussed in Heckman and Vytlacil (2005), and for one such choice results in the average treatment effect.

We make the following model assumptions, which are based on those in JPX11 and CH05.

Assumption A. $\boldsymbol{u}, \boldsymbol{v}$ *have (marginal)* $U(0,1]$*–distributions.*

Assumption B. $\boldsymbol{u}, \boldsymbol{v}$ *are independent of* $\boldsymbol{z}$.

Assumption C. $\phi(z, v)$ *is left–continuous and nondecreasing in* v *for all values of* z.

Assumption D. $g(x^*, u)$ *is nondecreasing in* u *on* $(0, 1]$.

Assumption E. $g(x^*, u)$ *is left–continuous in* u *at* $\mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v^*)$.

Assumption F. *For all* $\tau \in (0, 1]$, $\mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau|v)$ *is nondecreasing in* v.

Assumption A is fairly standard in the literature and is essentially a normalization. Assumption B is strong, but indispensable here. The conditions on ϕ in assumption C are also common. Assumptions A to C imply that one can represent the relationship between $\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{v}$ as $\boldsymbol{x} = \mathbb{1}\{\boldsymbol{v} > p(\boldsymbol{z})\}$, where $\mathbb{1}$ is the indicator function and $p(z) = \mathbb{P}(\boldsymbol{x} = 0|\boldsymbol{z} = z)$ is one minus the propensity score; see e.g. Vytlacil (2006). Assumptions D and E are needed for the last equality in (2), as noted before. Note that assumption D is weaker than strict monotonicity of g in u on $(0, 1]$. In particular, if for instance $\boldsymbol{y}$ represents earnings then assumptions D and E allow for the case where there is a mass point at the minimum wage and the minimum wage is below the desired quantile. Note also that $g(x^*, \cdot)$ is allowed to have a discontinuous jump at $\mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v^*)$.

The positive dependence condition in assumption F is used in both JPX11 and CH05.[3] We use assumption F to obtain identifiable bounds for ψ^*, but it is not needed to establish point identification of ψ^* if there are continuous instruments and $g(x^*, u)$ is continuous at $\mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v^*)$.

[2] We define quantiles in the standard way, i.e. $\mathbb{Q}_{\boldsymbol{a}|\boldsymbol{b}}(\tau|b) = \inf\{a : \mathbb{P}(\boldsymbol{a} \leq a|\boldsymbol{b} = b) \geq \tau\}$.

[3] Negative dependence can be dealt with similarly. The essence of this assumption is the monotonicity of $\mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau|\cdot)$.

Let V_L, V_U be arbitrary subsets (of positive measure) of $(0, v^*]$ and $(v^*, 1]$, respectively. Then assumptions A, D and F imply that

$$g\{x^*, \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|V_L)\} \leq \psi^* \leq g\{x^*, \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|V_U)\}, \tag{4}$$

which can be seen by inverting the inequality e.g.

$$\begin{aligned}\mathbb{P}\{g(x^*, \boldsymbol{u}) \leq y|\boldsymbol{v} \in V_U\} &= \frac{1}{\mathbb{P}(\boldsymbol{v} \in V_U)} \int_{V_U} \mathbb{P}\{g(x^*, \boldsymbol{u}) \leq y|\boldsymbol{v} = v\} dv \\ &\leq \frac{1}{\mathbb{P}(\boldsymbol{v} \in V_U)} \int_{V_U} \mathbb{P}\{g(x^*, \boldsymbol{u}) \leq y|\boldsymbol{v} = v^*\} dv = \mathbb{P}\{g(x^*, \boldsymbol{u}) \leq y|\boldsymbol{v} = v^*\}.\end{aligned}$$

The bounds in (4) are discussed in detail in CH05 and JPX11. Since it is important for our discussion to understand when these bounds are identified, we briefly discuss CH05 and JPX11 focusing on the case $x^* = 0$.

Let $\mathcal{V}(0) = \{(0, p(z)] : z \in \mathcal{S}_z\}$ and $\mathcal{V}(1) = \{(p(z), 1] : z \in \mathcal{S}_z\}$. If $V_L \in \mathcal{V}(0)$ then assumptions B, D and E imply that

$$g\{0, \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|V_L)\} = \mathbb{Q}_{g(0,\boldsymbol{u})|\boldsymbol{v}}(\tau^*|V_L) = \mathbb{Q}_{\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z}}(\tau^*|0, z),$$

which is identified, where $z \in \mathcal{S}_z$ is such that $V_L = (0, p(z)]$. However, exclusively relying on sets of the form $(0, p(z)]$ leads to a trivial upper bound of the identified set because there is no set of that form that lies in its entirety above v^*. Similarly, relying on $\mathcal{V}(1)$ leads to a trivial lower bound in the case of $x^* = 1$. CH05 stops here and interprets this problem as a violation of his rank condition.

JPX11 go on to show that the bounds in (4) are identified when sets not belonging to $\mathcal{V}(0)$ are utilized. For instance, suppose that there exist z and $\tilde{z}$ in $\mathcal{S}_z$ such that $v^* \leq p(\tilde{z}) < p(z)$. Then $(p(\tilde{z}), p(z)] = (0, p(z)] - (0, p(\tilde{z})] \geq v^*$, i.e. all elements in the majorant side set are no less than v^*. Hence one can choose $V_U = (p(\tilde{z}), p(z)]$ in (4) to obtain an upper bound, namely the τ^* quantile of the conditional distribution given by

$$\begin{aligned}&\mathbb{P}\{g(0, \boldsymbol{u}) \leq y|\boldsymbol{v} \in V_U\} \\ &\quad = \frac{1}{p(z) - p(\tilde{z})}\{\mathbb{P}(\boldsymbol{y} \leq y|\boldsymbol{x} = 0, \boldsymbol{z} = z)p(z) - \mathbb{P}(\boldsymbol{y} \leq y|\boldsymbol{x} = 0, \boldsymbol{z} = \tilde{z})p(\tilde{z})\}.\end{aligned} \tag{5}$$

A Dynkin system $\mathcal{D}(x^*)$ generated by $\mathcal{V}(x^*)$ can be obtained by applying various set operations to $\mathcal{V}(x^*)$ and ensures that $g\{x^*, \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|V)\}$ is identified whenever $V \in \mathcal{D}(x^*)$. Such a Dynkin

system can used to identify the tightest bounds in (4). The following definition was first introduced in JPX11.

Definition 1 (Dynkin System, JPX11)**.** *A Dynkin system $\mathcal{D}(x^*)$ is defined by the collection $\mathcal{D}_\infty$ in the following iterative scheme. Let $\mathcal{D}_0 = \mathcal{V}(x^*)$. Then for all $t \geq 0$, $\mathcal{D}_{t+1}$ consists of all sets A^* such that at least one of the following three conditions is satisfied.*

(i) $A^* \in \mathcal{D}_t$,

(ii) $\exists A_1, A_2 \in \mathcal{D}_t : A_1 \subset A_2,\ \mu(A_2 - A_1) > 0,\ A^* = A_2 - A_1$,

(iii) $\exists A_1, A_2 \in \mathcal{D}_t : A_1 \cap A_2 = \emptyset,\ \mu(A_1 \cup A_2) > 0,\ A^* = A_1 \cup A_2$.

Since $\{\mathcal{D}_t : t = 0, 1, \cdots\}$ is an increasing sequence of collections of sets, we have $\mathcal{D}(x^*) = \cup_{t=0}^{\infty} \mathcal{D}_t$. It can be shown that the conditional distribution function of $g(x^*, \boldsymbol{u})$ given $\boldsymbol{v} \in V$ is identified whenever $V \in \mathcal{D}(x^*)$.

Let $\mathcal{D}_L(x^*, v^*) = \{V \in \mathcal{D}(x^*) : V \leq v^*\}$ and let $\mathcal{D}_U(x^*, v^*)$ be similarly defined. Then JPX11 have shown[4] that under assumptions A to F identified bounds for ψ^* are given by

$$\sup_{V \in \mathcal{D}_L(x^*, v^*)} g\{x^*, \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|V)\} \leq \psi^* \leq \inf_{V \in \mathcal{D}_U(x^*, v^*)} g\{x^*, \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|V)\}. \tag{6}$$

KM14 have shown that there are circumstances under which the above bounds can be tightened further.

We now discuss how additional continuity conditions can be used to obtain the point identification of ψ^*.

Assumption G. *$\mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v)$ is left–continuous in v at v^*.*

Assumption H. *There exists a sequence $\{z_t\}$ in $\mathcal{S}_z$ such that $p(z_t)$ is (strictly) increasing in t with supremum v^*.*

Left–continuity in assumptions E and G can be replaced with right–continuity. Please note that left–continuity does not rule out the presence of discrete jumps in the function g and it hence allows for mass points in the distribution of $\boldsymbol{y}$. A sufficient condition for assumption H is that at least one element of the vector of instruments $\boldsymbol{z}$ has continuous variation (given the other elements) and p is continuous in that element of $\boldsymbol{z}$.

We can now strengthen the result in (6).

[4]The conditions in JPX11 are slightly different from the ones here.

Theorem 1. *(i) Suppose that assumptions E to H are satisfied. Then*

$$\sup_{V_L \in \mathcal{D}_L(x^*,v^*)} g\{x^*, \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|V_L)\} = \psi^* \leq \inf_{V_U \in \mathcal{D}_U(x^*,v^*)} g\{x^*, \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|V_U)\}. \tag{7}$$

(ii) If moreover assumptions A to D are satisfied then ψ^* is point–identified.

Theorem 1 implies that the Dynkin system approach of JPX11 can be used to achieve point identification of ψ^* using continuous variation in the propensity score $p(\boldsymbol{z})$. This result can be compared to CH03 where point identification is achieved assuming strong monotonicity of $\phi(z, v)$ in v, which implies that $\boldsymbol{x}$ is continuously distributed. If left–continuity (assumptions E and G) is strengthened to continuity then the inequality in (7) becomes an equality such that the intersection bounds in (6) collapse to ψ^*.

Continuity of $g(x^*, u)$ at $u = \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v^*)$ is more helpful for the identification of ψ^* than continuity of $\mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v)$ at v^*. Indeed, continuity of $\mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v)$ is of at most modest help to relax assumption H whereas continuity of $g(x^*, u)$ obviates the need for monotonicity assumptions on $\mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau|v)$ in v (assumption F) for the identification of ψ^*, as theorem 2 demonstrates. However, as we have pointed out before, assuming continuity rather than left–continuity of g will impose restrictions on the support of the outcome, which would undesirable.

Assumption I. *$g(x^*, u)$ is continuous at $\mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v^*)$.*

Theorem 2. *Suppose that assumptions A to D and G to I are satisfied. Then ψ^* is identified.*

Please note that theorems 1 and 2 rely on monotonicity/continuity but not on smoothness (i.e. differentiability). We now discuss how these results are connected with the existing results of *identification via local instruments* in the treatment effects literature.

It is well–known in the treatment effects literature that differentiability can result in point identification of the distribution functions of counterfactual outcomes conditional on $p(\boldsymbol{z}) = v^*$.[5] We now explain how this treatment effects literature result is related to the results in theorems 1 and 2. Let $\partial_z p(z^*)$ be the partial derivative of p with respect to z at z^* and $G^*(y|x^*, z) = \mathbb{P}(\boldsymbol{y} \leq y|\boldsymbol{x} = x^*, \boldsymbol{z} = z)$.

Assumption J. *For any $y \in \mathbb{R}$, $G^*(y|x^*, z)$ is continuously differentiable in z at z^*.*

Assumption K. *For some z^* in the interior of $\mathcal{S}_z$, (i) $p(z^*) = v^*$ and (ii) $\partial_z p(z^*) \neq 0$.*[6]

[5] See e.g. CL09 and Heckman and Vytlacil (1999, 2001).

[6] Please note that $\partial_z p$ is vector–valued and we only require one of its elements to be nonzero.

It is useful to compare assumptions E, G and H with assumptions J and K. If the propensity score is differentiable then it follows from (8) that assumption J is equivalent to continuity of $F_{\boldsymbol{u}|\boldsymbol{v}}(u|v)$ in v at v^* and indeed to the differentiability of $F_{\boldsymbol{uv}}(u,v)$ in v at v^*.

We now show that the smoothness conditions in assumptions J and K provide an alternative path to identification. Suppose that $\boldsymbol{z}$ is scalar–valued. Since $\boldsymbol{x} = 0$ and $\boldsymbol{z} = z$ is equivalent to $\boldsymbol{v} \in \big(0, p(z)\big]$ and $\boldsymbol{z} = z$, we have that for any $y \in \mathbb{R}$,

$$G^*(y|x^*, z) = \mathbb{P}\big\{g(0,\boldsymbol{u}) \leq y | \boldsymbol{v} \in (0, p(z)]\big\} = \frac{1}{p(z)} \int_0^{p(z)} \mathbb{P}\big\{g(0,\boldsymbol{u}) \leq y | \boldsymbol{v} = v\big\} dv. \tag{8}$$

Differentiating both sides in (8) and evaluating at z^* yields

$$\mathbb{P}\big\{g(0,\boldsymbol{u}) \leq y | \boldsymbol{v} = v^*\big\} = G^*(y|x^*, z^*) + v^* \frac{\partial_z G^*(y|x^*, z^*)}{\partial_z p(z^*)}. \tag{9}$$

The right hand side in (9) is identified and ψ^* is defined as the smallest value of y for which the left hand side in (9) is equal to τ^*. An expression similar to (9) can be found in CL09.

For vector–valued $\boldsymbol{z}$ it is more natural to work with the propensity score.[7] Thus, let $G(y|x, p) = \mathbb{P}\{\boldsymbol{y} \leq y | \boldsymbol{x} = x, p(\boldsymbol{z}) = p\}$ so that assumption B implies $G^*(y|x, z) = G\{y|x, p(z)\}$. Then, as was shown by CL09, we have

$$\begin{aligned} \mathbb{P}\big\{g(0,\boldsymbol{u}) \leq y | \boldsymbol{v} = v^*\big\} &= G(y|0, v^*) + v^* \partial_p G(y|0, v^*), \\ \mathbb{P}\big\{g(1,\boldsymbol{u}) \leq y | \boldsymbol{v} = v^*\big\} &= G(y|1, v^*) - (1 - v^*) \partial_p G(y|1, v^*). \end{aligned} \tag{10}$$

Theorem 3. *If assumptions A to D, J and K are satisfied then ψ^* is identified.*

Theorems 1 to 3 articulate a trade–off between monotonicity, continuity, and smoothness assumptions. Continuity of $F_{\boldsymbol{u}|\boldsymbol{v}}(u|v)$ in v and differentiability of the propensity score are convenient for estimation but neither condition is necessary for identification.

Finally, we note that the Dynkin system idea in theorems 1 and 2 has applications far beyond the simple binary endogenous variable model of this paper: see e.g. JPX11; Jun, Pinkse, and Xu (2012); Jun, Pinkse, Xu, and Yildiz (2010).

3. Estimation

3.1. **Assumptions.** We now proceed to describe and motivate our estimation procedure, for which we will focus on the case $x^* = 0$. We add a subscript i to $\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{u}, \boldsymbol{v}$ and assume that we have an

[7] Assumption J is sufficient but not necessary for the differentiability of $G(y|x, \cdot)$.

i.i.d. sample of size n. We allow for the presence of exogenous covariates $\boldsymbol{a}_i \in \mathbb{R}^{d_a}$ in the function g, i.e. we now consider

$$\boldsymbol{y}_i = g(\boldsymbol{x}_i, \boldsymbol{a}_i, \boldsymbol{u}_i). \tag{11}$$

The covariates $\boldsymbol{a}_i$ are contained in the vector of instruments $\boldsymbol{z}_i$, which contains one or more additional elements $\boldsymbol{q}_i$ and is assumed to be independent of $\boldsymbol{u}_i, \boldsymbol{v}_i$, as is formally assumed here. However, in our proofs, we explicitly allow for the possibility that $\boldsymbol{a}_i$ and $\boldsymbol{z}_i$ have the same dimension, because in the semiparametric version of our estimator introduced in section 4 we consider $\boldsymbol{a}_i^\intercal \theta_0$ and $\boldsymbol{z}_i^\intercal \gamma_0$ (with $\boldsymbol{a}_i$ a subvector of $\boldsymbol{z}_i$) in lieu of $\boldsymbol{a}_i$ and $\boldsymbol{z}_i$, respectively.

Assumption L. *Assumptions A to D and K are for some q^* satisfied with $\boldsymbol{z}_i = [\boldsymbol{q}_i^\intercal, \boldsymbol{a}_i^\intercal]^\intercal$, $z^* = [q^{*\intercal}, a^{*\intercal}]^\intercal$, and with $g(x^*, a^*, u)$ in lieu of $g(x^*, u)$.*

Let $\mathcal{F}_j : \mathbb{R}^{d_z} \to \mathbb{R}$ denote the class of functions which are j times continuously differentiable on $\mathcal{Z}$ and $j + 2$ times boundedly differentiable with respect to z_1. We replace assumption J with the stronger assumption M.

Assumption M. $G^*(y|0, z) \in \mathcal{F}_2$.

The addition of the covariates $\boldsymbol{a}_i$ does not complicate the identification argument much. Indeed, one can simply condition on $\boldsymbol{a}_i = a^*$ in which case the entire argument of section 2 can be repeated with $\boldsymbol{q}_i$ assuming the role of $\boldsymbol{z}_i$. For estimation we adopt the identification argument of theorem 3, because it is the most convenient. Assumption M is introduced to obtain the desired convergence rate.

Thus, let $\boldsymbol{b}_i = [\boldsymbol{p}_i, \boldsymbol{a}_i^\intercal]^\intercal$, $\boldsymbol{z}_i = [\boldsymbol{q}_i^\intercal, \boldsymbol{a}_i^\intercal]^\intercal$, $\boldsymbol{p}_i = p(\boldsymbol{z}_i)$, and (re)define

$$G(y|x, a, p) = \mathbb{P}(\boldsymbol{y}_i \leq y | \boldsymbol{x}_i = x, \boldsymbol{a}_i = a, \boldsymbol{p}_i = p). \tag{12}$$

We start by estimating

$$\psi^* = \psi^*(x^*, \tau^*|a^*, v^*) = \mathbb{Q}_{g(x^*, a^*, \boldsymbol{u}_i)|\boldsymbol{v}_i}(\tau^*|v^*) = g\big\{x^*, a^*, \mathbb{Q}_{\boldsymbol{u}_i|\boldsymbol{v}_i}(\tau^*|v^*)\big\}. \tag{13}$$

We propose estimating ψ^* by inverting the functions $H(\cdot|a^*, v^*)$ defined by

$$H(y|a, v) = \mathbb{P}\{g(0, a, \boldsymbol{u}_i) \leq y | \boldsymbol{v}_i = v\},$$

which under assumption L satisfies

$$H(y|a,v) = G(y|0,a,v) + v\partial_p G(y|0,a,v).$$

So whereas the estimator in CL09 is semiparametric and the object of interest is the mean,[8] our approach is nonparametric and we estimate quantiles which entails an additional inversion step which requires some empirical process theory. However, in section 4 we discuss the possibility of using single index restrictions allowing for the possibility of semiparametric estimation, albeit in a more structural fashion than CL09.

Let $G(y|a,v) = G(y|0,a,v)$ and by assumptions B to D for $\boldsymbol{w}_i = w(\boldsymbol{z}_i)$ for some function w to be introduced later,

$$G(y|a^*,v^*) = \frac{\mathbb{E}\{\mathbb{1}(\boldsymbol{y}_i \leq y)\boldsymbol{w}_i|\boldsymbol{x}_i = 0, \boldsymbol{a}_i = a^*, \boldsymbol{p}_i = v^*\}}{\mathbb{E}(\boldsymbol{w}_i|\boldsymbol{x}_i = 0, \boldsymbol{a}_i = a^*, \boldsymbol{p}_i = v^*)}. \tag{14}$$

Since the function p is estimable, so is the function G, and thence H. We propose estimating both G and $\partial_v G$ by nonparametric kernel (derivative) regression estimation and inverting the resulting estimator of $H(\cdot|a^*,v^*)$ to estimate ψ^*.

It is well–known that kernel regression estimation has problems in the tails of the distribution, or more precisely wherever the density of conditioning variables is close to zero. In the estimation we hence only use observations i for which $\boldsymbol{z}_i$ belongs to some convex and compact set $\mathcal{Z}$ on which the density f of $\boldsymbol{z}_i$ is bounded away from zero and which is further constrained below. Not using all data does have efficiency implications, but the commonly used alternative of sample–size dependent trimming is practically cumbersome, technically messy, and any meaningful gains of such a procedure in empirical work are phantasmic. In what follows we will assume $\boldsymbol{z}_i$ to be continuously distributed even though in empirical work discrete covariates and instruments are prevalent. Kernel estimation with discrete regressors can be accommodated (see e.g. Delgado and Mora, 1995) at the expense of longer proofs. However, because in practice the index version of the estimator proposed in section 4 will often be more attractive and since for the index version only one of the elements of $\boldsymbol{a}_i$ and one of the elements in $\boldsymbol{z}_i$ that are not in $\boldsymbol{a}_i$ must be continuously distributed, we do not weaken the assumption here.

[8] or a quantile under an additional additive separability assumption.

The function w in (14) is chosen to be nonnegative on $\mathcal{Z}$ and zero elsewhere. Let $\boldsymbol{I}_{xi} = \mathbb{1}(\boldsymbol{x}_i = 0)\boldsymbol{w}_i$, $\boldsymbol{I}_i(y) = \boldsymbol{I}_{xi}\mathbb{1}(\boldsymbol{y}_i \le y)$, let f_{ap} be the joint density of $\boldsymbol{a}_i$, $\boldsymbol{p}_i$, and let

$$S_{0x} = S_{0x}(a^*, v^*) = \mathbb{E}(\boldsymbol{I}_{xi}|\boldsymbol{a}_i = a^*, \boldsymbol{p}_i = v^*) f_{ap}(a^*, v^*), \qquad S_{1x} = \partial_v S_{0x},$$
$$S_0(y) = S_0(y; a^*, v^*) = \mathbb{E}\{\boldsymbol{I}_i(y)|\boldsymbol{a}_i = a^*, \boldsymbol{p}_i = v^*\} f_{ap}(a^*, v^*), \qquad S_1(y) = \partial_v S_0(y).$$

Then, noting that $G(y|a^*, v^*) = S_0(y)/S_{0x}$, it follows that

$$H(y|a^*, v^*) = \frac{S_0(y)S_{0x} + v^* S_1(y)S_{0x} - v^* S_0(y)S_{1x}}{S_{0x}^2}.$$

We now develop our estimator. Let k be a kernel, K be a product kernel based on k whose dimension is determined by its argument, and let h_0, h_1, h_z be bandwidths. Define $\boldsymbol{K}_{zi}(z) = K\{(z - \boldsymbol{z}_i)/h_z\}/h_z^{d_z}$, and $\hat{\boldsymbol{p}}_i = \hat{\boldsymbol{p}}(\boldsymbol{z}_i)$, where

$$\hat{\boldsymbol{p}}(z) = \frac{\sum_{i=1}^n \boldsymbol{K}_{zi}(z)\mathbb{1}(\boldsymbol{x}_i = 0)}{\sum_{i=1}^n \boldsymbol{K}_{zi}(z)}. \tag{15}$$

Let further $\boldsymbol{K}_{aij} = K\{(a^* - \boldsymbol{a}_i)/h_j\}/h_j^{d_a}$, $\boldsymbol{k}_{ij}^{(s)} = k^{(s)}\{(v^* - \boldsymbol{p}_i)/h_j\}/h_j^{s+1}$, $\hat{\boldsymbol{k}}_{ij}^{(s)} = k^{(s)}\{(v^* - \hat{\boldsymbol{p}}_i)/h_j\}/h_j^{s+1}$, and

$$\hat{\boldsymbol{S}}_s(y; p) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{k}_{is}^{(s)} \boldsymbol{K}_{ais} \boldsymbol{I}_i(y), \qquad \hat{\boldsymbol{S}}_s(y; \hat{p}) = \frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{k}}_{is}^{(s)} \boldsymbol{K}_{ais} \boldsymbol{I}_i(y), \tag{16}$$

The proposed estimator is then given by

$$\hat{\hat{\boldsymbol{H}}}(y|a^*, v^*) = \frac{\hat{\hat{\boldsymbol{S}}}_0(y)\hat{\hat{\boldsymbol{S}}}_{0x} + v^* \hat{\hat{\boldsymbol{S}}}_1(y)\hat{\hat{\boldsymbol{S}}}_{0x} - v^* \hat{\hat{\boldsymbol{S}}}_0(y)\hat{\hat{\boldsymbol{S}}}_{1x}}{\hat{\hat{\boldsymbol{S}}}_{0x}^2}, \tag{17}$$

where $\hat{\hat{\boldsymbol{S}}}_0(y) = \hat{\boldsymbol{S}}_0(y; \hat{p})$, $\hat{\hat{\boldsymbol{S}}}_{0x} = \hat{\hat{\boldsymbol{S}}}_0(\infty)$, $\hat{\hat{\boldsymbol{S}}}_1(y) = \hat{\boldsymbol{S}}_1(y; \hat{p})$, and $\hat{\hat{\boldsymbol{S}}}_{1x} = \hat{\hat{\boldsymbol{S}}}_1(\infty)$.

The bandwidths h_0, h_1, and h_z vary with n according to

$$h_0 \sim n^{-\eta_0}, h_1 \sim n^{-\eta_1}, \text{ and } h_z \sim n^{-\eta_z} \tag{18}$$

for some $\eta_0, \eta_1, \eta_z > 0$ to be constrained in assumption S.

So there are a total of five different input parameters here: a kernel, the w–function, and three bandwidths. The number of bandwidths can be reduced to two by choosing $h_0 = h_z$, but our conditions require that h_0 and h_z converge to zero faster than h_1.

We make the following assumptions.

Assumption N. *$G(\cdot|a^*, v^*)$ and $\partial_p G(\cdot|a^*, v^*)$ are differentiable in y, and hence so is $H(\cdot|a^*, v^*)$.*

Assumption N presumes a continuous outcome, which is not necessary for identification but is convenient for estimation. It is sufficient for the quantile of interest to be uniquely defined and is needed for the empirical process results that are used.

Assumption O. *$\mathcal{Z} = \mathcal{Z}_1 \times \tilde{\mathcal{Z}}$ is a subset of the interior of the support $\mathcal{S}_z$ of $\boldsymbol{z}_i = [z_{i1}, \tilde{\boldsymbol{z}}_i^\intercal]^\intercal$, for which $\tilde{\mathcal{Z}} \subset \mathbb{R}^{d_z-1}$, $\mathcal{Z}_1 = [\underline{z}_1, \bar{z}_1]$ for some $\underline{z}_1 < \bar{z}_1$ are compact and convex. On $\mathcal{Z}$ the density $f \in \mathcal{F}_2$ of $\boldsymbol{z}_i$ is bounded away from zero. Finally, $\mathcal{Z}$ contains points of the form (q, a^*) and for any such points and any vector $\xi \in \mathbb{R}^{d_a}$ there exists an $\epsilon > 0$ such that $(q, a^* - \epsilon\xi)$ and $(q, a^* + \epsilon\xi)$ are also in $\mathcal{Z}$.*

Assumption P. *$p \in \mathcal{F}_2$ is (strictly) increasing in its first argument and $0 < \mathbb{P}\{p(\bar{z}_1, \tilde{\boldsymbol{z}}_i) \leq v^*\} < \mathbb{P}\{p(\underline{z}_1, \tilde{\boldsymbol{z}}_i) \leq v^*\} < 1$.*

Assumptions M, O and P are typical for the kernel derivative estimation literature, albeit that we require the existence of one extra derivative in the first argument. There is nothing special about the first argument (other than that it is an element of $\boldsymbol{q}_i$ rather than $\boldsymbol{a}_i$); one of the instruments used must satisfy this condition, but there is no need to know, or indeed specify, which one. The number of required derivatives with respect to z_1 can possibly be reduced by one at the expense of much more restrictive conditions on the bandwidth sequences (assumption S) and permitted dimensions d_a, d_z.

The remaining assumptions (assumptions Q to S below) pertain to the choice of input parameters and are hence of lesser importance as long as input parameters that satisfy the properties exist.

Assumption Q. *$w \in \mathcal{F}_2$ is positive on the interior of $\mathcal{Z}$, zero everywhere else, and nowhere greater than one.*

Assumption Q is there both to ensure that only observations i with $\boldsymbol{z}_i \in \mathcal{Z}$ are used (the need for which was explained earlier) and to allow us to use standard kernel bias expansions by removing discontinuities on the boundaries of $\mathcal{Z}$.

We now state our conditions for the kernel and bandwidth choices.

Assumption R. *The kernel k is even, everywhere nonnegative, infinitely many times boundedly differentiable, and integrates to one. It further satisfies $\kappa_{s2} = \int \{k^{(s)}(t)\}^2 dt < \infty$ for $s = 0, 1$, and $\kappa_2 = \int k(t)t^2 dt < \infty$.*

Conditions on the kernel similar to those in assumption R are standard in the kernel estimation literature. Since we get to choose k, assumption R is innocuous. It is possible to require a smaller number of derivatives at the expense of longer proofs and possibly stronger restrictions on the bandwidths than those found in assumption S.

Assumption S. *The constants η_0, η_1, η_z defined in (18) are such that for $\eta^* = \max(2d_z\eta_z - 1, 0)$, $2\eta_0 < \min(4\eta_z, 1 - \eta_z d_z)$, and*

$$\max\{\eta^*, 1 - 4\max(\eta_z, \eta_0), (d_a + 1)\eta_0, 2(d_a + 2)\eta_0 + \eta^* - 1\} < (d_a + 3)\eta_1 < \min\{(d_a + 3)\min(\eta_0, \eta_z), 1 - \eta^*\}.$$

The choice of bandwidths in assumption S results in the convergence rate

$$\rho_n = n^{\{1-(3+d_a)\eta_1\}/2}. \tag{19}$$

While assumption S allows for undersmoothing, the choice of $\eta_1 = 1/(7 + d_a)$ leads to the optimal rate of $\rho_n = n^{2/(7+d_a)}$ for kernel derivative estimators (using second order kernels). Faster convergence rates are feasible under additional smoothness conditions (more derivatives) using bias reduction techniques such as higher order kernels or local polynomial estimation. Such an extension is a well–trodden path, which adds no new theoretical insights, and its promised performance improvements are not often realized in samples of finite size.

To see that η_0, η_1, η_z exist for many (but certainly not all) combinations of d_a, d_z, we present table 1 which for $\eta_1 = 1/(7 + d_a)$ lists the values of 1,000 times the values of η_0, η_z which are the 'points of gravity' of the regions of η_0, η_z combinations for which assumption S is satisfied and which are in some sense hence farthest from violating assumption S. If there is no entry in the table for a particular $d_a \leq d_z - 1$ combination then that means that for $\eta_1 = 1/(7+d_a)$ there are no values of η_0, η_z to satisfy assumption S. Of course, η_0, η_1, η_z only indicate a rate; the constant multiplying $n^{-\eta_0}$ for instance still needs to be chosen.

3.2. **Limit results for our estimator of H.** Before stating our formal results, we introduce some notation. Let $\lambda_b = \lim_{n\to\infty}(\rho_n h_1^2)$, $\lambda_v = \lim_{n\to\infty}(\rho_n^2/nh_1^{3+d_a})$, $p_y(y|z) = \mathbb{P}(\boldsymbol{y}_i \leq y|\boldsymbol{x}_i = 0, \boldsymbol{z}_i = z)$,

$$\Gamma(y, a^*, v^*) = v^* f_{ap}(a^*, v^*)\kappa_{12}\mathbb{E}\{p_y(y|\boldsymbol{z}_i)\boldsymbol{w}_i^2|\boldsymbol{a}_i = a^*, \boldsymbol{p}_i = v^*\},$$

	d_z								
$d_a \downarrow$	1	2	3	4	5	6	7	8	9
	226	209	190	166					
0	142	142	142	142					
	345	216	174	154					
	181	180	176	166	145				
1	125	125	125	125	125				
	380	221	169	145	133				
		145	144	142	136	128			
2		111	111	111	111	111			
		223	168	140	125	115			
			122	121	120	115			
3			100	100	100	100			
			160	132	116	107			
				107	106	104	100		
4				90	90	90	90		
				126	110	100	93		
					96	95	92		
5					83	83	83		
					105	95	88		
						86	85	83	
6						76	76	76	
						90	83	78	
							79	77	
7							71	71	
							80	75	
								72	70
8								66	66
								72	68

TABLE 1. Suggested choices for 1,000 times η_0, η_1, η_z for various combinations of d_a, d_z.

and

$$\mathcal{C}(y, y^*) = \mathcal{C}(y, y^*; a^*, v^*) = \Gamma\{\min(y, y^*), a^*, v^*\} - \Gamma(y, a^*, v^*)G(y^*|a^*, v^*) \\ - \Gamma(y^*, a^*, v^*)G(y|a^*, v^*) + G(y|a^*, v^*)G(y^*|a^*, v^*)\Gamma(\infty, a^*, v^*). \quad (20)$$

Let further

$$\mathcal{C}(y, y^*) = \mathcal{C}(y, y^*; a^*, v^*) = \lambda_v \frac{v^{*2}\mathcal{C}(y, y^*; a^*, v^*)}{S_{0x}^2(a^*, v^*)},$$

and

$$\mathcal{B}(y) = \mathcal{B}(y; a^*, v^*) = \frac{\lambda_b v^* \kappa_2}{2 S_{0x}} \mathrm{tr}\Big\{ \partial_v \partial_{bb^\intercal} S_0(y; a^*, v^*) - G(y|a^*, v^*) \partial_v \partial_{bb^\intercal} S_{0x}(a^*, v^*) \Big\}.$$

Theorem 4. *Under assumptions L to S,*

$$\rho_n \{ \hat{\hat{\boldsymbol{H}}}(\cdot|a^*, v^*) - H(\cdot|a^*, v^*) \} \xrightarrow{w} \mathbb{G},$$

on the space of bounded functions on $\mathcal{Y} = \{y : \exists u \in \mathcal{U} : g(0, a^*, u) = y\}$, $\mathcal{L}^\infty(\mathcal{Y})$, *where* $\mathbb{G}$ *is a Gaussian process with mean* $\mathcal{B}$ *and covariance kernel* $\mathcal{C}$.

Please note that table 1 implies that it is possible for the limit distribution not to be affected by the first step estimation of p — the 'oracle property' — even in some cases in which $d_z > d_a + 1$. This may appear to be at odds with other results in the voluminous literature on nonparametric generated regressors (Rilstone, 1996; Pinkse, 2001; Mammen, Rothe, and Schienle, 2012, inter multa alia) in which nonparametrically estimated regressors do affect the optimal convergence rate unless the estimated regressors are functions whose vector of arguments is of smaller dimension than the vector of arguments of the function of interest. However, here we are not evaluating $\hat{\boldsymbol{p}}$ at a fixed point, say z^*, to obtain our estimate $\hat{\hat{\boldsymbol{H}}}(y|a^*, v^*)$. Instead we *only* use $\hat{\boldsymbol{p}}_i$'s which are averaged in some sense which reduces their contribution to the variance, thereby allowing us to use smaller h_z values to reduce the bias, also.

3.3. **Limit results for our estimator of ψ^*.** We finally turn to our estimator of ψ^* itself. We use the standard definition of quantile using the estimated conditional distribution function, i.e.

$$\hat{\boldsymbol{\psi}}^* = \inf\{\tilde{\psi} : \hat{\hat{\boldsymbol{H}}}(\tilde{\psi}|a^*, v^*) \geq \tau^*\}.$$

The asymptotic behavior can then be inferred from theorem 4. Indeed, we have theorem 5.

Theorem 5. *Under assumptions L to S,* $\rho_n(\hat{\boldsymbol{\psi}}^* - \psi^*) \xrightarrow{d} N(\mathcal{B}_\psi, \mathcal{V}_\psi)$, *where*

$$\mathcal{B}_\psi = -\frac{\mathcal{B}(\psi^*)}{H'(\psi^*|a^*, v^*)}, \qquad \mathcal{V}_\psi = \frac{\mathcal{C}(\psi^*, \psi^*)}{\{H'(\psi^*|a^*, v^*)\}^2}.$$

3.4. **Bias and variance estimation.** The bias and variance in theorem 5 can be consistently estimated by standard methods. Since the bias can be removed by undersmoothing, the Jackknife, or

other methods, we focus on estimation of the variance below. Note that

$$H'(y|a^*, v^*) = \frac{S_0^{(1)}(y)S_{0x} + v^* S_1^{(1)}(y)S_{0x} - v^* S_0^{(1)}(y)S_{1x}}{S_{0x}^2},$$

where letting $f_y(\cdot|x, z)$ be the conditional density of $\boldsymbol{y}_i$ given $\boldsymbol{x}_i = x$ and $\boldsymbol{z}_i = z$,

$$\begin{cases} S_0^{(1)}(y) = \partial_y S_0(y) = \mathbb{E}\{\boldsymbol{I}_{xi} f_y(y|\boldsymbol{x}_i, \boldsymbol{z}_i)|\boldsymbol{p}_i = v^*\} f_{ap}(a^*, v^*), \\ S_1^{(1)}(y) = \partial_y S_1(y) = \partial_{yv} S_0(y). \end{cases}$$

For $s = 0, 1$, we can estimate $S_s^{(1)}(y)$ by

$$\hat{\hat{\boldsymbol{S}}}_s^{(1)}(y) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{K}_{ais}\hat{\boldsymbol{k}}_{is}^{(s)}\boldsymbol{k}_{yi}(y)\boldsymbol{I}_{xi},$$

where $\boldsymbol{k}_{yi}(y) = k\{(y - \boldsymbol{y}_i)/h_y\}/h_y$ with h_y a bandwidth. Also, $\mathcal{C}(y, y; a^*, v^*)$ can be estimated by

$$\hat{\boldsymbol{\mathcal{C}}}(y, y; a^*, v^*) = \frac{h_1^{1+d_a}}{n}\sum_{i=1}^n \boldsymbol{K}_{ai1}^2\hat{\boldsymbol{k}}_{i1}^2\mathbb{1}(\boldsymbol{x}_i = 0)\boldsymbol{w}_i^2\left\{\mathbb{1}(\boldsymbol{y}_i \le y) - \frac{\hat{\boldsymbol{S}}_0(y)}{\hat{\boldsymbol{S}}_{0x}}\right\}^2.$$

The final estimator of $\mathcal{V}_\psi$ can be obtained by using $\hat{\hat{\boldsymbol{S}}}_s^{(1)}$ and $\hat{\boldsymbol{\mathcal{C}}}$ evaluated at $y = y^* = \hat{\boldsymbol{\psi}}^*$. The following theorem establishes the consistency of $\hat{\hat{\boldsymbol{S}}}_s^{(1)}(\hat{\boldsymbol{\psi}}^*)$ and $\hat{\boldsymbol{\mathcal{C}}}(\hat{\boldsymbol{\psi}}^*, \hat{\boldsymbol{\psi}}^*)$.

Theorem 6. *Suppose that assumptions L and N to S are satisfied with* $h_y = o(1)$, $1 = o(\rho_n h_y)$, *and* $\sup_s |k(s)| < \infty$. *Then, for* $s = 0, 1$,

$$\hat{\hat{\boldsymbol{S}}}_s^{(1)}(\hat{\boldsymbol{\psi}}^*) \xrightarrow{p} S_s^{(1)}(\psi^*) \quad \textit{and} \quad \hat{\boldsymbol{\mathcal{C}}}(\hat{\boldsymbol{\psi}}^*, \hat{\boldsymbol{\psi}}^*) \xrightarrow{p} \mathcal{C}(\psi^*, \psi^*, v^*).$$

4. Index

In most applications the dimensions of the $\boldsymbol{a}_i, \boldsymbol{z}_i$ vectors are too large for estimates to be sufficiently precise. One solution to this problem is to impose semiparametric restrictions on the g and p functions or, said differently, to assume that $\boldsymbol{a}_i, \boldsymbol{z}_i$ enter as indices. As a leading example, we consider[9]

$$\begin{cases} \boldsymbol{y}_i &= g\{\boldsymbol{x}_i, \boldsymbol{a}_i^\intercal\theta_0, \boldsymbol{u}_i\}, \\ \boldsymbol{x}_i &= \mathbb{1}\{\boldsymbol{v}_i > p(\boldsymbol{z}_i^\intercal\gamma_0)\}, \end{cases} \tag{21}$$

albeit that in our application we allow the value of θ_0 in $g(1, \cdot, \cdot)$ to be different from that in $g(0, \cdot, \cdot)$.

[9]Other parametric link functions or multiple indices can be accommodated but they complicate the identification conditions.

It follows from the copious work on index models that several normalizations are needed. First, $\boldsymbol{a}_i, \boldsymbol{z}_i$ should not include a constant term and even then the vectors θ_0, γ_0 are (at best) identified up to scale. Second, one should be able to move $\boldsymbol{x}_i$ exogenously without changing $\boldsymbol{a}_i$, i.e. at least one of the γ–coefficients on the $\boldsymbol{q}_i$ component of $\boldsymbol{z}_i$ should be nonzero. Indeed, if one lets $\boldsymbol{z}_i^\intercal \gamma_0 = \boldsymbol{a}_i^\intercal \gamma_{0a} + \boldsymbol{q}_i^\intercal \gamma_{0q}$ then the conditions of sections 2 and 3 can be verified conditional on $\boldsymbol{a}_i^\intercal \theta_0 = a^{*\intercal}\theta_0$ and taking $\boldsymbol{z}$ in sections 2 and 3 to equal $\boldsymbol{q}_i^\intercal \gamma_{0q}$; doing this requires that $\gamma_{0q} \neq 0$.

From now on, we take identification of ψ^* and that of γ_0, θ_0 as given. We also take as given that $\sqrt{n}$–consistent estimators $\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\theta}}$ of γ_0, θ_0 exist. We are not generally fans of high–level assumptions. However, the structure of (21) fits well into the index model estimation literature of which Powell, Stock, and Stoker (1989); Ichimura (1993); Klein and Spady (1993) are prominent examples. Indeed, $\mathbb{P}(\boldsymbol{x}_i = 1 | \boldsymbol{z}_i = z) = p(z^\intercal \gamma_0)$, which yields an estimate of γ_0. Further, $\mathbb{E}(\boldsymbol{y} | \boldsymbol{x} = x, \boldsymbol{z} = z)$ is an unknown function of $x, z^\intercal \gamma_0, a^\intercal \theta_0$, which can be used to construct an estimate of θ_0.

The main task for this section, then, is to establish that the estimation of the nuisance parameters γ_0, θ_0 does not affect the limit distribution of the estimator of ψ^*.

Let $\hat{\hat{\boldsymbol{\psi}}}^*$ be defined as $\hat{\boldsymbol{\psi}}^*$, replacing $\boldsymbol{a}_i$ with $\boldsymbol{a}_i^\intercal \hat{\boldsymbol{\gamma}}$ and $\boldsymbol{z}_i$ with $\boldsymbol{z}_i^\intercal \hat{\boldsymbol{\theta}}$.

Theorem 7. *Suppose that $\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\theta}}$ are $\sqrt{n}$–consistent estimates of γ_0, θ_0, respectively. Then theorems 5 and 6 hold for $\hat{\hat{\boldsymbol{\psi}}}^*$ when $\boldsymbol{a}_i, \boldsymbol{z}_i$ are replaced with $\boldsymbol{a}_i^\intercal \hat{\boldsymbol{\gamma}}$ and $\boldsymbol{z}_i^\intercal \hat{\boldsymbol{\theta}}$, respectively.*

5. Application

5.1. **Data.** We now apply our method to estimate the returns to college education using the NLSY 1979 data, which were used by CL09 and CHV11 inter alia. Indeed, we use the code provided by CHV11 to obtain exactly the same data set and therefore we compare our results with CHV11: the results of CL09 are not qualitatively different from those in CHV11. The sample consists of 1,747 white males. The data set contains a number of demographic and education–related variables as can be gleaned from our results tables, tables 2 and 3. Even though the sample is fairly small relative to the nonparametric specification of the model that we are estimating, we are finding several interesting results.

A detailed description of the data can be found in CHV11 and its supplementary material,[10] which also contains summary statistics separated by education group (college versus noncollege).

[10] See https://www.aeaweb.org/aer/data/oct2011/20061111_app.pdf

5.2. **Methodology.** As noted in the introduction, our approach differs from that of CL09 and CHV11 in several respects. The most important difference is that in both those papers the error in the earnings equation enters additively whereas in our model it enters nonparametrically, i.e. nonadditively. Hence in CL09 and CHV11 there can be no interaction between the earnings equation error and regressors such that potential wage curves (as a function of v^*) for different covariate values are constrained to be vertical shifts of each other, whereas ours can vary freely. This shape restriction is also used in CHV11, which takes a more parametric (and therefore more restrictive) approach than CL09.

Like CL09 and CHV11 we distinguish between two groups: those with a college education (group 1) and those without (group 0). For the schooling equation we use the second half of (21). For the wage equation we allow the θ–coefficients in (21) to be different across education groups, i.e. we consider

$$\begin{cases} \boldsymbol{y}_i = g_1(\boldsymbol{a}_i^\intercal \theta_{10}, \boldsymbol{u}_{i1}) \mathbb{1}\{\boldsymbol{x}_i = 1\} + g_0(\boldsymbol{a}_i^\intercal \theta_{00}, \boldsymbol{u}_{i0}) \mathbb{1}\{\boldsymbol{x}_i = 0\}, \\ \boldsymbol{x}_i = \mathbb{1}\{\boldsymbol{v}_i > p(\boldsymbol{z}_i^\intercal \gamma_0)\}, \end{cases}$$

where the outcome variable is a multi–year (1989–1993) average of log hourly wages deflated to 1983$: our identification analysis focuses on each of the two potential outcomes and it naturally extends to the case where the errors in each of the potential outcome equations are different.

Since $\boldsymbol{x}_i$ is *increasing* in $\boldsymbol{v}_i$ for fixed $\boldsymbol{z}_i$ and $\boldsymbol{v}_i$ is assumed independent of $\boldsymbol{z}_i$, we interpret $\boldsymbol{v}_i$ as an unobserved measure of an individual's inclination to attend college, i.e. a measure of such inclination that is not accounted for by $\boldsymbol{z}_i$.

We estimate γ_0 (normalized to have a unit length) in the schooling equation using Ichimura's semiparametric least squares estimator (SLSE), which uses the fact that

$$\gamma_0 = \underset{\|\gamma\|=1}{\operatorname{argmin}} \, \mathbb{E}\big\{\boldsymbol{x}_i - \mathbb{E}(\boldsymbol{x}_i | \boldsymbol{z}_i^\intercal \gamma)\big\}^2.$$

Ichimura (1993) states conditions under which the SLSE is $\sqrt{n}$–consistent.

The specification of the θ–parameters is in essence a double index model (Ichimura and Lee, 1991) since

$$\begin{cases} \mathbb{E}(\boldsymbol{y}_i | \boldsymbol{x}_i = 0, \boldsymbol{a}_i = a, \boldsymbol{z}_i = z) = \mathbb{E}\{g_0(0, a^\intercal \theta_{00}, \boldsymbol{u}_{i0}) | \boldsymbol{v}_i \leq p(z^\intercal \gamma_0)\} = G_0(a^\intercal \theta_{00}, z^\intercal \gamma_0), \\ \mathbb{E}(\boldsymbol{y}_i | \boldsymbol{x}_i = 1, \boldsymbol{a}_i = a, \boldsymbol{z}_i = z) = \mathbb{E}\{g_1(1, a^\intercal \theta_{10}, \boldsymbol{u}_{i1}) | \boldsymbol{v}_i > p(z^\intercal \gamma_0)\} = G_1(a^\intercal \theta_{10}, z^\intercal \gamma_0), \end{cases}$$

albeit that in our case we already have $\sqrt{n}$–consistent estimates of γ_0. We therefore use Escanciano, Jacho-Chávez, and Lewbel (2010, EJL10) instead of Ichimura and Lee (1991); $\sqrt{n}$–consistency obtains under conditions stated in EJL10.

Since our analysis is semiparametric, we scale the exogenous variables (i.e. controls and instruments) by their standard deviations. All computations are done in Matlab using its global optimization toolbox with 100 initial values. For the bandwidth choice we follow Härdle, Hall, and Ichimura (1993), which entails optimizing over the coefficients for each bandwidth and then doing a grid search over the bandwidth. The search range is $(0.01, 0.41)$ noting that $2n^{-1/5} \approx 0.4$. There is, as is not unusual, some sensitivity to the choice of bandwidth, but changing the bandwidth does not affect the qualitative conclusions of our study.

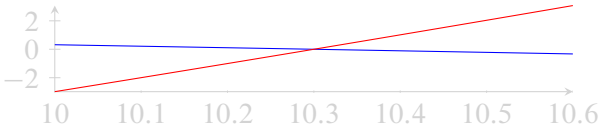


FIGURE 1. Effect of permanent local log earnings at age 17

5.3. **Results pertaining to the schooling decision.** We now turn to a discussion of our estimation results. Table 2 shows the estimates of the schooling equation coefficients. We consider two specifications: one includes proxies of intellectual ability (AFQT[11] and mother's education) and the other does not. The full model is consistent with specifications used by others and is our main specification. We include the models without proxies mainly to demonstrate that *(i)* v^* should be interpreted as an inclination to attend college conditional on ability and that *(ii)* the AFQT scores and mother's education do a great job proxying for ability.

The estimates can be interpreted as average partial effects up to scale normalization, because the schooling equation is a single index model. The coefficients in the model without proxies are normalized to have the same norm for the corresponding visible coefficients in the full model. The last column of table 2 contains the corresponding estimates of the first stage logit regression reported in the online appendix of CHV11 normalized to have the same norm as the visible coefficients of the full model after correcting for our scale normalization of the covariates mentioned earlier. The most striking difference is in the effect of permanent local log earnings at 17 (PLLE17): figure 1 displays the index as a function of PLLE17 holding all other regressors constant, where the domain

[11] The Armed Forces Qualification Test was administered to all subjects.

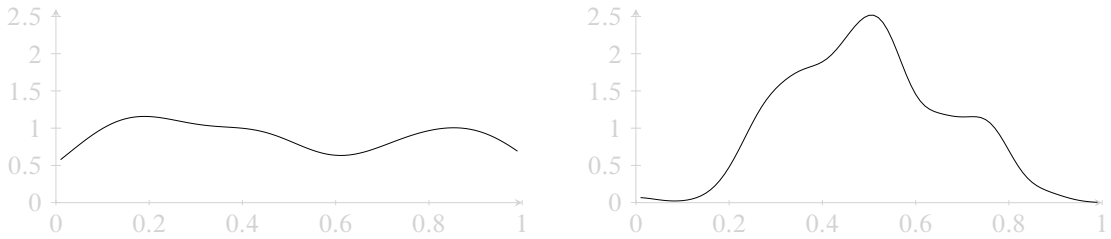
	Full model		No proxies		
	$\hat{\gamma}$	S.E.	$\hat{\gamma}$	S.E.	CHV11
Corrected AFQT	0.4630	(0.0421)**			0.2237
Corrected AFQT Squared	0.0411	(0.0634)			0.0118
Mother's Years of Schooling	−0.1852	(0.0532)**			-0.1897
Mother's Years of Schooling Squared	0.1883	(0.0680)**			0.0258
Number of Siblings	−0.4398	(0.0586)**	−0.4376	(0.0355)**	-0.3833
Number of Siblings Squared	0.0010	(0.0401)	−0.0235	(0.0302)	0.0016
Urban Residence at 14	−0.0130	(0.0330)	0.1243	(0.0596)*	0.0023
Permanent Local Log Earnings at 17	0.0559	(0.0239)*	0.0524	(0.0284)	-0.5112
Permanent Local Log Earnings at 17 Sq.	−0.0551	(0.0234)*	−0.0014	(0.0325)	0.5146
Permanent State Unempl. Rate at 17	0.0282	(0.0329)	−0.0415	(0.0323)	-0.0145
Permanent State Unempl. Rate at 17 Sq.	−0.0415	(0.0319)	0.0266	(0.0310)	-0.0139
Presence of a College at 14	0.0135	(0.0352)	0.0500	(0.0475)	0.0174
× AFQT	0.0046	(0.0495)			0.0059
× Mother's Education	0.0087	(0.0383)			-0.0180
× Number of Siblings	−0.0249	(0.0430)	−0.0212	(0.0396)	0.0019
Local Log Earnings at 17	−0.0618	(0.0431)	−0.0686	(0.0488)	-0.0305
× AFQT	−0.1882	(0.1100)			-0.1981
× Mother's Education	0.1665	(0.0293)**			0.1953
× Number of Siblings	0.5344	(0.0756)**	0.3375	(0.0779)**	0.3843
Local Unemployment Rate at 17	0.1097	(0.0428)*	−0.0214	(0.0372)	0.0377
× AFQT	−0.0099	(0.0499)			-0.0052
× Mother's Education	−0.0553	(0.0380)			-0.0398
× Number of Siblings	0.0292	(0.0447)	0.0072	(0.0355)	-0.0046
Tuition in 4 Year Public Colleges at 17	−0.1385	(0.0514)**	−0.0076	(0.0389)	-0.0279
× AFQT	0.1783	(0.0720)*			0.0081
× Mother's Education	0.1667	(0.0355)**			0.0297
× Number of Siblings	−0.2352	(0.0498)**	−0.1113	(0.0439)*	-0.0062
Bandwidth	0.03		0.04		

Notes: The birth year cohort dummy coefficients are omitted from the table. The numbers in parentheses are standard errors that are computed using the bootstrap with 500 replications. The coefficients in the full model are normalized to have norm one with the birth year cohort dummies included. The no proxies coefficients are normalized to have the same norm as the full model coefficients for the visible coefficients, e.g. the norm of the coefficients in the *no proxies* column is the same as that of the corresponding coefficients in the full model column. The CHV11 coefficients are taken from their online appendix, rescaled to be consistent with our choice to normalize covariates to have unit variance and then renormalized to have the same norm as the visible vector of full model coefficients. Significance at the 5% and 1% level is indicated with asterisks.

TABLE 2. Schooling equation coefficients.

is mean±two standard deviations and where the functions are vertically centered at zero. Both are close to linear with one increasing and the other decreasing, albeit that the magnitudes of the CHV11 coefficients far exceeds ours. We have done further experiments (not tabulated) to determine the source of this difference, which suggest that the differences are mainly due to the nonparametric and

nonseparable second stage in our estimation method: if we replace our semiparametric first stage with a fully parametric first stage then the results are largely unchanged.



(a) Full model (b) No proxies

FIGURE 2. Density of $p(z_1^\intercal \gamma_0)$

Figure 2 depicts the density estimates of $p(z_i^\intercal \gamma_0)$. When the intelligence proxies are excluded, the density of the propensity score is small at extreme quantiles, i.e. at extreme levels of the objects' inclination to attend college. This is intuitive since there are no purely demographic variables that would explain the college education decision with (near) certainty; individuals with very low AFQT scores will not attend college with certainty but once AFQT scores and mother's education are omitted there are no covariates left to predict the outcome with an equal degree of certainty.

5.4. **Results pertaining to returns to education.** Regression coefficients for the second stage can be found in table 3. The last two columns contain the coefficients of the normal switching regressions reported in the online appendix of CHV11, rescaled to make them comparable to ours. The coefficients in our full model, which as noted is our main specification, seem sensible. A few coefficients differ in sign from what is found in the CHV11 regressions, especially for group 0. The most noticeable difference is in the effect of PLLE17, which is the same variable that had different coefficients in the schooling decision equation. But most coefficients, including those on PLLE17, are insignificant in CHV11, it is unclear what one would expect the signs of the coefficients to be, and our model nests CHV11 so the differences may be due to misspecification or indeed noise due to overspecification.

We now turn our attention to the potential wage curves depicted in figure 3.[12] Like in the existing literature potential wages for group 0 in the full model decrease as v^* increases. There are however some important differences: our potential wage curves for the two education groups do not cross —

[12]Note that averaging the covariates or v^* do not yield 'unconditional' quantiles of potential wages.

	Full model		No proxies		CHV11	
	no coll.	college	no coll.	college	no coll.	college
Experience	0.7998	0.6668	0.0131	0.6511	0.0575	0.5384
Experience Squared	−0.2504	−0.4453	−0.8404	−0.4946	0.0073	-0.4068
Log Local Earnings in 1991	0.2685	0.3839	−0.4730	0.4316	0.0283	0.2605
Local Unemployment in 1991	−0.1372	0.0748	0.0273	0.0766	-0.0014	0.0077
Corrected AFQT	−0.0321	0.1562			0.0143	0.2202
Corrected AFQT Squared	0.0169	0.1060			-0.0171	0.1202
Mother's Years of Schooling	−0.0433	−0.0990			-0.0129	-0.0429
Mother's Years of Schooling Sq.	0.0191	0.0336			0.0152	0.1508
Number of Siblings	0.0577	0.0024	−0.0648	−0.0131	0.0024	-0.0369
Number of Siblings Squared	0.0421	0.0306	0.0423	0.0121	-0.0036	0.0066
Urban Residence at 14	−0.0778	0.0595	0.0459	0.1747	0.0010	0.0452
Perm. Local Log Earnings at 17	−0.1781	−0.2023	−0.1534	−0.0462	0.7027	0.3201
Perm. Local Log Earnings at 17 Sq.	0.2842	0.2558	−0.0143	0.1173	-0.6936	-0.3144
Perm. State Unempl. Rate at 17	0.2230	0.1344	0.0744	0.0403	0.0210	0.2724
Perm. State Unempl. Rate at 17 Sq.	−0.1034	−0.1531	−0.0955	-0.1980	-0.0222	-0.3241
Bandwidth	0.21	0.39	0.40	0.29		

Notes: Birth year cohort dummy coefficients are not reported. The coefficients in the last two columns were scaled to have the same norm as the vector containing the coefficients to the same variables in the first two columns. The dependent variable is log hourly wages.

TABLE 3. Earnings equation coefficients

returns on a college education are consistently positive — and potential wages for group 1 decrease as v^* increases. There are several explanations for the decreasing potential wage curves.

One explanation is that the intelligence proxies adequately measure ability and that v^* should now be interpreted as one's inclination to attend college. Those with an unfulfilled desire to attend college may do poorly in the labor market, as do those who attend college despite their lack of ability (as measured by AFQT scores and mother's education level). Such explanations are consistent with the graphs for the model without the intelligence proxies in figure 3 which feature (mostly) increasing potential wage curves because absent proxies v^* is correlated with overall ability. We conclude that AFQT and mother's education do a good job proxying for ability and focus on the full model from hereon.

The return on a college education at different quantiles of education equation (v^*) and earnings equation (τ^*) unobservables are presented in figure 4, where the controls are fixed at their sample means. Figure 5 depicts the same information in greater detail at three values of τ^* and figure 6 at five values of τ^* but without confidence bands.

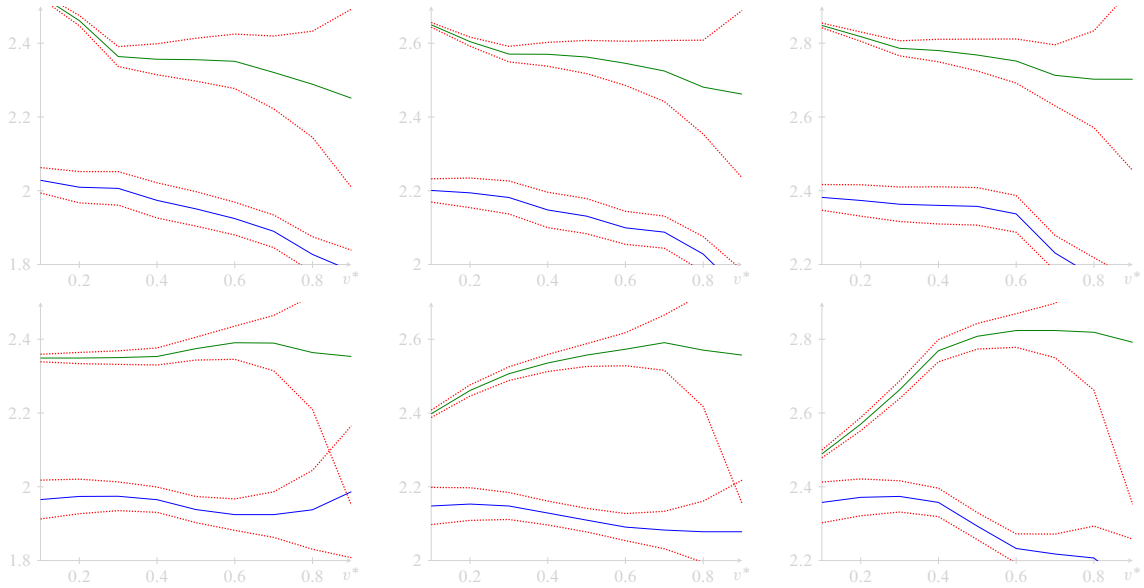


FIGURE 3. Potential wages as a function of v^* for $\tau^* = 0.3$ (left), $\tau^* = 0.5$ (middle), and $\tau^* = 0.7$ (right), with (top) and without (bottom) intelligence proxies. Covariates are evaluated at their means. Dotted lines show (pointwise) 95% confidence intervals. $2.2 \approx \$9.03$, $2.4 \approx \$11.02$, $2.6 \approx \$13.46$, $2.8 \approx \$16.44$.

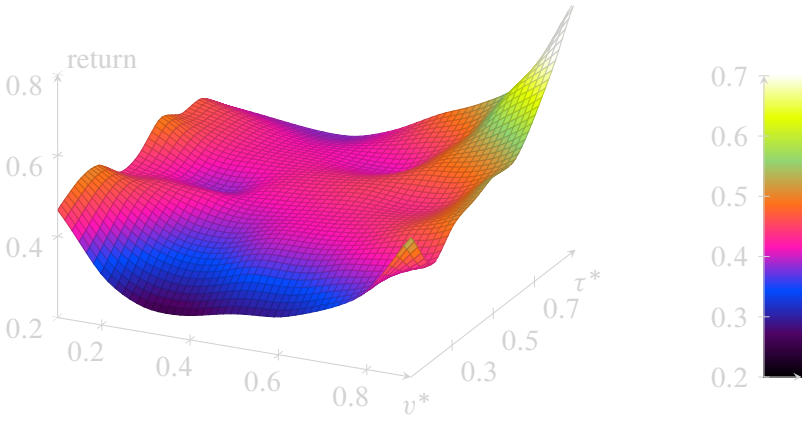


FIGURE 4. Return on a college education as a function of v^*, τ^*; covariates evaluated at their means; 0.2 corresponds to a 22% premium, $0.4 \approx 49\%$, $0.6 \approx 82\%$, and $0.7 \approx 100\%$.

Returns to college vary substantially with both τ^* and v^*, albeit that figure 5 suggests that at the moderate τ^* quantiles depicted returns are similar except at high v^* values. More striking is the $\tau^* = 0.1$ graph in figure 6, which indicates that the returns on education are substantially less at the bottom end once inclination to attend college and demographics are controlled for.

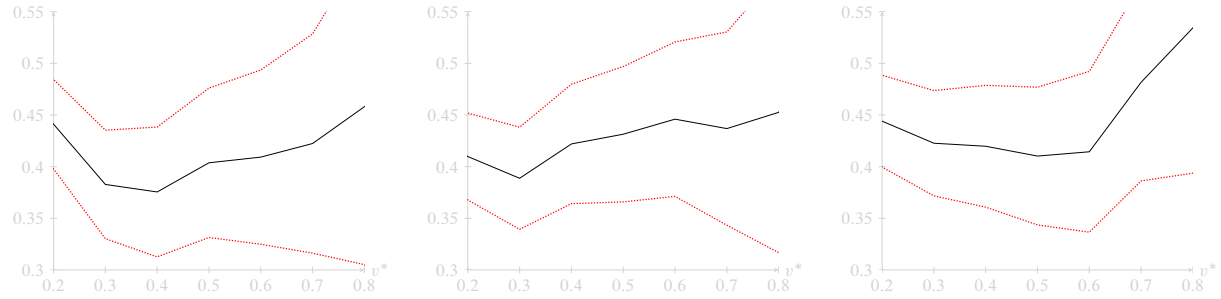


FIGURE 5. Returns to college as a function of v^* for $\tau^* = 0.3$ (left), $\tau^* = 0.5$ (middle), and $\tau^* = 0.7$ (right) for the full model. Covariates are evaluated at their means. Dotted lines show (pointwise) 95% confidence intervals

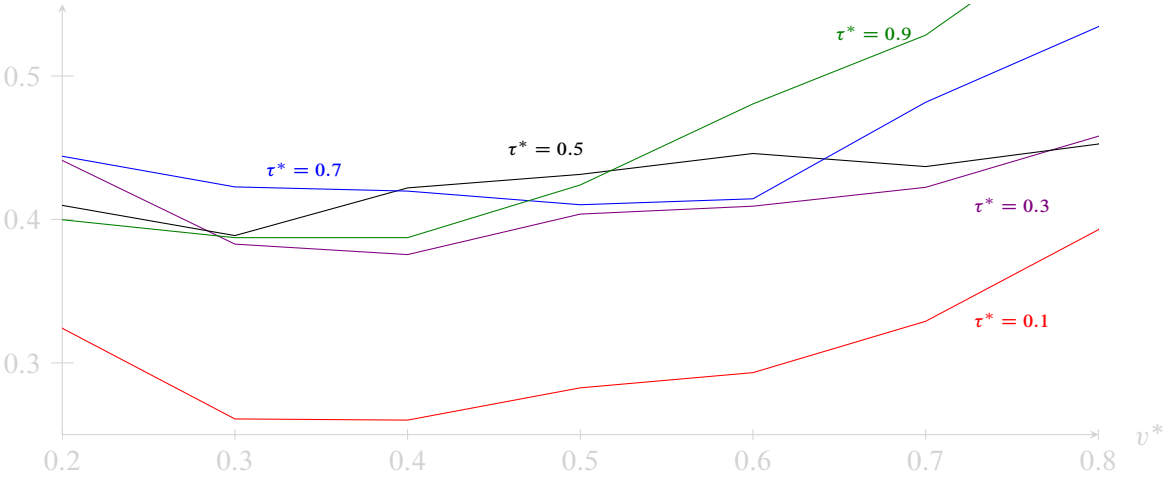


FIGURE 6. Returns for various τ^* values in the full model. Covariates are evaluated at their means.

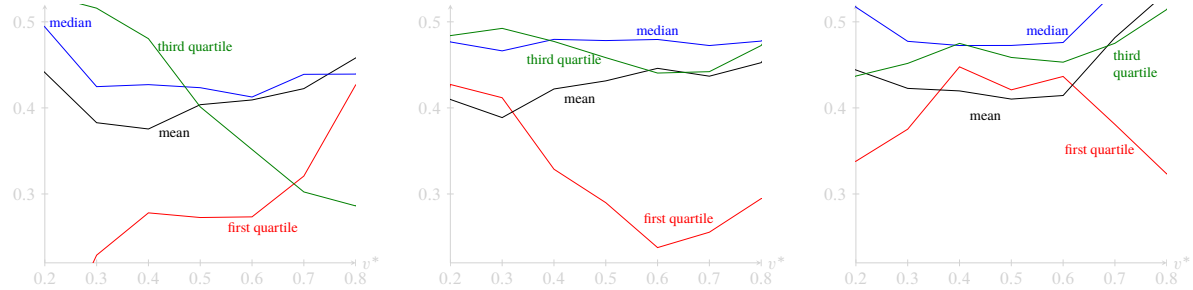


FIGURE 7. Returns for $\tau^* = 0.3$ (left), $\tau^* = 0.5$ (middle), and $\tau^* = 0.7$ (right) in the full model at different covariate values. Covariates are evaluated as indicated.

All of the graphs discussed thus far have covariates evaluated at their means. Figure 7 introduces variation across covariate values and depicts returns as a function of v^* for three moderate values of τ^* where the covariates are measured at four different values: three quartiles and the mean.

CHV11 and CL09 impose shape invariance restrictions that are not imposed by us. The shape invariance restrictions in both of those papers imply that the returns curves (and in fact also the potential wage curves) are identical to each other up to a vertical shift if one changes covariate values. Such shape invariance conditions look inconsistent with figure 7.

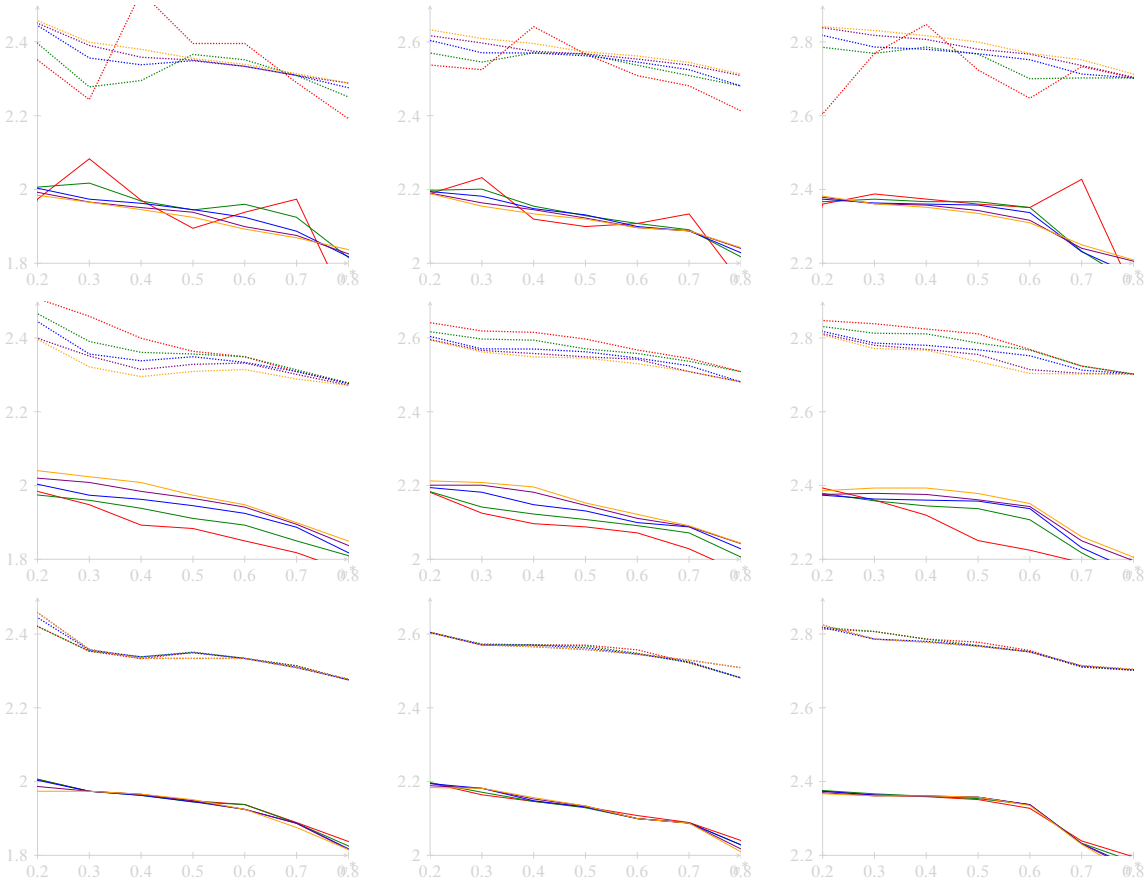


FIGURE 8. Potential wages for different values of τ^* (0.3, 0.5, 0.7 from left to right) and different bandwidth choices. At the top we vary $h_0 \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$ with $h_1 = 0.35, h_z = 0.15$, in the middle we vary $h_1 = \{0.25, 0.30, 0.35, 0.40, 0.45\}$ with $h_0 = 0.15, h_z = 0.15$, and at the bottom we vary $h_z \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$ with $h_0 = 0.15, h_1 = 0.35$. Each colored curve denotes a different bandwidth ordered from small to large (red, green, blue, purple, and orange). Dotted curves are for the college educated case.

We conclude our discussion with an analysis of the sensitivity of our results to the choice of input parameters. The results are depicted in figure 8. The results are unusually robust to the choice of h_z, which we attribute to the fact that the semiparametric estimation procedure averages across $\hat{\boldsymbol{p}}_{\boldsymbol{i}}$'s.[13]

[13]The word 'averaging' here should be interpreted in the sense of inter alia Linton and Nielsen (1995), not an immediate sample mean. We average over functions in which $\hat{\boldsymbol{p}}_{\boldsymbol{i}}$ enters as one of the arguments.

With severe undersmoothing the h_1 curves and especially the h_0 curves get the expected nonsmooth appearance which is exacerbated by the fact that we only computed estimates at 0.1 increments of the v^*–values.

References

Arcones, M., and E. Giné (1993): "Limit theorems for U–processes," *Annals of Probability*, 21(3), 1494–1542.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011): "Estimating marginal returns to education," *American Economic Review*, 101, 2754–2781.

Carneiro, P., and S. Lee (2009): "Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality," *Journal of Econometrics*, 149(2), 191–208.

Chernozhukov, V., and C. Hansen (2005): "An IV model of quantile treatment effects," *Econometrica*, 73(1), 245–261.

——— (2006): "Instrumental quantile regression inference for structural and treatment effect models," *Journal of Econometrics*, 132(2), 491–525.

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

——— (2005): "Nonparametric identification under discrete variation," *Econometrica*, 73, 1525–1550.

De la Peña, V., and E. Giné (1999): *Decoupling: from dependence to independence*. Springer Verlag.

Delgado, M., and J. Mora (1995): "Nonparametric and semiparametric estimation with discrete regressors," *Econometrica*, 63(6), 1477–1484.

Doksum, K. (1974): "Empirical probability plots and statistical inference for nonlinear models in the two sample case," *Annals of Statistics*, 2(2), 266–277.

Escanciano, J. C., D. Jacho-Chávez, and A. Lewbel (2010): "Identification and estimation of semiparametric two step models," *Unpublished manuscript*.

Härdle, W., P. Hall, and H. Ichimura (1993): "Optimal smoothing in single-index models," *Annals of Statistics*, 21(1), 157–178.

Heckman, J., and E. Vytlacil (1999): "Local instrumental variables and latent variable models for identifying and bounding treatment effects," *Proceedings of the National Academy of Sciences*, 96(8), 4730.

——— (2001): "Local instrumental variables," in *Nonlinear statistical modeling: Proceedings of the thirteenth international symposium in economic theory and econometrics: essays in honor of Takeshi Amemiya*, vol. 13, pp. 1–15. Cambridge University Press.

——— (2005): "Structural equations, treatment effects, and econometric policy evaluation," *Econometrica*, 73(3), 669–738.

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

Ichimura, H. (1993): "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models," *Journal of Econometrics*, 58(1), 71–120.

Ichimura, H., and L.-F. Lee (1991): "Semiparametric least squares estimation of multiple index models: single equation estimation," in *Nonparametric and semiparametric methods in econometrics and statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics. Cambridge*, pp. 3–49.

Imbens, G. W., and W. K. Newey (2009): "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica*, 77(5), 1481–1512.

Jun, S. (2009): "Local structural quantile effects in a model with a nonseparable control variable," *Journal of Econometrics*, 151(1), 82–97.

Jun, S. J., J. Pinkse, and H. Xu (2011): "Tighter bounds in triangular systems," *Journal of Econometrics*, 161, 122–128.

——— (2012): "Discrete endogenous variables in weakly separable models," *Econometrics Journal*, 15, 288–303.

Jun, S. J., J. Pinkse, H. Xu, and N. Yildiz (2010): "Identification of average treatment effects in a triangular system of equations," Discussion paper, Pennsylvania State University.

Kasy, M. (2011): "Identification in triangular systems using control functions," *Econometric Theory*, 27, 663–671.

Kédagni, D., and I. Mourifié (2014): "Tightening Bounds in Triangular Systems," Discussion paper, University of Toronto.

Klein, R., and R. Spady (1993): "An efficient semiparametric estimator for binary response models," *Econometrica*, 61, 387–421.

Lee, A. (1990): *U–statistics: theory and practice*. Marcel Dekker.

Linton, O., and J. P. Nielsen (1995): "A kernel method of estimating structured nonparametric regression based on marginal integration," *Biometrika*, pp. 93–100.

Ma, L., and R. Koenker (2006): "Quantile regression methods for recursive structural equation models," *Journal of Econometrics*, 134(2), 471–506.

Mammen, E., C. Rothe, and M. Schienle (2012): "Nonparametric regression with nonparametrically generated covariates," *Annals of Statistics*, 40(2), 1132–1170.

Pagan, A., and A. Ullah (1999): *Nonparametric econometrics*. Cambridge University Press.

Pinkse, J. (2001): "Nonparametric regression estimation using weak separability," Discussion paper, UBC.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989): "Semiparametric estimation of index coefficients," *Econometrica: Journal of the Econometric Society*, pp. 1403–1430.

Rilstone, P. (1996): "Nonparametric estimation of models with generated regressors," *International Economic Review*, pp. 299–313.

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

van der Vaart, A., and J. Wellner (1996): *Weak convergence and empirical processes*. Springer.

Vytlacil, E. (2006): "A note on additive separability and latent index models of binary choice: representation results," *Oxford Bulletin of Economics and Statistics*, 68(4), 515–518.

Appendix A. Lemmas for Identification

Lemma A1. *Suppose that assumptions G and H are satisfied. Then,*

$$\lim_{t\to\infty} \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|V_t) = \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v^*),$$

where $V_t = \big(p(z_t), p(z_{t+1})\big]$ *and* $\{p(z_t)\}$ *is as in assumption H.*

Proof. Choose $\epsilon > 0$. By assumption G there exists a $v_\epsilon < v^*$ such that for all $v \in (v_\epsilon, v^*]$,

$$\mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v^*) - \epsilon < \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v) < \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v^*) + \epsilon.$$

Recalling that $\mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v^*)$ is the smallest value of u for which $\mathbb{P}(\boldsymbol{u} \le u|\boldsymbol{v} = v^*) \ge \tau^*$, it follows that for all $v \in (v_\epsilon, v^*]$,

$$\mathbb{P}\big\{\boldsymbol{u} \le \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v^*) - \epsilon|\boldsymbol{v} = v\big\} < \tau^* \le \mathbb{P}\big\{\boldsymbol{u} \le \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v^*) + \epsilon|\boldsymbol{v} = v\big\}.$$

Hence, if one picks t large enough to ensure that $v_\epsilon < p(z_t) < p(z_{t+1}) < v^*$ then

$$\mathbb{P}\big\{\boldsymbol{u} \le \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v^*) + \epsilon|\boldsymbol{v} \in V_t\big\} = \frac{\int_{p(z_t)}^{p(z_{t+1})} \mathbb{P}\big\{\boldsymbol{u} \le \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v^*) + \epsilon|\boldsymbol{v} = v\big\}dv}{p(z_{t+1}) - p(z_t)} \ge \tau^*, \quad (24)$$

and similarly

$$\mathbb{P}\big\{\boldsymbol{u} \le \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v^*) - \epsilon|\boldsymbol{v} \in V_t\big\} < \tau^*. \quad (25)$$

Hence, it follows from (24) and (25) that

$$\mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v^*) - \epsilon < \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|V_t) \leq \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v^*) + \epsilon.$$

Since $\epsilon > 0$ was arbitrarily chosen, the proof is done. □

Lemma A2. *Suppose that assumptions F to H are satisfied. Then,*

$$\sup_{V \in \mathcal{D}_L(x^*, v^*)} \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|V) = \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v^*) \leq \inf_{V \in \mathcal{D}_U(x^*, v^*)} \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|V). \tag{26}$$

Proof. By assumption F and definition of $\mathcal{D}_L(x^*, v^*)$ and $\mathcal{D}_U(x^*, v^*)$, we have

$$\sup_{V \in \mathcal{D}_L(x^*, v^*)} \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|V) \leq \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v^*) \leq \inf_{V \in \mathcal{D}_U(x^*, v^*)} \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|V). \tag{27}$$

Now let $V_t = \big(p(z_t), p(z_{t+1})\big]$, where $\{p(z_t)\}$ be as in assumption G. Since $V_t \in \mathcal{D}_L(x^*, v^*)$, we have

$$\forall t, \quad \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|V_t) \leq \sup_{V \in \mathcal{D}_L(x^*, v^*)} \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|V).$$

Therefore, it follows from lemma A1 that

$$\mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|v^*) \leq \sup_{V \in \mathcal{D}_L(x^*, v^*)} \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|V). \tag{28}$$

Combining (27) with (28) completes the proof. □

Appendix B. Technical Lemmas

Lemma B1. *Let $\Xi_n = \sum_{i=1}^n \boldsymbol{\xi}_{ni}$, where $\{\boldsymbol{\xi}_{ni}\}$ is an i.i.d. mean zero sequence of functions whose elements can depend on n. For any compact set Υ, suppose that $\tilde{\zeta}_n \succ 1$ is such that $\sup_{\upsilon \in \Upsilon} \|\partial_\upsilon \Xi_n(\upsilon)\| \preceq \tilde{\zeta}_n$, let $\sigma^2_{n\xi} = \sup_{\upsilon \in \Upsilon} \mathbb{V}\boldsymbol{\xi}_{ni}(\upsilon)$ and let $\bar{\xi}_n$ be such that $\mathbb{P}\{\sup_{\upsilon \in \Upsilon} |\boldsymbol{\xi}_{ni}(\upsilon)| > \bar{\xi}_n\} = 0$. If $\sigma^2_{n\xi} \prec 1/n \log \tilde{\zeta}_n$ and $\bar{\xi}_n \prec 1/\log \tilde{\zeta}_n$ then $\sup_{\upsilon \in \Upsilon} |\Xi_n(\upsilon)| \prec 1$.*

Proof. Cover Υ using ζ_n balls $\Upsilon_1, \ldots, \Upsilon_{\zeta_n}$ with centroids $\upsilon_1, \ldots, \upsilon_{\zeta_n}$, in such a way that for any n, $\max_{t=1,\ldots,\zeta_n} \sup_{\upsilon \in \Upsilon_t} \|\upsilon - \upsilon_t\| \leq C/\zeta_n^{1/d_\upsilon}$ for some C independent of n. Then

$$\sup_{\upsilon \in \Upsilon} |\Xi_n(\upsilon)| \leq \max_{t=1,\ldots,\zeta_n} \sup_{\upsilon \in \Upsilon_t} |\Xi_n(\upsilon) - \Xi_n(\upsilon_t)| + \max_{t=1,\ldots,\zeta_n} |\Xi_n(\upsilon_t)|. \tag{29}$$

Choose any $\epsilon > 0$. For $\delta > 0$ to be chosen, let $\zeta_n = (\delta \tilde{\zeta}_n/\epsilon)^{d_\upsilon}$.

For RHS2 in (29) we have by the Bernstein inequality that

$$\mathbb{P}\Big\{\max_{t=1,\ldots,\zeta_n}|\Xi_n(\upsilon_t)|>\epsilon\Big\}\le\sum_{t=1}^{\zeta_n}\mathbb{P}\big\{|\Xi_n(\upsilon_t)|>\epsilon\big\}\le 2\zeta_n\exp\Big\{-\frac{\epsilon^2}{2(n\sigma_{n\xi}^2+\bar{\xi}_n\epsilon)}\Big\}\prec 1.$$

Finally, for RHS1 in (29) we have by the mean value theorem that

$$\begin{aligned}\mathbb{P}\Big\{\max_{t=1,\ldots,\zeta_n}\sup_{\upsilon\in\Upsilon_t}|\Xi_n(\upsilon)-\Xi_n(\upsilon_t)|>\epsilon\Big\}&\le\mathbb{P}\Big\{\sup_{\upsilon\in\Upsilon}\|\partial_\upsilon\Xi_n(\upsilon)\|\max_{t=1,\ldots,\zeta_n}\sup_{\upsilon\in\Upsilon_t}\|\upsilon-\upsilon_t\|>\epsilon\Big\}\\&\le\mathbb{P}\Big\{\sup_{\upsilon\in\Upsilon}\|\partial_\upsilon\Xi_n(\upsilon)\|>\frac{\epsilon\zeta_n^{1/d_\upsilon}}{C}\Big\}\le\mathbb{P}\Big\{\sup_{\upsilon\in\Upsilon}\|\partial_\upsilon\Xi_n(\upsilon)\|>\frac{\delta\tilde{\zeta}_n}{C}\Big\}.\end{aligned}$$

Let $n\to\infty$ followed by $\delta\to\infty$. □

Lemma B2. *Let $\{\boldsymbol{\xi}_{\boldsymbol{ni}}^*\}$ be an i.i.d. sequence of mean zero functions defined on a compact set Υ for which for some $C<\infty$, $\sup_n\big[n^{-C}E\{\sup_{\upsilon\in\Upsilon}\|\partial_\upsilon\boldsymbol{\xi}_{\boldsymbol{ni}}^*(\upsilon)\|\}\big]<\infty$. Let further $\sigma_{n\xi^*}^2=\sup_{\upsilon\in\Upsilon}\mathbb{V}\boldsymbol{\xi}_{\boldsymbol{ni}}^*(\upsilon)$ and $\sup_n\mathbb{P}\big(\sup_{\upsilon\in\Upsilon}|\boldsymbol{\xi}_{\boldsymbol{ni}}^*(\upsilon)|>\bar{\xi}_n^*\big)=0$. Then for any $\zeta_n\succ\max\big(\sqrt{\sigma_{n\xi^*}^2\log n/n},\bar{\xi}_n^*\log n/n\big)$,*

$$\sup_{\upsilon\in\Upsilon}\Big|n^{-1}\sum_{i=1}^n\boldsymbol{\xi}_{\boldsymbol{ni}}^*(\upsilon)\Big|\prec\zeta_n.$$

Proof. In lemma B1 take $\boldsymbol{\xi}_{ni}=\boldsymbol{\xi}_{\boldsymbol{ni}}^*/n\zeta_n$. □

Lemma B3. *Let $\{\boldsymbol{\zeta}_i\}$ be an i.i.d. sequence, let $\boldsymbol{\zeta}_i$ include $\boldsymbol{y}_i$ as an element, and let $\{\hat{\boldsymbol{A}}_i\}$ be such that $\hat{\boldsymbol{A}}_i=A_n(\boldsymbol{\zeta}_i,\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{i-1},\boldsymbol{\zeta}_{i+1},\ldots,\boldsymbol{\zeta}_n)$ for arbitrary function A_n. If $\mathbb{P}\big(\|\hat{\boldsymbol{A}}_1\|>\epsilon\big)\prec 1/n$ and $\sup_{y\in\mathcal{Y}}\mathbb{P}\Big\{\Big\|\sum_{i=1}^{n-1}\hat{\boldsymbol{A}}_i\boldsymbol{I}_i(y)\Big\|>\epsilon\Big\}\prec 1/n$, then $\sup_{y\in\mathcal{Y}}\big\|\sum_i\hat{\boldsymbol{A}}_i\boldsymbol{I}_i(y)\big\|\prec 1$.*

Proof.

$$\begin{aligned}\mathbb{P}\Big\{\sup_{y\in\mathcal{Y}}\Big\|\sum_{i=1}^n\hat{\boldsymbol{A}}_i\boldsymbol{I}_i(y)\Big\|>2\epsilon\Big\}&=\mathbb{P}\Big\{\max_{t=1,\ldots,n}\Big\|\sum_{i=1}^n\hat{\boldsymbol{A}}_i\boldsymbol{I}_i(\boldsymbol{y}_t)\Big\|>2\epsilon\Big\}\le\sum_{t=1}^n\mathbb{P}\Big\{\Big\|\sum_{i=1}^n\hat{\boldsymbol{A}}_i\boldsymbol{I}_i(\boldsymbol{y}_t)\Big\|>2\epsilon\Big\}\\&\le n\sup_{y\in\mathcal{Y}}\mathbb{P}\Big\{\Big\|\sum_{i=1}^{n-1}\hat{\boldsymbol{A}}_i\boldsymbol{I}_i(y)\Big\|>\epsilon\Big\}+n\mathbb{P}\big\{\|\hat{\boldsymbol{A}}_1\|>\epsilon\big\}\prec 1.\end{aligned}$$ □

Appendix C. V–statistics

Let $\aleph_n=\{1,\ldots,n\}$, $\Upsilon_{n\ell}=\aleph_n^\ell$, and let $\Upsilon_{n\ell j}$ be the set of vectors in $\Upsilon_{n\ell}$ with exactly j distinct elements. Let further for any $\iota\in\Upsilon_{n\ell}$, $\boldsymbol{\zeta}_{\boldsymbol{\iota}}=(\boldsymbol{\zeta}_{\boldsymbol{\iota_1}},\ldots,\boldsymbol{\zeta}_{\boldsymbol{\iota_\ell}})^\intercal$.

Lemma C1. *For* $\boldsymbol{V}_{n\ell} = \sum_{\iota\in\Upsilon_{n\ell}} m(\boldsymbol{\zeta}_\iota)$ *and* $\boldsymbol{U}_n^{(\ell,j)} = \sum_{\iota\in\Upsilon_{n\ell j}} m(\boldsymbol{\zeta}_\iota)$*, we have* $\boldsymbol{V}_{n\ell} = \sum_{j=1}^{\ell} \boldsymbol{U}_n^{(\ell,j)}$*, where* $\boldsymbol{U}_n^{(\ell,j)}$ *is a U–statistic of order* j *whose kernel* $m^{(\ell,j)}$ *consists of a sum of* $\sum_{t=1}^{j}(-1)^{j-t}t^{\ell-1}/\{(j-t)!(t-1)!\}$ *elements.*[14]

Proof. See Lee (1990, theorem 1 on p.183). □

Lemma C2. *For any symmetric* j*–th order U–statistic kernel* $m^{(j)}$*, let* $\boldsymbol{U}_{nj} = \sum_{\iota\in\Upsilon_{njj}} m^{(j)}(\boldsymbol{\zeta}_\iota)$*. Let further for* $0 \le t \le j$ *and any* $a_1,\ldots,a_t$*,* $m_t^{(j)}(a) = \mathbb{E}m^{(j)}(a_1,\ldots,a_t,\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{j-t})$*,* $\boldsymbol{U}_{njt} = \sum_{\iota\in\Upsilon_{ntt}} m_t^{(j)}(\boldsymbol{\zeta}_\iota)$*, and* $\boldsymbol{U}_{njt}^c$ *the corresponding canonical U–statistic (De la Peña and Giné, 1999). Then if* $\mu^{(j)} = \mathbb{E}m^{(j)}(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_j)$*,*

$$\boldsymbol{U}_{nj} = \frac{n!}{(n-j)!}\mu^{(j)} + \sum_{t=1}^{j}\binom{j}{t}\frac{(n-t)!}{(n-j)!}\boldsymbol{U}_{njt}^c.$$

Proof. This is essentially the Hoeffding decomposition (Lee, 1990, theorem 1 on p.26) combined with a rearrangement of terms.[15] □

Lemma C3. *For* $\boldsymbol{U}_{njt}^c$ *defined in lemma C2, we have*

$$\mathbb{P}\big(|\boldsymbol{U}_{njt}^c| > \epsilon\big) \le C_t \exp\Big[-\frac{\epsilon^{2/t}/C_t}{n\sigma_{jt}^{2/t} + \beta_{jt}^{2/(t+1)}n^{(t-1)/(t+1)}\epsilon^{2/\{t(t+1)\}}}\Big],$$

where C_t *is a constant which only depends on* t*,* $\beta_{jt} = \sup m_t^{(j)}(\cdot)$*,* $\sigma_{jt}^2 = \mathbb{V}m_t^{(j)}(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_t)$*.*

Proof. Follows from Arcones and Giné (1993, proposition 2.3(c)). □

Lemma C4. *For an* ℓ*–th order V–statistic* $\boldsymbol{V}_{n\ell}$ *as defined in lemma C1 with symmetric kernel* m*, let for* $1 \le t \le j \le \ell$*,* $m^{(\ell,j)}$ *be defined as in lemma C1,* $\mu^{(\ell,j)} = \mathbb{E}m^{(\ell,j)}(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_j)$*,* $m_t^{(\ell,j)}(a) = \mathbb{E}m^{(\ell,j)}(a_1,\ldots,a_t,\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{j-t})$*,* $\beta_t^{(\ell,j)} = \sup m_t^{(\ell,j)}$*, and* $\sigma_t^{(\ell,j)} = \sqrt{\mathbb{V}m_t^{(\ell,j)}(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_t)}$*. Then* $\mathbb{P}(\boldsymbol{V}_{n\ell} > \epsilon_n)$ *decreases faster than any polynomial of* n*, where*

$$\epsilon_n = \max_{1\le t\le j\le \ell}\Big[(\log n)^{t+1}\max\Big\{n^{t/2}\sigma_t^{(\ell,j)}, n^{(t-1)/2}\beta_t^{(\ell,j)}, n^j\mu^{(\ell,j)}\Big\}\Big].$$

[14] These are Stirling numbers of the second kind.

[15] The representation is slightly different here from the one in Lee (1990) because the U–statistic kernel incorporates a number of permutations in his case.

Proof. In lemma C1 the V–statistic is separated into a number (independent of n) of U–statistics. Each of these U–statistics is further separated into a number (again independent of n) of canonical U–statistics in lemma C2 plus a mean. Finally, apply lemma C3 to each element individually.[16] □

Appendix D. $\mathcal{Z}$

Let π be such that $\pi\{p(z_1, \tilde{z}), \tilde{z}\} = z_1$ for all $(z_1, \tilde{z}) \in \mathcal{Z}$. The function π is well–defined by assumption P.

Lemma D1. *For all p and any $c > 0$, $f_p(p|\tilde{z})$ is four times boundedly differentiable with respect to p, uniformly in $\tilde{z}$ for which $\tilde{f}(\tilde{z}) \geq c$.*

Proof. Note that $F_p(p|\tilde{z}) = \mathbb{P}(\boldsymbol{p}_i \leq p|\tilde{\boldsymbol{z}}_i = \tilde{z}) = \mathbb{P}\{\boldsymbol{z}_{i1} \leq \pi(p, \tilde{z})|\tilde{\boldsymbol{z}}_i = \tilde{z}\}$, such that $f_p(p|\tilde{z}) = \partial_p \pi(p, \tilde{z}) f\{\pi(p, \tilde{z}), \tilde{z}\}/\tilde{f}(\tilde{z})$. The stated result then follows from assumption O. □

Lemma D2. *For all p for which $\exists z \in \mathcal{Z} : p(z) = p$ and all t times boundedly differentiable functions ν for which $\nu(z) = 0$ for all $z \notin \mathcal{Z}$, $\mathbb{E}\{\nu(\boldsymbol{z}_i)|\boldsymbol{p}_i = p\} f_p(p)$ is $\min(t, 3)$ times boundedly differentiable in p.*

Proof. Let π be as in lemma D1. Then for any $z \in \mathcal{Z}$ and $p = p(z)$,

$$\mathbb{E}\{\nu(\boldsymbol{z}_i)|\boldsymbol{p}_i = p\} f_p(p) = \int \nu\{\pi(p, \tilde{z}), \tilde{z}\} f\{\pi(p, \tilde{z}), \tilde{z}\} \partial_p \pi(p, \tilde{z}) d\tilde{z}.$$ □

Appendix E. Kernels

Lemma E1. *Let $\{(\boldsymbol{\xi}_i, \boldsymbol{z}_i)\}$ be i.i.d., and suppose that $\mu(z) f(z)$ with $\mu(z) = \mathbb{E}(\boldsymbol{\xi}_i|\boldsymbol{z}_i = z)$ has two bounded derivatives. Then*

$$\sup_{z \in \mathcal{Z}} \left| \mathbb{E}\left\{ \boldsymbol{K}_{zi}(z) \boldsymbol{\xi}_i \right\} - \mu(z) f(z) \right| \preceq h_z^2.$$

Proof. This follows from a standard kernel bias expansion. □

Lemma E2 can be found, often in slightly different form, in many other sources, including Pagan and Ullah (1999).

[16] Because the number of canonical U–statistics has an upper bound independent of n, looking at each individual term separately is sufficient.

Lemma E2. *Let* $\{(\boldsymbol{\xi}_i, \boldsymbol{z}_i)\}$ *be i.i.d.,* $\boldsymbol{\xi}_i$ *uniformly bounded, and* $\sigma^2_{\boldsymbol{\xi}}(z) = \mathbb{V}(\boldsymbol{\xi}_i | \boldsymbol{z}_i = z)$ *is continuous on* $\mathcal{Z}$*. Then*

$$\sup_{z\in\mathcal{Z}} \left| \frac{1}{n} \sum_{i=1}^{n} \left[\boldsymbol{K}_{zi}(z)\boldsymbol{\xi}_i - \mathbb{E}\{\boldsymbol{K}_{zi}(z)\boldsymbol{\xi}_i\} \right] \right| \prec \frac{\log n}{\sqrt{n h_z^{d_z}}}.$$

Proof. Follows directly from lemma B2. □

Let

$$\alpha_n = \log n / \sqrt{n h_z^{d_z}} + h_z^2. \tag{30}$$

Let further $\hat{\boldsymbol{r}} = \hat{\boldsymbol{p}}\hat{\boldsymbol{f}}$, where $\hat{\boldsymbol{f}}$ is the kernel density estimator of f using bandwidth h_z and kernel K.

Lemma E3. *(i)* $\sup_{z\in\mathcal{Z}} |\hat{\boldsymbol{f}}(z) - f(z)| \preceq \alpha_n$*, (ii)* $\sup_{z\in\mathcal{Z}} |\hat{\boldsymbol{r}}(z) - r(z)| \preceq \alpha_n$*, (iii)* $\sup_{z\in\mathcal{Z}} |\hat{\boldsymbol{p}}(z) - p(z)| \preceq \alpha_n$*,*

Proof. The first two results follows by combining lemmas E1 and E2 and the third one from the first two by noting that for any $\tilde{z} \in \mathcal{Z}$,

$$\frac{r(\tilde{z}) - \sup_{z\in\mathcal{Z}} |\hat{\boldsymbol{r}}(z) - r(z)|}{f(\tilde{z}) + \sup_{z\in\mathcal{Z}} |\hat{\boldsymbol{f}}(z) - f(z)|} \leq \hat{\boldsymbol{p}}(\tilde{z}) \leq \frac{r(\tilde{z}) + \sup_{z\in\mathcal{Z}} |\hat{\boldsymbol{r}}(z) - r(z)|}{f(\tilde{z}) - \sup_{z\in\mathcal{Z}} |\hat{\boldsymbol{f}}(z) - f(z)|}.$$

□

Lemma E4. *For some* $\epsilon > 0$*,* $\lim_{n\to\infty} \mathbb{P}\{\inf_{z\in\mathcal{Z}} \hat{\boldsymbol{f}}(z) < \epsilon\} = 0$*.*

Proof. Note that

$$\mathbb{P}\left\{ \inf_{z\in\mathcal{Z}} \hat{\boldsymbol{f}}(z) < \epsilon \right\} \leq I\left\{ \inf_{z\in\mathcal{Z}} f(z) < 2\epsilon \right\} + \mathbb{P}\left\{ \sup_{z\in\mathcal{Z}} |\hat{\boldsymbol{f}}(z) - f(z)| > \epsilon \right\}.$$

Apply lemma E3. □

Appendix F. Expansions

Recalling (16), let $\mathcal{E}(y; p) = \mathbb{E}\{\boldsymbol{I}_i(y) | \boldsymbol{a}_i = a^*, \boldsymbol{p}_i = p\}$, $\mathcal{E}(y) = \mathcal{E}(y; v^*)$, and

$$\bar{\boldsymbol{S}}_s(y; p) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{k}_{is}^{(s)} \boldsymbol{K}_{ais} \mathcal{E}(y), \qquad \bar{\boldsymbol{S}}_s(y; \hat{p}) = \frac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{k}}_{is}^{(s)} \boldsymbol{K}_{ais} \mathcal{E}(y). \tag{31}$$

Let further $\boldsymbol{K}_{zij} = \boldsymbol{K}_{zi}(\boldsymbol{z}_j)$.

Lemmas F1 to F3 serve as inputs into establishing two results, namely

$$\sup_y \left| \hat{\boldsymbol{S}}_s(y; \hat{p}) - \bar{\boldsymbol{S}}_s(y; \hat{p}) - \hat{\boldsymbol{S}}_s(y; p) + \bar{\boldsymbol{S}}_s(y; p) \right| \prec 1/\rho_n \tag{32}$$

for $s = 0, 1$ and

$$\sup_y \left| \bar{\boldsymbol{S}}_0(y; \hat{p}) - \bar{\boldsymbol{S}}_0(y; p) \right| \prec 1/\rho_n, \tag{33}$$

i.e. lemmas F6 and F7. Each of these expression is expanded using the mean value theorem to some order J to apply lemmas F1 to F3. For instance, by assumption R, the RHS of (32) is bounded above by

$$\sum_{j=1}^{J-1} \frac{1}{j!} \sup_{y \in \mathcal{Y}} \left| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{k}_{is}^{(s+j)} \boldsymbol{K}_{ais} (\boldsymbol{p}_i - \hat{\boldsymbol{p}}_i)^j \left\{ \boldsymbol{I}_{i_0}(y) - \mathcal{E}(y) \right\} \right| + \sup_{y \in \mathcal{Y}} \frac{1}{n h_s^{s+1+J} J!} \sum_{i=1}^{n} \left| \boldsymbol{K}_{ais} (\boldsymbol{p}_i - \hat{\boldsymbol{p}}_i)^J \left\{ \boldsymbol{I}_{i_0}(y) - \mathcal{E}(y) \right\} \right|, \tag{34}$$

where the first term is covered by lemmas F1 and F2 and the second term is dealt with in lemma F3.

Lemma F1. *Let $\zeta_1, \ldots, \zeta_j, \zeta^* \in \mathcal{F}_2$, and $\mathcal{T}_j \subset \mathbb{R}^j$ consist of vectors whose elements are all either equal to one or zero and let $\boldsymbol{u}_{\ell i}^*$ be such that $\mathbb{E}(\boldsymbol{u}_{\ell i}^* | \boldsymbol{z}_i) = 0$ a.s. and $\mathbb{E}(\boldsymbol{u}_{\ell i}^{*2} | \boldsymbol{z}_i = z)$ is continuous on $\mathcal{Z}$. Then for $s = 0, 1$, $j = 1, 2, \ldots$, and all $t \in \mathcal{T}_j$,*

$$\sup_{y \in \mathcal{Y}} \left| \frac{1}{n^{j+1}} \sum_{i_0=1}^{n} \cdots \sum_{i_j=1}^{n} \boldsymbol{k}_{i_0 s}^{(s+j)} \boldsymbol{K}_{a i_0 s} \boldsymbol{\zeta}_{i_0}^* \left\{ \boldsymbol{I}_{i_0}(y) - \mathcal{E}(y) \right\} \prod_{\ell=1}^{j} \boldsymbol{K}_{z i_0 i_\ell} \boldsymbol{u}_{\ell i_\ell}^{*t_\ell} (\boldsymbol{\zeta}_{\ell i_\ell} - \boldsymbol{\zeta}_{\ell i_0})^{1-t_\ell} \right| \prec 1/\rho_n, \tag{35}$$

where $\boldsymbol{\zeta}_{\ell i} = \zeta_\ell(\boldsymbol{z}_i)$ and similarly for other ζ symbols.

Proof. As will become apparent in lemma F6, for every j the LHS in (35) corresponds to the j–th term in a Taylor expansion of $\hat{\boldsymbol{S}}_s(y; \hat{p}) - \bar{\boldsymbol{S}}_s(y; \hat{p})$ around $\hat{\boldsymbol{S}}_s(y; p) - \bar{\boldsymbol{S}}_s(y; p)$; see (34). Because by lemma E3 $\hat{\boldsymbol{p}}_i - \boldsymbol{p}_i$ converges faster (uniformly in i) than the extra $1/h_s$ incurred for each additional derivative, the convergence rate is slowest for $s = j = 1$, so we establish convergence at the promised rate for that case; all other cases can be verified similarly, albeit sometimes more painfully.

Thus, we use lemma C4 to obtain a rate for

$$\sup_{y \in \mathcal{Y}} \left| \frac{1}{n^2} \sum_{i_0=1}^{n} \sum_{i_1=1}^{n} \boldsymbol{k}_{i_0 1}'' \boldsymbol{K}_{a i_0 1} \boldsymbol{\zeta}_{i_0}^* \left\{ \boldsymbol{I}_{i_0}(y) - \mathcal{E}(y) \right\} \boldsymbol{K}_{z i_0 i_1} \boldsymbol{u}_{1 i_1}^{*t_1} (\boldsymbol{\zeta}_{1 i_1} - \boldsymbol{\zeta}_{1 i_0})^{1-t_1} \right|. \tag{36}$$

Let $\boldsymbol{\xi}_i$ contain all random variables pertaining to observation i. Noting that (36) is a V statistic and that lemma C4 is based on a decomposition of the V statistic into a sum of U statistics, we have

for the m–symbols of lemma C4 and for some $\bar{\zeta}, \tilde{\zeta} \in \mathcal{F}_0$,

$$
\begin{aligned}
m^{(2,2)}(\boldsymbol{\xi}_{i_0}, \boldsymbol{\xi}_{i_1}) &= \frac{1}{2n^2}\Big[\boldsymbol{k}''_{i_0 1}\boldsymbol{K}_{a i_0 1}\boldsymbol{\zeta}^*_{i_0}\{\boldsymbol{I}_{i_0}(y) - \mathcal{E}(y)\}\boldsymbol{K}_{z i_0 i_1}\boldsymbol{u}^{*t_1}_{1 i_1}(\boldsymbol{\zeta}_{1 i_1} - \boldsymbol{\zeta}_{1 i_0})^{1-t_1} \\
&\quad + \boldsymbol{k}''_{i_1 1}\boldsymbol{K}_{a i_1 1}\boldsymbol{\zeta}^*_{1 i_1}\{\boldsymbol{I}_{i_1}(y) - \mathcal{E}(y)\}\boldsymbol{K}_{z i_0 i_1}\boldsymbol{u}^{*t_1}_{1 i_0}(\boldsymbol{\zeta}_{1 i_0} - \boldsymbol{\zeta}_{1 i_1})^{1-t_1}\Big], \\
m_1^{(2,2)}(\boldsymbol{\xi}_i) &= \mathbb{E}\{m^{(2,2)}(\boldsymbol{\xi}_i, \boldsymbol{\xi}_{i_1})|\boldsymbol{\xi}_i\} \qquad (i_1 \neq i) \\
&\simeq \begin{cases} \frac{h_z^2}{2n^2}\Big[\boldsymbol{k}''_{i1}\boldsymbol{K}_{ai1}\boldsymbol{\zeta}^*_i\{\boldsymbol{I}_i(y) - \mathcal{E}(y)\}\bar{\boldsymbol{\xi}}_i, \\ \qquad +\boldsymbol{k}''_{i1}\boldsymbol{K}_{ai1}\{\mathcal{E}(y; \boldsymbol{p}_i, \boldsymbol{a}_i) - \mathcal{E}(y; v^*, a^*)\}\tilde{\bar{\boldsymbol{\xi}}}_i\Big], & t_1 = 0, \\ \frac{h_z^2}{2n^2}\boldsymbol{k}''_{i1}\boldsymbol{K}_{ai1}\boldsymbol{\xi}^*_i\{\boldsymbol{I}_i(y) - \mathcal{E}(y)\}\bar{\boldsymbol{\xi}}_i\boldsymbol{u}^*_{i1}, & t_1 = 1, \end{cases} \\
m_1^{(2,1)}(\boldsymbol{\xi}_i) &= \begin{cases} 0, & t_1 = 0, \\ \frac{1}{n^2}\boldsymbol{k}''_{i1}K_{ai1}\boldsymbol{\zeta}^*_i\{\boldsymbol{I}_i(y) - \mathcal{E}(y)\}\boldsymbol{K}_{zii}\boldsymbol{u}^*_{i1}, & t_1 = 1. \end{cases} \\
n^2\mu^{(2,2)} &\preceq h_z^2, \qquad (37) \\
n\mu^{(2,1)} &\preceq \frac{1}{nh_z^{d_z}}, \qquad (38) \\
\sqrt{n}\beta_2^{(2,2)} &\preceq \frac{1}{n^{3/2}h_z^{d_z}h_1^{3+d_a}}, \qquad (39) \\
\beta_1^{(2,2)} &\preceq \frac{h_z^2}{n^2h_1^{3+d_a}}, \qquad (40) \\
\beta_1^{(2,1)} &\preceq \frac{1}{n^2h_z^{d_z}h_1^{3+d_a}}, \qquad (41) \\
n\sigma_2^{(2,2)} &\preceq \frac{1}{nh_1^{3+d_a}}, \qquad (42) \\
\sqrt{n}\sigma_1^{(2,2)} &\preceq \frac{h_z^2}{n^{3/2}h_1^{(5+d_a)/2}}, \qquad (43) \\
\sqrt{n}\sigma_1^{(2,1)} &\preceq \frac{1}{n^{3/2}h_1^{(5+d_a)/2}h_z^{d_z}}. \qquad (44)
\end{aligned}
$$

Sufficient conditions for (37) to (44) to converge at a rate faster than $1/\rho_n$ are respectively

$$(3 + d_a)\eta_1 > 1 - 4\eta_z, \qquad (45)$$

$$(3 + d_a)\eta_1 > 2d_z\eta_z - 1, \qquad (46)$$

$$(3 + d_a)\eta_1 < 2 - 2\eta_z d_z, \qquad (47)$$

$$(3+d_a)\eta_1 < 3+4\eta_z, \tag{48}$$

$$(3+d_a)\eta_1 < 3-2\eta_z d_z, \tag{49}$$

$$(3+d_a)\eta_1 < 1, \tag{50}$$

$$\eta_1 < 1+2\eta_z, \tag{51}$$

$$\eta_1 < 1-\eta_z d_z. \tag{52}$$

Conditions (48) to (52) are implied by (47) and/or standard kernel estimation conditions needed for consistency of the estimator of H without nuisance parameters. Thus, only (45) to (47) are potentially relevant and the lemma statement holds if

$$\max(1-4\eta_z, 2d_z\eta_z - 1, 0) < (3+d_a)\eta_1 < \min(2-2\eta_z d_z, 1),$$

which is satisfied by assumption S. □

Lemma F2. *Using essentially the same notation and conditions as in lemma F1, for* $j = 1, 2, \ldots,$

$$\left| \frac{1}{n^{j+1}} \sum_{i_0=1}^{n} \cdots \sum_{i_j=1}^{n} \boldsymbol{k}_{i_0 0}^{(j)} \boldsymbol{K}_{a i_0 0} \prod_{\ell=1}^{j} \boldsymbol{K}_{z i_0 i_\ell} \boldsymbol{u}_{\ell i_\ell}^{*t_\ell} (\zeta_{\ell i_\ell} - \zeta_{\ell i_0})^{1-t_\ell} \right| \prec 1/\rho_n. \tag{53}$$

Proof. This lemma is used in lemma F7 to deal with the expansion of $\bar{\boldsymbol{S}}_0(y;\hat{p})$ around $\bar{\boldsymbol{S}}_0(y;p)$. Using the same strategy and rationale for focusing on the case $s=1$ as in lemma F1, we have

$$\begin{aligned} m^{(2,2)}(\boldsymbol{\xi}_{i_0}, \boldsymbol{\xi}_{i_1}) = \frac{1}{2n^2}\Big[& \boldsymbol{k}'_{i_0 0} \boldsymbol{K}_{a i_0 0} \boldsymbol{K}_{z i_0 i_1} \boldsymbol{u}_{1 i_1}^{*t} (\zeta_{i_1} - \zeta_{i_0})^{1-t} \\ & + \boldsymbol{k}'_{i_1 0} \boldsymbol{K}_{a i_1 0} \boldsymbol{K}_{z i_0 i_1} \boldsymbol{u}_{1 i_0}^{*t} (\zeta_{i_0} - \zeta_{i_1})^{1-t} \Big], \end{aligned}$$

$$m_1^{(2,2)}(\boldsymbol{\xi}_i) \simeq \begin{cases} \frac{h_z^2}{n^2} \boldsymbol{k}'_{i0} \boldsymbol{K}_{ai0} \bar{\boldsymbol{\xi}}_i, & t=0, \\ \frac{1}{2n^2} \boldsymbol{k}'_{i0} \boldsymbol{K}_{ai0} \bar{\boldsymbol{\xi}}_i \boldsymbol{u}_{i1}^*, & t=1. \end{cases}$$

$$m_1^{(2,1)}(\boldsymbol{\xi}_i) = \begin{cases} 0, & t=0, \\ \frac{1}{n^2} \boldsymbol{k}'_{i0} \boldsymbol{K}_{ai0} \boldsymbol{K}_{zii} \boldsymbol{u}_{i1}^*, & t=1, \end{cases}$$

$$n^2 \mu^{(2,2)} \preceq h_z^2, \tag{54}$$

$$n\mu^{(2,1)} \preceq \frac{1}{n h_z^{d_z}}, \tag{55}$$

$$\sqrt{n}\beta_2^{(2,2)} \preceq \frac{1}{n^{3/2} h_0^{d_a+2} h_z^{d_z}}, \tag{56}$$

$$\beta_1^{(2,2)} \preceq \frac{1}{n^2 h_0^{d_a+2}}, \tag{57}$$

$$\beta_1^{(2,1)} \preceq \frac{1}{n^2 h_0^{d_a+2} h_z^{d_z}}, \tag{58}$$

$$n\sigma_2^{(2,2)} \preceq \frac{1}{n h_0^{d_a+2}}, \tag{59}$$

$$\sqrt{n}\sigma_1^{(2,2)} \preceq \frac{1}{n^{3/2} h_0^{(d_a+3)/2}}, \tag{60}$$

$$\sqrt{n}\sigma_1^{(2,1)} \preceq \frac{1}{n^{3/2} h_0^{(d_a+3)/2}}. \tag{61}$$

Some of the numbered equations above were already shown to be $\prec 1/\rho_n$ in lemma F1. The remaining conditions are implied by

$$(3+d_a)\eta_1 > 2\eta_0(d_a+2) + 2\eta_z d_z - 2, \tag{62}$$

$$(3+d_a)\eta_1 > 2\eta_0(d_a+2) - 1, \tag{63}$$

$$(3+d_a)\eta_1 > (d_a+3)\eta_0 - 2, \tag{64}$$

which follow from assumption S. □

Lemma F3. *Using the same notation as in lemma F1, for some* $0 < J < \infty$ *and* $s = 0, 1$,

$$\sup_{y\in\mathcal{Y}} \frac{1}{n h_s^{s+1+J}} \sum_{i=1}^{n} \left| \boldsymbol{K}_{ais} \xi_i^* \{ \boldsymbol{I}_i(y) - \mathcal{E}(y) \} \right| \sup_{z\in\mathcal{Z}} \left| \hat{\boldsymbol{p}}(z) - p(z) \right|^J \prec 1/\rho_n.$$

Proof. By lemma E3, the stated result is implied by α_n / h_s decreasing to zero at a polynomial rate since for any polynomial π_n^* then $\pi_n^*(\alpha_n/h_s)^J \prec 1$. The requirement that α_n/h_s decrease to zero at a polynomial rate is guaranteed by $\eta_s < \min\{2\eta_z, (1-\eta_z d_z)/2\}$, which was assumed in assumption S. □

Lemma F4.

$$\sup_{z\in\mathcal{Z}} \left| \hat{\boldsymbol{p}}(z) - p(z) - \frac{\hat{\boldsymbol{r}}(z) - r(z)}{f(z)} \sum_{j=0}^{J-2} \left(\frac{f(z) - \hat{\boldsymbol{f}}(z)}{f(z)} \right)^j - p(z) \sum_{j=1}^{J-1} \left(\frac{f(z) - \hat{\boldsymbol{f}}(z)}{f(z)} \right)^j \right| \preceq \alpha_n^J.$$

Proof. From the recursion of $f/\hat{\boldsymbol{f}} = 1 + (f - \hat{\boldsymbol{f}})/\hat{\boldsymbol{f}}$, we have $f/\hat{\boldsymbol{f}} = \sum_{j=0}^{J-1}\{(f - \hat{\boldsymbol{f}})/f\}^j + \{(f - \hat{\boldsymbol{f}})/f\}^J f/\hat{\boldsymbol{f}}$. Therefore, the LHS in the lemma statement is bounded by

$$\sup_{z\in\mathcal{Z}}\left|\Big(\frac{\hat{\boldsymbol{r}}(z) - r(z)}{f(z)}\Big)\Big(\frac{f(z) - \hat{\boldsymbol{f}}(z)}{f(z)}\Big)^{J-1}\frac{f(z)}{\hat{\boldsymbol{f}}(z)} + \Big(\frac{f(z) - \hat{\boldsymbol{f}}(z)}{f(z)}\Big)^{J}\frac{f(z)}{\hat{\boldsymbol{f}}(z)}\right|.$$

Apply lemmas E3 and E4. □

Let $\bar{f}(z) = \mathbb{E}\hat{\boldsymbol{f}}(z)$, $\bar{r}(z) = \mathbb{E}\hat{\boldsymbol{r}}(z)$, $\hat{\boldsymbol{f}}^* = \hat{\boldsymbol{f}}/f$, $\hat{\boldsymbol{r}}^* = \hat{\boldsymbol{r}}/f$, $\bar{r}^* = \bar{r}/f$, $\bar{f}^* = \bar{r}/f$.

Lemma F5. *For given j, J, let Λ_j be the collection of vectors ℓ of dimension four containing non-negative integers satisfying $\ell_1 + \ell_2 \le j$ and $\ell_1 + \ell_2 + \ell_3 + \ell_4 < J$. Then for all sufficiently large n and any $1 \le j < J$ and some constants $C_{j\ell}$ independent of n, z,*

$$\sup_{z\in\mathcal{Z}}\Big|\{\hat{\boldsymbol{p}}(z) - p(z)\}^j - \sum_{\ell\in\Lambda_j} C_{j\ell}\{\hat{\boldsymbol{r}}^*(z) - \bar{r}^*(z)\}^{\ell_1}\{\bar{r}^*(z) - p(z)\}^{\ell_2}\{\hat{\boldsymbol{f}}^*(z) - \bar{f}^*(z)\}^{\ell_3}\{\bar{f}^*(z) - 1\}^{\ell_4}\Big| \preceq \alpha_n^J.$$

Proof. Follows directly from lemma F4 combined with the multinomial theorem. □

Lemma F6. *For $s = 0, 1$, $\sup_{y\in\mathcal{Y}}\big|\hat{\boldsymbol{S}}_s(y;\hat{p}) - \bar{\boldsymbol{S}}_s(y;\hat{p}) - \hat{\boldsymbol{S}}_s(y;p) + \bar{\boldsymbol{S}}_s(y;p)\big| \prec 1/\rho_n$.*

Proof. Let J be sufficiently large as in lemma F3. Then, expand the LHS of the lemma statement to order J using the mean value theorem to obtain an upper bound of (34). The second term in (34) is covered by lemma F3 and the first term in (34) is dealt with in lemma F1, using lemmas F4 and F5. □

Lemma F7. $\sup_{y\in\mathcal{Y}}\big|\bar{\boldsymbol{S}}_0(y;\hat{p}) - \bar{\boldsymbol{S}}_0(y;p)\big| \prec 1/\rho_n$.

Proof. The proof is entirely analogous to that of lemma F6, albeit using lemma F2 instead of lemma F1, and is hence omitted. □

Below we will write $S_s(y;p)$ for $S_s(y) = S_s(y;a^*,v^*)$ for $s = 0, 1$.

Lemma F8.

$$\sup_{y\in\mathcal{Y}}\big|\hat{\boldsymbol{S}}_0(y;p) - S_0(y;p)\big| \prec 1/\rho_n. \tag{65}$$

Proof. By standard kernel estimation theory, the squared LHS in (65) is $\preceq h_0^4 + 1/nh_0^{1+d_a} \prec 1/\rho_n^2$ by assumption S. □

Lemma F9.

$$\sup_{y\in\mathcal{Y}}\big|\hat{\boldsymbol{S}}_0(y;\hat{p}) - S_0(y;p)\big| \prec 1/\rho_n. \tag{66}$$

Proof. The LHS in (66) is bounded above by the sum of

$$\sup_{y\in\mathcal{Y}}\left|\hat{\boldsymbol{S}}_0(y;\hat{p})-\bar{\boldsymbol{S}}_0(y;\hat{p})-\hat{\boldsymbol{S}}_0(y;p)+\bar{\boldsymbol{S}}_0(y;p)\right|, \tag{67}$$

$$\sup_{y\in\mathcal{Y}}\left|\bar{\boldsymbol{S}}_0(y;\hat{p})-\bar{\boldsymbol{S}}_0(y;p)\right|. \tag{68}$$

$$\sup_{y\in\mathcal{Y}}\left|\hat{\boldsymbol{S}}_0(y;p)-S_0(y;p)\right|, \tag{69}$$

Apply lemmas F6 to F8. □

Lemma F10. *For all* $y\in\mathcal{Y}$, $\{\bar{\boldsymbol{S}}_1(y,\hat{p})-\bar{\boldsymbol{S}}_1(y,p)\}S_0(\infty,p)=\{\bar{\boldsymbol{S}}_1(\infty,\hat{p})-\bar{\boldsymbol{S}}_1(\infty,p)\}S_0(y,p)$.

Proof. Trivial. □

Lemma F11. *Let* $\hat{\boldsymbol{H}}(y;\hat{p})=\hat{\hat{\boldsymbol{H}}}(y|a^*,v^*)$ *and* $H(y;p)=H(y|a^*,v^*)$. *Then*

$$\sup_{y\in\mathcal{Y}}\left|\hat{\boldsymbol{H}}(y;\hat{p})-H(y;p)-v^*\frac{\{\hat{\boldsymbol{S}}_1(y;p)-S_1(y;p)\}S_0(\infty;p)-\{\hat{\boldsymbol{S}}_1(\infty;p)-S_1(\infty;p)\}S_0(y;p)}{S_0^2(\infty;p)}\right| \prec 1/\rho_n. \tag{70}$$

Proof. For the remainder of this lemma, let $\simeq_\rho$ mean that the left and right hand sides differ by a term $\prec 1/\rho_n$, uniformly in y. By lemma F9,

$$\hat{\boldsymbol{H}}(y,\hat{p})\simeq_\rho\frac{S_0(y;p)S_0(\infty;p)+v^*\{\hat{\boldsymbol{S}}_1(y;\hat{p})S_0(\infty;p)-\hat{\boldsymbol{S}}_1(\infty;\hat{p})S_0(y;p)\}}{S_0^2(\infty;p)}.$$

Since by lemma F6 $\hat{\boldsymbol{S}}_1(y;\hat{p})\simeq_\rho\hat{\boldsymbol{S}}_1(y;p)+\bar{\boldsymbol{S}}_1(y;\hat{p})-\bar{\boldsymbol{S}}_1(y;p)$, it follows from lemma F10 that

$$\hat{\boldsymbol{H}}(y,\hat{p})\simeq_\rho\frac{S_0(y;p)S_0(\infty;p)+v^*\{\hat{\boldsymbol{S}}_1(y;p)S_0(\infty;p)-\hat{\boldsymbol{S}}_1(\infty;p)S_0(y;p)\}}{S_0^2(\infty;p)}. \tag{71}$$

Claim (70) then follows by subtracting and adding $S_1(y;p)$ and $S_1(\infty;p)$ in the numerator of (71). □

Appendix G. Weak Convergence

Let $\hat{\boldsymbol{S}}_s(y)=\hat{\boldsymbol{S}}_s(y;p)$. We first show the weak convergence of $\hat{\mathbb{G}}^*_{ns}(\cdot)=\sqrt{nh_s^{2s+1+d_a}}\{\hat{\boldsymbol{S}}_s(\cdot)-\mathbb{E}\hat{\boldsymbol{S}}_s(\cdot)\}$ in $\mathcal{L}^\infty(\mathcal{I})$, where $\mathcal{I}$ is an arbitrary compact subset of $\mathbb{R}$. Let $\omega_{nsc}(x,y,z,p)=w(z)\mathbb{1}(x=0)\mathbb{1}(y\le c)K\{(a^*-a)/h_s\}k^{(s)}\{(v^*-p)/h_s\}/\sqrt{h_s^{1+d_a}}$ and consider

$$\mathcal{F}_{ns}=\mathcal{F}_{ns}(\mathcal{I})=\{(x,y,z,p)\mapsto\omega_{nsc}(x,y,z,p):c\in\mathcal{I}\}.$$

Define $\mathcal{E}_{ns}$ by $\mathcal{E}_{ns}(x,y,z,p) = \left|K\{(a^*-a)/h_s\}k^{(s)}\{(v^*-p)/h_s\}\right|/\sqrt{h_s^{1+d_a}}$ so that it is an envelope function of $\mathcal{F}_{ns}$. Below we will write $\mathcal{E}_{ns}(a,p)$ for $\mathcal{E}_{ns}(x,y,z,p)$ given that $\mathcal{E}_{ns}(x,y,z,p)$ depends only on a, p.

Lemma G1. *For $s = 0, 1$, $\mathbb{E}\mathcal{E}_{ns}^2(\boldsymbol{a}_i, \boldsymbol{p}_i) \preceq 1$. Also, for any $\epsilon > 0$, $\mathbb{E}\big[\mathcal{E}_{ns}^2(\boldsymbol{a}_i, \boldsymbol{p}_i)\mathbb{1}\{\mathcal{E}_{ns}(\boldsymbol{a}_i, \boldsymbol{p}_i) > \epsilon\sqrt{n}\}\big] \prec 1$.*

Proof. The first statement follows from a change of variables and assumption R. The second statement follows from $\mathbb{1}\{\mathcal{E}_{ns}(a,p) > \epsilon\sqrt{n}\} \le \mathbb{1}\{\sup_{t_1,t_2}|K(t_1)k^{(s)}(t_2)| > \epsilon\sqrt{nh_s^{1+d_a}}\} = 0$ for sufficiently large n by assumption S. □

Lemma G2. *For any $\delta_n \prec 1$ and $s = 0, 1$,*

$$\sup_{|c-c^*|\le\delta_n} \mathbb{E}\Big[\boldsymbol{I}_{xi}\{\mathbb{1}(\boldsymbol{y}_i \le c) - \mathbb{1}(\boldsymbol{y}_i \le c^*)\}K\Big(\frac{a^*-\boldsymbol{a}_i}{h_s}\Big)k^{(s)}\Big(\frac{v^*-\boldsymbol{p}_i}{h_s}\Big)\Big]^2/h_s^{1+d_a} \prec 1.$$

Proof. The LHS of the lemma statement is bounded by twice of

$$\sup_{|c-c^*|\le\delta_n} \mathbb{E}\Big[\mathbb{1}\{\min(c,c^*) < \boldsymbol{y}_i \le \max(c,c^*)\}w(\boldsymbol{z}_i)K\Big(\frac{a^*-\boldsymbol{a}_i}{h_s}\Big)k^{(s)}\Big(\frac{v^*-\boldsymbol{p}_i}{h_s}\Big)\Big]^2/h_s^{1+d_a}$$
$$\preceq C\delta_n \sup_{y,z,p} f_{yzp}(y,z,p) \prec 1,$$

where C is a constant and f_{yzp} is the density of $\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{p}_i$. □

Lemma G3. *For $s = 0, 1$, $\mathcal{F}_{ns}$ is a Vapnik–Cervonenkis (VC) class with VC index uniformly bounded in n.*

Proof. Let $\mathcal{J} = \{y \mapsto \mathbb{1}(y \le c) : c \in \mathcal{I}\}$ and let $\chi_{ns}(x,z,p) = \mathbb{1}(x = 0)w(z)K\{(a^*-a)/h_s\}k^{(s)}\{(v^*-p)/h_s\}/\sqrt{h_s^{1+d_a}}$. Then, by van der Vaart and Wellner (1996, lemma 2.6.18), the VC index of $\mathcal{F}_{ns} = \chi_{ns}\cdot\mathcal{J} = \{\chi_{ns}\bar{J} : \bar{J} \in \mathcal{J}\}$ is bounded by the VC index of $\mathcal{J}$ times 2 minus 1. Therefore, the VC index of $\mathcal{F}_{ns}$ is bounded and independent of n, because $\mathcal{J}$ is a VC class that does not depend on n. □

Lemma G4. *For $s = 0, 1$, $\hat{\mathbb{G}}_{ns}^* \xrightarrow{w} \mathbb{G}_s^*$ in $\mathcal{L}^\infty(\mathbb{R})$, where $\mathbb{G}_s^*$ is a mean–zero Gaussian process.*

Proof. Convergence of finite marginals easily follows by a central limit theorem. Now, for $\ell = 1, 2$, let $\mathcal{F}_{ns,\delta}^\ell$ be a set defined by

$$\Big\{(x,y,z,p) \mapsto \{\omega_{nsc}(x,y,z,p) - \omega_{nsc^*}(x,y,z,p)\}^\ell : |c-c^*| < \delta, \omega_{nsc}, \omega_{nsc^*} \in \mathcal{F}_{ns}(\mathcal{I})\Big\}$$

Since $\{\mathbb{1}(z \in \mathcal{Z}, x = 0, y \leq c) - \mathbb{1}(z \in \mathcal{Z}, x = 0, y \leq c^*)\}^\ell$ is left– or right–continuous for every c, c^* and since $\mathcal{I}$ is separable, $\mathcal{F}^\ell_{ns,\delta}$ contains a countable subclass $\mathcal{G}^\ell_{ns,\delta}$ such that for every $\chi \in \mathcal{F}^\ell_{ns,\delta}$ there exists a sequence $\{\chi_j\} \subset \mathcal{G}^\ell_{ns,\delta}$ with $\chi_j(x, y, z, p) \to \chi(x, y, z, p)$. Therefore, by the same reasoning as van der Vaart and Wellner (1996, example 2.3.4), $\mathcal{F}^\ell_{ns,\delta}$ for $\ell = 1, 2$ is a measurable class for every $\delta > 0$. Therefore, it follows from lemmas G1 to G3 and van der Vaart and Wellner (1996, theorem 2.11.22) that $\hat{\mathbb{G}}^*_{ns} \xrightarrow{w} \mathbb{G}^*_s$ in $\mathcal{L}^\infty(\mathcal{I})$. Since $\mathcal{I}$ is an arbitrary compact set in $\mathbb{R}$, we know by van der Vaart and Wellner (1996, theorem 1.6.1) that $\hat{\mathbb{G}}^*_{ns} \xrightarrow{w} \mathbb{G}^*_s$ in $\mathcal{L}^\infty(\mathcal{I}_1, \mathcal{I}_2, \cdots)$, where $\{\mathcal{I}_j\}$ is an increasing sequence of compact sets such that $\cup_j \mathcal{I}_j = \mathbb{R}$. Finally note that for all n, $\hat{\mathbb{G}}^*_{ns}, \mathbb{G}^*_s \in \mathcal{L}^\infty(\mathbb{R}) \subset \mathcal{L}^\infty(\mathcal{I}_1, \mathcal{I}_2, \cdots)$. □

Lemma G5. *$\hat{\mathbb{G}}^*_{n1}(\cdot) - \hat{\mathbb{G}}^*_{n1}(\infty) G(\cdot|a^*, v^*) \xrightarrow{w} \mathbb{G}^*(\cdot)$ in $\mathcal{L}^\infty(\mathbb{R})$, where $\mathbb{G}^*$ is a mean–zero Gaussian process with the covariance kernel given by $\mathcal{C}$ in (20).*

Proof. By a central limit theorem, $\hat{\mathbb{G}}^*_{n1}(\infty) G(y|a^*, v^*) \xrightarrow{d} \boldsymbol{\Psi} G(y|a^*, v^*)$ for a mean–zero normal random variable $\boldsymbol{\Psi}$. Since $G(\cdot|a^*, v^*)$ is uniformly continuous in $\mathcal{I}$, where $\mathcal{I}$ is an arbitrary compact set in $\mathbb{R}$, we have $\hat{\mathbb{G}}^*_{n1}(\infty) G(\cdot|a^*, v^*) \xrightarrow{w} \boldsymbol{\Psi} G(\cdot|a^*, v^*)$ in $\mathcal{L}^\infty(\mathcal{I})$. Therefore, by van der Vaart and Wellner (1996, theorem 1.6.1), we have $\hat{\mathbb{G}}^*_{n1}(\infty) G(\cdot|a^*, v^*) \xrightarrow{w} \boldsymbol{\Psi} G(\cdot|a^*, v^*)$ in $\mathcal{L}^\infty(\mathcal{I}_1, \mathcal{I}_2, \cdots)$, where $\{\mathcal{I}_j\}$ is an increasing sequence of compact sets such that $\cup_j \mathcal{I}_j = \mathbb{R}$. Now note that for all n, $\hat{\mathbb{G}}^*_{n1}(\infty) G(\cdot|a^*, v^*)$ and $\boldsymbol{\Psi} G(\cdot|a^*, v^*)$ are in $\mathcal{L}^\infty(\mathbb{R}) \subset \mathcal{L}^\infty(\mathcal{I}_1, \mathcal{I}_2, \cdots)$ and the lemma statement follows from the continuous mapping theorem. □

Lemma G6. *For $s = 0, 1$,*

$$\sup_{y \in \mathcal{Y}} \left| \mathbb{E}\{\boldsymbol{K}_{ais} \boldsymbol{k}^{(s)}_{is} \boldsymbol{I}_i(y)\} - \partial^s_v S_0(y) - \frac{h_s^2 \kappa_2}{2} \operatorname{tr}\{\partial^s_v \partial_{bb^\intercal} S_0(y; a^*, v^*)\} \right| \prec \frac{1}{\rho_n}.$$

Proof. This is nothing but a standard kernel bias expansion after noting that $\mathbb{E}\{\boldsymbol{I}_i(y) | \boldsymbol{a}_i = a, \boldsymbol{p}_i = p\} = S_0(y; a, p)/f_{ap}(a, p)$. □

Appendix H. Semiparametric Estimation

Let $\hat{\tilde{\boldsymbol{p}}}(z) = \hat{\tilde{\boldsymbol{r}}}_L(z)/\hat{\tilde{\boldsymbol{f}}}_L(z)$ and redefine $\hat{\boldsymbol{p}}(z) = \hat{\boldsymbol{r}}_L(z)/\hat{\boldsymbol{f}}_L(z)$ (compare with (15)), where

$$\begin{cases} \hat{\tilde{\boldsymbol{f}}}_L(z) = n^{-1} \sum_{i=1}^n \hat{\boldsymbol{K}}_{zLi}, & \hat{\tilde{\boldsymbol{r}}}_L(z) = n^{-1} \sum_{i=1}^n \hat{\boldsymbol{K}}_{zLi} \mathbb{1}(\boldsymbol{x}_i = 0), \\ \hat{\boldsymbol{f}}_L(z) = n^{-1} \sum_{i=1}^n \boldsymbol{K}_{zLi}, & \hat{\boldsymbol{r}}_L(z) = n^{-1} \sum_{i=1}^n \boldsymbol{K}_{zLi} \mathbb{1}(\boldsymbol{x}_i = 0), \end{cases}$$

with $\hat{\boldsymbol{K}}_{zLi} = \hat{\boldsymbol{K}}_{zLi}(z) = K\{(z-\boldsymbol{z}_i)^\intercal \hat{\boldsymbol{\gamma}}/h_z\}/h_z$ and $\boldsymbol{K}_{zLi} = K\{(z-\boldsymbol{z}_i)^\intercal \gamma_0/h_z\}/h_z$. $\hat{\boldsymbol{K}}_{aLi}$ and $\boldsymbol{K}_{aLi}$ are similarly defined.

Lemma H1. *Let*

$$\tilde{\mu}(z) = \frac{1}{\gamma_{01}^2 f_L(z)} \int \partial_{z_1} p\Big(\frac{z^\intercal\gamma_0 - \tilde{t}^\intercal\tilde{\gamma}_0}{\gamma_{01}}, \tilde{t}\Big) f\Big(\frac{z^\intercal\gamma_0 - \tilde{t}^\intercal\tilde{\gamma}_0}{\gamma_{01}}, \tilde{t}\Big)\Big(\frac{z^\intercal\gamma_0 - \tilde{t}^\intercal\tilde{\gamma}_0}{\gamma_{01}}, \tilde{t}\Big) d\tilde{t}$$

Then

$$\sup_{z\in\mathcal{Z}} \Big| \hat{\hat{\boldsymbol{p}}}(z) - \hat{\boldsymbol{p}}(z) - \tilde{\mu}^\intercal(z)(\hat{\boldsymbol{\gamma}} - \gamma_0) \Big| \preceq \frac{1}{\pi_n^* \sqrt{n}},$$

for some π_n^* increasing as fractional power of n.

Proof. Note that both $\hat{\hat{\boldsymbol{f}}}_L(z) - \hat{\boldsymbol{f}}_L(z)$ and $\hat{\hat{\boldsymbol{r}}}_L(z) - \hat{\boldsymbol{r}}_L(z)$ can be expanded as

$$\sum_{j=1}^{J-1} \frac{1}{nj!} \sum_{i=1}^{n} \boldsymbol{K}_{zLi}^{(j)} \boldsymbol{\xi}_i \{(z-\boldsymbol{z}_i)^\intercal(\hat{\boldsymbol{\gamma}} - \gamma_0)\}^j + \frac{1}{nJ!h_z^{J+1}} \sum_{i=1}^{n} K_{zLi}^{(J)}(\cdot)\boldsymbol{\xi}_i \{(z-\boldsymbol{z}_i)^\intercal(\hat{\boldsymbol{\gamma}} - \gamma_0)\}^J, \quad (72)$$

for bounded $\boldsymbol{\xi}_i$.

The J–th order term in (72) is of order $n^{-J/2}h_z^{-J-1}$ (uniformly in $z \in \mathcal{Z}$) which, for sufficiently large J, is $\prec 1/\sqrt{n}$. For $1 \le j \le J-1$, note that for any $1 \le \ell_1, \ldots, \ell_j \le d_z$,

$$\sup_{z\in\mathcal{Z}} \Big| \mathbb{E}\Big\{ \boldsymbol{K}_{zLi}^{(j)} \boldsymbol{\xi}_i \prod_{t=1}^{j} (z_{\ell_t} - \boldsymbol{z}_{i\ell_t}) \Big\} \Big| \preceq \frac{1}{h_z^{\max(0,j-2)}}, \qquad \sup_{z\in\mathcal{Z}} \Big| \mathbb{V}\Big\{ \boldsymbol{K}_{zLi}^{(j)} \boldsymbol{\xi}_i \prod_{t=1}^{j} (z_{\ell_t} - \boldsymbol{z}_{i\ell_t}) \Big\} \Big| \preceq \frac{1}{h_z^{2j+1}}.$$

Consequently, analogous to lemma E2, the j–th term in (72) is for $j \ge 2$ of order no greater than

$$\frac{1}{n^{j/2}} \max\left(\frac{1}{h_z^{\max(0,j-2)}}, \frac{\log n}{\sqrt{n h_z^{2j+1}}} \right) \prec \frac{1}{\sqrt{n}},$$

uniformly in $z \in \mathcal{Z}$.

Finally for $j = 1$ note that by standard kernel theory for any function $\mu \in \mathcal{F}_1$,

$$\sup_{z\in\mathcal{Z}} \Big| \frac{1}{h_z^2} \int k'\Big(\frac{(z-t)^\intercal\gamma_0}{h_z}\Big)\mu(t)dt - \frac{1}{\gamma_{01}^2} \int \partial_{z_1}\mu\Big(\frac{z^\intercal\gamma_0 - \tilde{t}^\intercal\tilde{\gamma}_0}{\gamma_{01}}, \tilde{t}\Big) d\tilde{t} \Big| \preceq h_z^2.$$

Hence it follows that as in lemma F4

$$\sup_{z\in\mathcal{Z}}\left|\hat{\hat{\boldsymbol{p}}}(z)-\hat{\boldsymbol{p}}(z)-\tilde{\mu}^{\mathsf{T}}(z)(\hat{\boldsymbol{\gamma}}-\gamma_0)\right|\simeq$$

$$\sup_{z\in\mathcal{Z}}\left|\frac{\{\hat{\hat{\boldsymbol{r}}}_L(z)-\hat{\boldsymbol{r}}_L(z)\}-p(z)\{\hat{\hat{\boldsymbol{f}}}_L(z)-\hat{\boldsymbol{f}}_L(z)\}}{f_L(z)}-\tilde{\mu}^{\mathsf{T}}(z)(\hat{\boldsymbol{\gamma}}-\gamma_0)\right|\preceq\frac{1}{\pi_n^*\sqrt{n}},$$

for some π_n^* increasing as a fractional power of n. □

Lemma H2. *Let $\boldsymbol{\xi}_i(y)$ be of the form $\zeta^*(\boldsymbol{z}_i)\boldsymbol{I}_i(y)+\zeta^{**}(\boldsymbol{z}_i)\mathcal{E}(y)$ with ζ^*,ζ^{**} depending only on $\boldsymbol{z}_i$ and be such that $\mathbb{E}\{\boldsymbol{\xi}_i(y)|\boldsymbol{z}_i=z\}\in\mathcal{F}_2$. Then for $s=0,1$,*

$$\sup_{y\in\mathcal{Y}}\left|\frac{1}{n}\sum_{i=1}^{n}\big(\hat{\hat{\boldsymbol{k}}}_{is}^{(s)}\hat{\boldsymbol{K}}_{aLi}-\hat{\boldsymbol{k}}_{is}^{(s)}\boldsymbol{K}_{aLi}\big)\boldsymbol{\xi}_i(y)\right|\prec 1/\rho_n.$$

Proof. We have to deal both with the presence of $\hat{\hat{\boldsymbol{p}}}_i$ in lieu of $\hat{\boldsymbol{p}}_i$ and with $\hat{\boldsymbol{\theta}}^{\mathsf{T}}\boldsymbol{a}_i$ in lieu of $\theta_0^{\mathsf{T}}\boldsymbol{a}_i$. Since the former is more difficult than the latter, we shall establish below that

$$\sup_{y\in\mathcal{Y}}\left|\frac{1}{n}\sum_{i=1}^{n}\big(\hat{\hat{\boldsymbol{k}}}_{is}^{(s)}-\hat{\boldsymbol{k}}_{is}^{(s)}\big)\boldsymbol{K}_{aLi}\boldsymbol{\xi}_i(y)\right|\prec 1/\rho_n,\tag{73}$$

where the remaining results can be established similarly but more simply. We again use lemma B3 repeatedly. By the mean value theorem, the LHS average in (73) can be expanded as

$$\sum_{j=1}^{J-1}\frac{1}{nj!}\sum_{i=1}^{n}\hat{\boldsymbol{k}}_{is}^{(s+j)}\boldsymbol{K}_{aLi}\boldsymbol{\xi}_i(y)\big(\hat{\hat{\boldsymbol{p}}}_i-\hat{\boldsymbol{p}}_i\big)^j$$

$$+\frac{1}{nJ!h_s^{s+1+J}}\sum_{i=1}^{n}k^{(s+J)}(\cdot)\boldsymbol{K}_{aLi}\boldsymbol{\xi}_i(y)\big(\hat{\hat{\boldsymbol{p}}}_i-\hat{\boldsymbol{p}}_i\big)^J.\tag{74}$$

The last term in (74) is of order $h_s^{-s-1-J}n^{-J/2}$ (uniformly in y) which, for sufficiently large J, is $\prec 1/\rho_n$. Further, for $1\le j<J$, we expand $\hat{\boldsymbol{k}}_{is}^{(s+j)}$ around $(v^*-\boldsymbol{p}_i)/h_s$ to obtain

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{k}}_{is}^{(s+j)}\boldsymbol{K}_{aLi}\boldsymbol{\xi}_i(y)\big(\hat{\hat{\boldsymbol{p}}}_i-\hat{\boldsymbol{p}}_i\big)^j=\sum_{j^*=0}^{J^*-1}\frac{1}{nj^*!}\sum_{i=1}^{n}\boldsymbol{k}_{is}^{(s+j+j^*)}\boldsymbol{K}_{aLi}\boldsymbol{\xi}_i(y)\big(\hat{\hat{\boldsymbol{p}}}_i-\hat{\boldsymbol{p}}_i\big)^j\big(\hat{\boldsymbol{p}}_i-\boldsymbol{p}_i\big)^{j^*}$$

$$+\frac{1}{nJ^*!h_s^{s+1+j+J^*}}\sum_{i=1}^{n}k^{(s+j+J^*)}(\cdot)\boldsymbol{K}_{ai}\boldsymbol{\xi}_i(y)\big(\hat{\hat{\boldsymbol{p}}}_i-\hat{\boldsymbol{p}}_i\big)^j\big(\hat{\boldsymbol{p}}_i-\boldsymbol{p}_i\big)^{J^*}\tag{75}$$

The last term in (75) is $\prec 1/\rho_n$ (uniformly in y) for sufficiently large J^* because α_n/h_s vanishes as a (fractional) power of n. For $j+j^*>1$ the terms in the RHS sum in (75) are of order $h_s^{-s-j-j^*}n^{-j/2}\alpha_n^{j^*}\prec h_s^{-s-j}n^{-j/2}\preceq 1/\rho_n$.

For $j = 0$ the expansion in (75) is redundant, which leaves the case $j = 1$, $j^* = 0$. Thus, we must analyze

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{k}_{is}^{(s+1)} \boldsymbol{K}_{aLi} \boldsymbol{\xi}_i(y)\big(\hat{\hat{\boldsymbol{p}}}_i - \hat{\boldsymbol{p}}_i\big),$$

which by lemma H1 and standard kernel estimation arguments equals

$$\frac{1}{n^{3/2}}\sum_{i=1}^{n} \boldsymbol{k}_{is}^{(s+1)} \boldsymbol{K}_{aLi} \boldsymbol{\xi}_i(y)\tilde{\mu}(\boldsymbol{z}_i) \tag{76}$$

plus a term of order $1/\rho_n\pi_n^* \prec 1/\rho_n$ (uniformly in y). Finally, (76) is of order $1/\sqrt{n}$, uniformly in y. □

Appendix I. Proofs of Theorems

Proof of Theorem 1. Part (i) follows from lemma A2 and assumption E. For (ii), please recall that the LHS in (6) was shown to be identified in JPX11. □

Proof of Theorem 2. It follows from lemma A1 and assumption I that $g\{x^*, \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|V_t)\} \to \psi^*$. Identification of $g\{x^*, \mathbb{Q}_{\boldsymbol{u}|\boldsymbol{v}}(\tau^*|V_t)\}$ follows from the fact that $V_t \in \mathcal{D}(x^*)$. □

Proof of Theorem 3. It follows from (10) and the monotonicity of g. □

Proof of Theorem 4. By lemma F11,

$$\begin{aligned}
&\rho_n\big\{\hat{\hat{\boldsymbol{H}}}(y|a^*, v^*) - H(y|a^*, v^*)\big\} \\
&\simeq \rho_n v^* \frac{\big\{\hat{\boldsymbol{S}}_1(y;p) - S_1(y;p)\big\} - \big\{\hat{\boldsymbol{S}}_1(\infty;p) - S_1(\infty;p)\big\}G(y;a^*,v^*)}{S_0(\infty;p)} \\
&= \frac{v^*}{S_{0x}}\Big[\big\{\hat{\mathbb{G}}_{n1}^*(y) - \hat{\mathbb{G}}_{n1}^*(\infty)G(y|a^*,v^*)\big\} \\
&\qquad + \big\{\mathbb{E}\hat{\boldsymbol{S}}_1(y;p) - S_1(y;p)\big\} - \big\{\mathbb{E}\hat{\boldsymbol{S}}_1(\infty;p) - S_1(\infty;p)\big\}G(y|a^*,v^*)\Big].
\end{aligned}$$

The stated result then follows from lemmas G5 and G6. □

Proof of Theorem 5. Let $\mathcal{D}$ be a collection of CADLAG functions and define a mapping $T : \mathcal{D} \times \mathcal{U} \to \mathbb{R}$ such that for $F^* \in \mathcal{D}$ and $\tau \in \mathcal{U}$, $T(F^*, \tau) = \inf\{\tilde{\psi} \in \mathcal{Y} : \ F^*(\tilde{\psi}) \geq \tau\}$. We then have $\hat{\boldsymbol{\psi}}^* = T\big\{\hat{\hat{\boldsymbol{H}}}(\cdot|a^*, v^*), \tau^*\big\}$ and $\psi^* = T\big\{H(\cdot|a^*, v^*), \tau^*\big\}$. We now use the functional delta–method; see e.g. Van der Vaart (2000, theorem 20.8). In particular, by Van der Vaart (2000, lemma 21.3), $T(\cdot, \tau^*)$ is Hadamard–differentiable at $H(\cdot|a^*, v^*)$ tangentially to the set of functions $F \in \mathcal{D}$

that are continuous at ψ^* with derivative $T_H(F,\tau^*) = -F(\psi^*)/H'(\psi^*|a^*,v^*)$. Therefore, by the functional delta–method and theorem 4, we have

$$\rho_n\big[T\big\{\hat{\hat{\boldsymbol{H}}}(\cdot|a^*,v^*),\tau^*\big\} - T\big\{H(\cdot|a^*,v^*),\tau^*\big\}\big] \xrightarrow{d} T_H(\mathbb{G},\tau^*) = -\frac{\mathbb{G}(\psi^*)}{H'(\psi^*|a^*,v^*)}. \qquad \square$$

Proof of Theorem 6. First consider $\hat{\boldsymbol{\mathcal{C}}}$. We have

$$\begin{aligned}\hat{\boldsymbol{\mathcal{C}}}(y,y) = \frac{1}{nh_1^{1+d_a}}\sum_{i=1}^n K^2\Big(\frac{a^*-\boldsymbol{a}_i}{h_1}\Big)k^2\Big(\frac{v^*-\hat{\boldsymbol{p}}_{\boldsymbol{i}}}{h_1}\Big)\boldsymbol{I}_{xi}\boldsymbol{I}_i(y) \\ - \frac{2}{nh_1^{1+d_a}}\sum_{i=1}^n K^2\Big(\frac{a^*-\boldsymbol{a}_i}{h_1}\Big)k^2\Big(\frac{v^*-\hat{\boldsymbol{p}}_{\boldsymbol{i}}}{h_1}\Big)\boldsymbol{I}_{xi}\boldsymbol{I}_i(y)\frac{\hat{\boldsymbol{S}}_0(y)}{\hat{\boldsymbol{S}}_{0x}} \\ + \frac{1}{nh_1^{1+d_a}}\sum_{i=1}^n K^2\Big(\frac{a^*-\boldsymbol{a}_i}{h_1}\Big)k^2\Big(\frac{v^*-\hat{\boldsymbol{p}}_{\boldsymbol{i}}}{h_1}\Big)\mathbb{1}(\boldsymbol{x}_i=0)\boldsymbol{w}_i^2\frac{\hat{\boldsymbol{S}}_0^2(y)}{\hat{\boldsymbol{S}}_{0x}^2}. \quad (77)\end{aligned}$$

By lemma F9 and as in the proof of theorem 4, we have $\sup_{y\in\mathcal{Y}}|\hat{\boldsymbol{S}}_0(y)/\hat{\boldsymbol{S}}_{0x} - S_0(y)/S_{0x}| \prec 1$ and

$$\begin{aligned}\sup_{y\in\mathcal{Y}}\Big|\frac{1}{nh_1^{1+d_a}}\sum_{i=1}^n K^2\Big(\frac{a^*-\boldsymbol{a}_i}{h_1}\Big)k^2\Big(\frac{v^*-\hat{\boldsymbol{p}}_{\boldsymbol{i}}}{h_1}\Big)\boldsymbol{I}_{xi}\boldsymbol{I}_i(y) \\ - \frac{1}{nh_1^{1+d_a}}\sum_{i=1}^n K^2\Big(\frac{a^*-\boldsymbol{a}_i}{h_1}\Big)k^2\Big(\frac{v^*-\boldsymbol{p}_i}{h_1}\Big)\boldsymbol{I}_{xi}\boldsymbol{I}_i(y)\Big| \prec 1.\end{aligned}$$

Therefore, $\hat{\boldsymbol{p}}_{\boldsymbol{i}}, \hat{\boldsymbol{S}}_0(y), \hat{\boldsymbol{S}}_{0x}$ in (77) can be replaced with $\boldsymbol{p}_i, S_0(y), S_{0x}$ without changing the (uniform) probability limit of $\hat{\boldsymbol{\mathcal{C}}}$. Then, standard kernel estimation theory the uniform consistency of $\hat{\boldsymbol{\mathcal{C}}}$.

For $\hat{\boldsymbol{S}}_{\boldsymbol{s}}^{(\mathbf{1})}$, let $\bar{\boldsymbol{S}}_{\boldsymbol{s}}^{(\mathbf{1})}$ be defined as $\hat{\boldsymbol{S}}_{\boldsymbol{s}}^{(\mathbf{1})}$ with $\hat{\boldsymbol{p}}_{\boldsymbol{i}}$ replaced with $\boldsymbol{p}_i$. Noting that $\sup_{y\in\mathcal{Y}}|k\{(y-\boldsymbol{y}_i)/h_y\}|/h_y \le C/h_y$ for some C, a slight modification of lemma F9 shows that

$$\sup_{y\in\mathcal{Y}}\big|\hat{\boldsymbol{S}}_{\boldsymbol{s}}^{(\mathbf{1})}(y;a^*,v^*) - \bar{\boldsymbol{S}}_{\boldsymbol{s}}^{(\mathbf{1})}(y;a^*,v^*)\big| \prec \frac{1}{\rho_n h_y} \prec 1,$$

using $\rho_n h_y \succ 1$. Then, standard kernel estimation theory shows the uniform consistency of $\bar{\boldsymbol{S}}_{\boldsymbol{s}}^{(\mathbf{1})}(\cdot;a^*,v^*)$.

$\square$

Proof of Theorem 7. Redefine $\hat{\hat{\boldsymbol{H}}}(\cdot|a^*, v^*)$ in (17) as using $\boldsymbol{a}_i^\intercal \gamma_0$ and $\boldsymbol{z}_i^\intercal \theta_0$ in lieu of $\boldsymbol{a}_i$ and $\boldsymbol{z}_i$. Let $\hat{\hat{\boldsymbol{H}}}(\cdot|a^*, v^*)$ be identically defined but using $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\theta}}$ instead of γ_0 and θ_0. It suffices to show that

$$\sup_{y \in \mathcal{Y}} |\hat{\hat{\boldsymbol{H}}}(y|a^*, v^*) - \hat{\hat{\boldsymbol{H}}}(y|a^*, v^*)| = o_p(1/\rho_n),$$

which follows from lemma H2. □