Proxy Controls and Panel Data*

Ben Deaner†

12/29/19

Abstract

We present a novel approach to nonparametric identification and consistent estimation in economic models using 'proxy controls'. Our approach is particularly well-suited to the context of panel data with a fixed time-dimension but also applies in cross-sectional settings. Proxy controls are proxies for unobserved 'perfect controls', where perfect controls are variables that are sufficient for the association between potential outcomes and treatments. Our identification strategy requires that the set of available proxy controls be split into two subsets, one subset acting as an instrument for the other. In the panel case, observations from different periods can be used as proxy controls and our key identifying assumptions follow from restrictions on the serial dependence of the data and confounding variables. We provide conditions under which our estimation problem is 'well-posed'. Our estimator is straight-forward to implement, the key step is penalized sieve minimum distance estimation. We derive simple convergence rates under high-level assumptions.

Like referee reports and ordinary conversations, confounding factors are frightening but unavoidable challenges for the empirical economist. The threat of confounding is familiar to quantitative researchers in all fields, but it is of particular concern to economists, who are interested almost exclusively in causal inference and whose data are usually drawn from observational studies. Confounding may be understood in terms of factors that impact both the assignment of treatments (variables in whose causal impact we are interested) and potential outcomes. These factors are often inherently unobservable, they are composed of features like innate ability and socio-economic status. Suppose that controlling for these factors the treatment assignments and potential outcomes are independent, then we say that these factors are a set of 'perfect controls'.

While perfect controls are often unobserved, the researcher may have access to covariates that proxy for the perfect controls. These 'proxy controls' could be a set of test scores in place of a measure of innate ability or some demographic

*An earlier draft of this paper was titled Nonparametric Estimation and Identification in Non-Separable Models Using Panel Data. The earlier version can be found at https://arxiv.org/pdf/1810.00283.pdf.

†PhD candidate, Massachusetts Institute of Technology. Email at bdeaner@mit.edu.

characteristics like self-reported wages and years in education in place of socio-economic status. In the context of panel data, observations from other time periods can be proxies for underlying confounding factors. To illustrate, suppose that innate ability is a confounding factor. Then an individual's innate ability is associated with both the individual's treatment assignments and potential outcomes. It follows that the history of treatment assignments is informative about innate ability. Therefore, the history of treatments may be a good proxy for innate ability.

A naive approach would treat the proxies as if they are perfect controls. For example, one could treat test scores as if they did in fact perfectly measure ability. However, if the proxies mismeasure the perfect controls then controlling for the proxies in the conventional manner would not remove all the confounding. Therefore, the resulting estimates may be asymptotically biased.

A recent literature provides conditions in which causal effects are nonparametrically identified when only proxy controls are available. Notably Miao *et al.* (2018) in the biometrics literature which in turn builds on Kuroki & Pearl (2014). In the economics literature identification with proxy controls is achieved in the context of regression discontinuity in Rokkanen (2015).

Key to these results is the observation that the use of proxy controls can be understood as a measurement error problem. The proxy controls mismeasure a set of latent perfect controls. To account for the measurement error Rokkanen (2015) and Miao *et al.* (2018) assume that the proxy controls can be split into groups that satisfy a set of conditional independence and 'completeness' conditions. The assumptions resemble some of those in Hu & Schennach (2008) which apply for general measurement error (i.e., in treatments rather than just controls). Identification with proxy controls is somewhat more amenable than the problem of measurement error in treatment variables. We are not interested in the causal effect of the perfect controls themselves and so we can weaken some of the assumptions in Hu & Schennach (2008) and provide a simpler, constructive identification of causal objects and an uncomplicated estimation method. Notably, we do not require a normalization like mean- or median-unbiasedness of the mis-measured variables which is required by Hu & Schennach (2008).

In Miao *et al.* (2018), as in our work, the researcher divides the available proxy controls into two groups and, in effect, uses one group of proxy controls to instrument for the other. The validity of this approach does not require that the proxy controls be valid instruments in the standard sense. Instead, the proxy controls must satisfy an exclusion restriction which, loosely speaking, states that the two sets of proxy controls are only related through their mutual association with the unobserved perfect controls. In addition to this assumption one requires that the proxy controls are sufficiently informative (defined in terms of statistical completeness) about the latent perfect controls.

The assumption that the two sets of proxy controls are related only through mutual association with underlying perfect controls must be assessed on a case-by-case basis. If multiple sets of test scores are available one may reasonably assume that the test scores reflect only underlying academic ability and random conditions on the day of the test. Therefore, if academic ability is the rele-

vant perfect control, the scores on the different tests are plausibly independent conditional on this latent factor (this argument is made in Cunha *et al.* (2010)).

In the panel case where past and future observations are used as proxy controls, the assumption can be understood in terms of the serial dependence structure of the data. Suppose that treatment assignment at each period t depends only on innate ability, the treatment assignment at period $t-1$ and some exogenous, serially independent factors. In other words, conditional on innate ability the treatment assignments follow a first-order Markov dependence structure. Let one set of proxy controls be the treatment assignments from periods prior to t and let the other set contain treatments in the periods subsequent to $t+1$. Then the two sets of proxy controls are related only through innate ability and the treatment at period $t+1$, which together are sufficient for the confounding and thus constitute a set of perfect controls. The use of observations from other periods to account for confounding originates in the work of Hausman & Taylor (1981) and is the basis of methods in Holtz-Eakin *et al.* (1988), Arellano & Bond (1991) and others in the linear panel case.

The assumption that each of the two sets of proxy controls are sufficiently informative about the perfect controls is analogous to an instrumental relevance condition. The proxy controls should be relevant instruments for the unobserved perfect controls. This generally places some restrictions on the number of unobserved perfect controls compared to proxy controls analogous to the order condition in linear instrumental variables. In the panel case the number of available observations from different time periods is limited by the panel length. If the proxy controls are observations from periods other than t, then the order condition implies a lower bound on the panel length.

We contribute new results on identification and estimation with proxy controls. With regards to identification, we identify a richer set of counterfactual objects than those identified by Miao *et al.* (2018) under similar assumptions, for example the average effect of treatment on the treated. We provide conditions under which the estimation problem suggested by our identification result is 'well-posed'. The well-posedness of our estimation problem is crucial for deriving simple rates of convergence comparable to those achieved in standard nonparametric regression. We provide a nonparametric estimator that builds on our identification results and analyze its properties. Our estimation method is, to the best of our knowledge novel. A key intermediate step is a Penalized Sieve Minimum Distance procedure of the type analyzed by Chen & Pouzo (2012) and others.

To summarize, our contribution is threefold. We add new results for identification with proxy controls that apply to both the cross-sectional and panel settings. We propose a novel estimation method based on our identification results. We show that in dynamic panel settings our exclusion restrictions follow from conditions on the serial dependence structure of the data.

The paper is structured as follows. In Section 1 we present a general model and define causal objects of interest. We define proxy controls and provide conditions under which our objects of interest are identified, we compare the relationship between our results and those of Rokkanen (2015) and Miao *et al.*

(2018). In Section 2 we present our estimation method and provide conditions for its consistency. In Section 3 we show how our identification results apply in panel settings with a fixed number of time periods. Section 4 concludes.

1 General Model and Identification

Consider the following structural model:

$$Y = y_0(X, U) \tag{1}$$

Y is an observed dependent variable, X is a column vector of observables that represents the levels of assigned treatments, and U is a (potentially infinite-dimensional) vector that represents unobserved heterogeneity. The 'structural function' y_0 is not assumed to be of any particular parametric form.

The model above incorporates both cross-sectional and panel settings. In the panel case the model applies for a particular period t, that is, for a particular cross-sectional slice of the panel data. In Section 3 we consider the panel case exclusively and add time-subscripts to Y, X, U and y_0 to make explicit the time-dependence of the model.

Throughout the discussion it is assumed that the structural function y_0 in (1) captures the causal effect of X on Y. For clarity, we situate our analysis in the potential outcomes framework. If a unit has realization of the heterogeneity U of u, then $y_0(x, u)$ is the 'potential outcome' from treatment level x. That is, the outcome that would have been observed had the treatment of that unit been set to level x. Thus U captures all heterogeneity in the potential outcomes. We assume (without loss of generality) that for any x and any $u_1 \neq u_2$ that $y_0(x, u_1) \neq y_0(x, u_2)$.

We assume throughout that the random pairs (X, U) are independently and identically distributed. The assumption that the distribution is identical across units is not restrictive in this setting because the distribution of the treatment variable X could depend strongly on the unobserved heterogeneity U.

The focus of this paper is on the identification and estimation of conditional average potential outcomes, where we condition on the assigned treatments X and possibly some additional variables S. By incorporating additional covariates S into our analysis we can define counterfactual objects like the average treatment effect for a particular demographic sub-group with $S = s$. The function that returns the conditional average potential outcomes is sometimes referred to as the 'conditional average structural function'.

Conditional average potential outcomes are defined formally as follows. The conditional average potential outcome from treatment level x_1, conditional on treatment assignment X equal to x_2 and additional covariates S equal to s is:

$$\bar{y}(x_1|x_2, s) = E\big[y_0(x_1, U)|X = x_2, S = s\big]$$

In words, suppose we draw a unit at random from the sub-population with additional covariates $S = s$ who were assigned treatment $X = x_2$. Then the

expected counterfactual outcome had the unit instead received treatment level x_1 is $\bar{y}(x_1|x_2, s)$.[1]

Many counterfactual objects of interest can be written in terms of the conditional average potential outcomes. For example, conditional average treatment effects and the average effect of treatment on the treated can both be expressed using the conditional average potential outcomes and the probability distributions of some observables. For a more involved example, consider the average outcome among agents in demographic group $S = s$ had they received treatments ten percent larger than those that were actually assigned. This can be written as:

$$E\big[y_0(1.1X, U)|S = s\big] = \int_{\mathcal{X}} \bar{y}(1.1x|x, s) F_{X|S=s}(dx)$$

Where $\mathcal{X}$ denotes the support of assigned treatments X and $F_{X|S=s}$ is the distribution of X conditional on $S = s$ (we use this notation for conditional probability laws throughout).

By transforming the model one can define an even richer set of counterfactual objects in terms of the conditional average potential outcomes of the transformed model. For example, let y be some fixed scalar and consider the transformation $w \mapsto 1\{w \leq y\}$. Let $\tilde{Y}$ be the transformed outcome variable, that is $\tilde{Y} = 1\{Y \leq y\}$, and let $\tilde{y}_0$ be the transformed structural function, that is $\tilde{y}_0(x, u) = 1\{y_0(x, u) \leq y\}$. The transformed model is:

$$\tilde{Y} = \tilde{y}_0(X, U)$$

The conditional cumulative distribution function of the potential outcomes in the original model can be written as:

$$\begin{aligned} P\big(y_0(x_1, U) \leq y|X = x_2, S = s\big) &= E\big[1\{y_0(x_1, U) \leq y\}|X = x_2, S = s\big] \\ &= E[\tilde{y}_0(x_1, U)|X = x_2, S = s] \end{aligned}$$

The right-hand side of the final equality above is the conditional average structural function for the transformed model. Our identifying assumptions do not refer to Y directly but instead to the latent variable U, and as such our assumptions are invariant to transformations of the kind above. That is, if our assumptions apply for the original model they also apply for the transformed model. Thus if we can identify the conditional average potential outcomes then we can also identify the conditional cumulative distribution function of the potential outcomes. Note that identification of the conditional cumulative distribution implies identification of the conditional quantiles.

Proxy Controls

The identification of the conditional average structural function is challenging when assigned treatments X may be associated with heterogeneity U. A common approach to identification in the presence of confounding relies on the

[1] Note that if X or S is continuously distributed then $\bar{y}(x_1|x_2, s)$ is only uniquely defined for x_2 and s up to a set of $F_{(X,S)}$-measure 1, where $F_{(X,S)}$ is the joint law of X and S.

presence of what we term 'perfect controls'. A vector of perfect controls is an observable random vector W^*, so that conditioning on W^* and the additional covariates S, the treatments X and the heterogeneity in potential outcomes are independent. That is:

$$U \perp X|(W^*, S)$$

Note that we use the notation above to denote conditional independence throughout this paper. Variables W^* with the property above are sometimes referred to simply as 'confounders' but due to the lack of consensus over this term (VanderWeele & Shpitser (2013)) we refer to them exclusively as 'perfect controls'.

Under the conditional independence assumption above and a full support assumption, the conditional average structural function is identified by:

$$\bar{y}(x_1|x_2, s) = \int_{\mathcal{W}^*} E[Y|X = x_1, W^* = w, S = s] F_{W^*|X=x_2,S=s}(dw)$$

Where $\mathcal{W}^*$ is the support of the perfect controls W^* and $F_{W^*|X=x_2,S=s}$ is the conditional probability law of controls W^* given assigned treatments $X = x_2$ and additional covariates $S = s$.

When perfect controls W^* are unavailable the researcher may have access to proxy controls W. W need not be a vector of perfect controls (i.e., $U \not\perp X|(W, S)$), however W may be informative about the perfect controls W^*.

We present assumptions that imply identification when only proxy controls W are available. The assumptions refer to the vector of perfect controls W^* for which W acts as a proxy. Since W^* is unobserved, the assumptions can be understood to state that a vector of latent variables W^* exists that simultaneously satisfies all the conditions in our assumptions. To argue persuasively that the assumptions are plausible in a given setting, a researcher will generally have to choose a particular set of unobserved perfect controls W^* and argue that the assumptions hold for those controls. Our identification results resemble those of Miao *et al.* (2018) and, to a lesser extent, Rokkanen (2015). We provide a detailed comparison later in this section.

As Rokkanen (2015) notes, the problem of identification with proxy controls can be understood as a measurement-error problem. The vector of proxy controls W can be understood as a measurement of W^* that is subject to non-classical (i.e., non-zero mean and non-additive) noise. Like Miao *et al.* (2018), we propose that the researcher split the vector of proxy controls W into two (possibly over-lapping) sub-vectors V and Z. The researcher in effect uses the proxy controls in Z as instruments for the proxy controls in V.

The instruments Z must be valid in that they satisfy an exclusion restriction involving the proxy controls V, the treatments X and the unobserved perfect controls W^*. We emphasize that, unlike in standard instrumental variables analysis, Z implicitly acts as an instrument for W^* and not for the treatments X. As such, Z is not required to be independent of W^*. In fact, Z must satisfy an informativeness assumption that is analogous to an instrumental relevance

condition and this generally precludes that Z and W^* be independent.[2]

We state our first two assumptions below. Assumption 1 is easily satisfied if W is a vector of perfect controls (i.e., $U \perp X|(W,S)$), in that case take the relevant perfect controls to be $W^* = W$, setting $V = Z = W$ the condition holds trivially. If, in addition, W has full support conditional on X and S then Assumption 2 holds with $V = Z = W$.

Assumption 1 (Conditional Independence)
i. $U \perp (X,Z)|(W^*,S)$ ii. $V \perp (X,Z)|(W^*,S)$

Assumption 2 (Informativeness)
i. For $F_{(X,S)}$-almost all (x,s), for any function $\delta \in L_2(F_{W^*|X=x,S=s})$:

$$E\big[E[\delta(W^*)|Z,X,S]^2|X=x,S=s\big]=0 \iff E[\delta(W^*)^2|X=x,S=s]=0$$

ii. For $F_{(X,S)}$-almost all (x,s), for any function $\delta \in L_2(F_{W^*|X=x,S=s})$:

$$E\big[E[\delta(W^*)|V,X,S]^2|X=x,S=s\big]=0 \iff E[\delta(W^*)^2|X=x,S=s]=0$$

Assumption 1 makes two assertions of conditional independence. In words, Assumption 1.i states that the perfect controls W^* and additional covariates S explain all the association between the heterogeneity U on the one hand, and the treatments X and proxy controls Z on the other. This assumption implies that W^* is in fact a perfect control (it implies $U \perp X|(W^*,S)$). Assumption 1.ii states that any dependence between V and (X,Z) is explained by their mutual association with the perfect controls W^* and additional conditioning variables S. We emphasize that the independence between V and (X,Z) in Assumption 1.ii is conditional. Without conditioning on the perfect controls W^* and additional covariates S, V could be strongly associated with both X and Z. Again, note that neither Assumption 1.i nor 1.ii requires either V or Z be independent of W^*.

Assumption 2, loosely speaking, states that both V and Z are sufficiently informative about the unobserved perfect controls W^*. The informativeness condition is in terms of 'completeness', or more precisely, L_2-completeness (Andrews (2017)). Completeness is used to achieve identification in the non-parametric instrumental variables (NPIV) models of Newey & Powell (2003) and Ai & Chen (2003). In the NPIV context, completeness is an instrumental relevance condition analogous to the rank condition for identification in linear IV (see Newey & Powell (2003) for discussion). With this interpretation, 2.i states that conditional on any given value of assigned treatments X and covariates S, Z is a relevant instrument for W^*, and 2.b. states that conditioning on X and S, V is a relevant instrument for W^*. Some sufficient conditions for statistical

[2] More precisely, Z and W^* must not be independent conditional on X and S (apart from in the trivial case of W^* non-random).

completeness can be found in D'Haultfoeuille (2011) and Hu & Shiu (2018). In some settings L_2-completeness is generic in a certain sense (Andrews (2017), Chen *et al.* (2014)).

In the linear IV case, the rank condition can only hold if the number of instruments exceeds the number of exogenous regressors for instrumental relevance to be possible (this is known as the 'order condition'). Such a condition is not, strictly speaking, necessary for L_2-completeness in nonparametric models.[3] However, an order condition is necessary in certain special cases, for example the conditional Gaussian case discussed in Newey & Powell (2003). As such, it would seem prudent to require that V and Z be of a weakly larger dimension than W^*.

Finally, we introduce some regularity conditions. In Theorem 1 we characterize the conditional average structural function as a limit of approximate solutions to a conditional moment restriction. The regularity conditions imply that this limit is well-defined and the estimation problem suggested by our characterization is well-posed.

Some of the perfect controls in W^* may be observed and used as proxy controls in Z, thus the vectors Z and W^* may have some components in common. We denote the shared components by '$\bar{W}$' and let $\tilde{Z}$ and $\tilde{W}^*$ respectively contain the entries of Z and W^* other than those in $\bar{W}$. Thus we can decompose $W^* = (\tilde{W}^*, \bar{W})$ and $Z = (\tilde{Z}, \bar{W})$. For each $(x, s, \bar{w})$ in the support of $(X, S, \bar{W})$ define a linear operator $A_{x,s,\bar{w}} : L_2(F_{\tilde{Z}|X=x,S=s,\bar{W}=\bar{w}}) \to L_2(F_{\tilde{W}^*|X=x,S=s,\bar{W}=\bar{w}})$ by:

$$A_{x,s,\bar{w}}[\delta](\tilde{w}^*) = E[\delta(\tilde{Z})|X = x, S = s, \bar{W} = \bar{w}, \tilde{W}^* = \tilde{w}^*]$$

For $F_{\tilde{W}^*|X=x,S=s,\bar{W}=\bar{w}}$-almost all $\tilde{w}^*$.

Assumption 3

i. The joint distribution of $\tilde{W}^*$, $\tilde{Z}$, $\bar{W}$, X, and S is absolutely continuous with respect to the product of their marginals. ii. For F_X-almost all x_1 and x_2 and F_S-almost all s and $F_{\bar{W}}$-almost all $\bar{w}$:

$$E\left[\frac{dF_{W^*|X=x_2,S=s}}{dF_{W^*|X=x_1,S=s}}(W^*)^2 \middle| X = x_1, S = s, \bar{W} = \bar{w}\right] < \infty$$

Where $\frac{dF_{W^*|X=x_2,S=s}}{dF_{W^*|X=x_1,S=s}}$ is the Radon-Nikodym derivative of $F_{W^*|X=x_2,S=s}$ with respect to $F_{W^*|X=x_1,S=s}$.[4]

iii. The following holds for F_X-almost all x, F_S-almost all s and $F_{\bar{W}}$-almost all $\bar{w}$. Let 'F_{prod}' denote the product measure of $\tilde{W}^*$ and $\tilde{Z}$ conditional on $X = x$, $S = s$ and $\bar{W} = \bar{w}$.[5] The conditional joint measure $F_{(\tilde{W}^*,\tilde{Z})|X=x,S=s,\bar{W}=\bar{w}}$ is absolutely continuous with respect to F_{prod}:

$$\int_{\tilde{\mathcal{W}}^* \times \tilde{\mathcal{Z}}} \left[\frac{dF_{(\tilde{W}^*,\tilde{Z})|X=x,S=s,\bar{W}=\bar{w}}}{dF_{prod}}(\tilde{w}^*, \tilde{z})\right]^2 F_{prod}(d\tilde{w}^*, d\tilde{z}) < \infty$$

[3] The methods of Andrews (2017) can be used to construct L_2-complete distributions even when the number of endogenous regressors exceeds the number of instruments.

[4] The absolute continuity in Assumption 3.i guarantees that the Radon-Nikodym derivative exists.

[5] In more conventional notation F_{prod} is equal to $F_{\tilde{W}^*|X=x,S=s,\bar{W}=\bar{w}} \otimes F_{\tilde{Z}|X=x,S=s,\bar{W}=\bar{w}}$.

Where $\tilde{\mathcal{W}}^*$ is the support of $\tilde{W}^*$ and $\tilde{\mathcal{Z}}$ the support of $\tilde{Z}$.

iv. There exists a finite constant $C > 0$ so that the following holds for F_X-almost all x_1 and x_2 , F_S-almost all s and $F_{\bar{W}}$-almost all $\bar{w}$. Let $\{(u_k, v_k, \mu_k)\}_{k=1}^{\infty}$ be the singular system for $A_{x_1,s,\bar{w}}$. That is, μ_k is the k^{th} singular value of $A_{x_1,s,\bar{w}}$ and $u_k : \tilde{\mathcal{W}}^* \to \mathbb{R}$ and $v_k : \tilde{\mathcal{Z}} \to \mathbb{R}$ are the k^{th} singular functions.[6] Then:

$$\sum_{k=1}^{\infty} \frac{1}{\mu_k^2} E[\frac{dF_{W^*|X=x_2,S=s}}{dF_{W^*|X=x_1,S=s}}(W^*)u_k(\tilde{W}^*)|X = x_1, S = s, \bar{W} = \bar{w}]^2 \leq C$$

The purpose of Assumptions 3.i, 3.ii and 3.iii is to ensure that the expansion in Assumption 3.iv is well-defined. Assumption 3.i ensures that the Radon-Nikodym derivatives in 3.ii and 3.iii exist. Assumption 3.ii states that the Radon-Nikodym derivative $\frac{dF_{W^*|X=x_2,S=s}}{dF_{W^*|X=x_1,S=s}}$ lies in the relevant space of mean square integrable functions. Assumption 3.iii guarantees that $A_{x,s,\bar{w}}$ is Hilbert-Schmidt and thus compact and so the singular system in Assumption 3.iv is well-defined (see Darolles *et al.* (2011) for some discussion).

Assumption 3.iv is crucial to our analysis because it ensures our characterization of the conditional average structural function is well-defined and that estimation of the conditional average structural function is not 'ill-posed'. This allows us to derive simple convergence rates for our estimation method that are comparable to those in standard non-parametric regression. Loosely speaking, Theorem 1 below characterizes the conditional average structural function as a linear functional of a nonparametric instrumental variables (NPIV) regression function. Estimation of an NPIV regression function is generally ill-posed but estimation of a sufficiently smooth linear functional of an NPIV regression function is well-posed.

Assumption 3.iv (combined with Assumption 2), implies that there exists a function φ so that:

$$E[\varphi(x_1, x_2, s, Z)|X = x_1, S = s, W^* = w^*] = \frac{dF_{W^*|X=x_2,S=s}}{dF_{W^*|X=x_1,S=s}}(w^*)$$

And:

$$E\big[\varphi(x_1, x_2, s, Z)^2|X = x_1, S = s\big] \leq C$$

This in turn implies a special case of a condition of Lemma 4.1 from Severini & Tripathi (2012). Severini & Tripathi (2012) and Ichimura & Newey (2017) show that a condition of this kind is closely related to root-n estimability. Deaner (2019) shows the same condition is (under mild additional assumptions) necessary and sufficient for robust estimation of the linear functional.

Theorem 1

Suppose Assumptions 1, 2, and 3 hold. Then the conditional average structural function $E[y_0(x_1, U)|X = x_2, S = s]$ is identified (for (x_1, x_2, s) up to a set of

[6] See, e.g., Kress (2014) Theorem 15.16 and associated discussion.

$F_X^2 \otimes F_S$-measure 1). In particular, there exists a sequence of functions $\{\gamma_k\}_{k=1}^{\infty}$ with $\gamma_k(x,s,\cdot) \in L_2(F_{V|X=x,S=s})$ so that:

$$\lim_{k\to\infty} E\Big[E[Y-\gamma_k(X,S,V)|X,S,Z]^2\Big|X=x,S=s\Big]=0$$

And for any such a sequence:

$$\bar{y}(x_1|x_2,s)=\lim_{k\to\infty} E\big[\gamma_k(x_1,S,V)|X=x_2,S=s\big]$$

In particular, for F_X-almost all x_1 and x_2 and F_S-almost all s:

$$\begin{aligned}&\big(\bar{y}(x_1|x_2,s)-E[\gamma_k(x_1,S,V)|X=x_2,S=s]\big)^2 \qquad (2)\\ \leq& CE\Big[\big(E[Y-\gamma_k(X,S,V)|X,S,Z]\big)^2\Big|X=x_1,S=s\Big]\end{aligned}$$

▲

The characterization of the conditional average structural function in Theorem 1 suggests a two-step approach to estimation. In a first stage, the researcher finds a function $\hat{\gamma}$ that approximately satisfies an empirical analogue of the following moment condition (with parameter γ):

$$E[\gamma(X,S,V)-Y|X=x,S=s,Z=z]=0 \qquad (3)$$

For $F_{(X,S,Z)}$-almost all (x,s,z). Equation (3) is equivalent to a non-parametric instrumental variables (NPIV) moment condition in which V is the vector of endogenous regressors, X and S are vectors of exogenous regressors, and Z is a vector of instruments. Thus we can attain an approximate solution using standard NPIV methods.

Note that Theorem 1 does not state that there exists a γ that satisfies the estimating equation (3) exactly. Instead it asserts that there is a γ that makes the two sides arbitrarily close (in a particular mean squared sense). For estimation this detail is of little consequence: an empirical analogue of the NPIV estimating equation generally has an exact solution.

So suppose $\hat{\gamma}$ solves an empirical analogue of the moment condition (3). In a second step Theorem 1 suggests we estimate the conditional average structural function by:

$$E[y_0(x_1,U)|X=x_2,S=s]\approx \hat{E}\big[\hat{\gamma}(x_1,S,V)|X=x_2,S=s\big]$$

where '$\hat{E}$' denotes some empirical analogue of the conditional expectation. The final assertion of Theorem 1 implies that our estimation problem is not ill-posed. In words, if $\hat{\gamma}$ satisfies the population moment condition (3) with small error, then $E\big[\hat{\gamma}(x_1,S,V)|X=x_2,S=s\big]$ is close to the conditional average potential outcome $\bar{y}(x_1|x_2,s)$. If, in addition, $E\big[\hat{\gamma}(x_1,S,V)|X=x_2,S=s\big]$ is close to the sample analogue $\hat{E}\big[\hat{\gamma}(x_1,S,V)|X=x_2,S=s\big]$, then the latter provides a good estimate of the conditional average potential outcome. This motivates our estimator in the next section.

Relationship to Existing Results

Our results are closely related to those in Miao *et al.* (2018). Suppose there are no additional covariates S. Then our Assumptions 2.i and 2.ii are similar to Conditions 2 and 3 in Miao *et al.* (2018) and in fact if either our Assumption 1 holds or their exclusion restriction (f) holds, then their Conditions 2 and 3 imply our Assumption 2. Our Assumption 1 is equivalent to their exclusion restriction (f) plus the assumption that $U \perp X|W^*$. Our characterization of the conditional average potential outcomes in Theorem 1 somewhat resembles their characterization of $p(y|do(x))$. Note that Miao *et al.* (2018) do not assume that $U \perp X|W^*$. They equate $p(y|do(x))$ with (for a binary Y) the object $E\big[E[Y|X = x, W^*]\big]$. For this to equal the average potential outcome $E[y_0(x, U)]$, one generally requires $U \perp X|W^*$. Therefore it seems the assumption is implicit in their analysis.

Our identification result differs from Miao *et al.* (2018) in key ways. Firstly, our analysis allows us to identify the conditional distribution of potential outcomes (conditional on assigned treatments X and additional characteristics S). This allows us to identify a richer set of counterfactual objects, for example the average effect of treatment on the treated or the policy counterfactual discussed earlier in this section in which units receive treatments a fixed percentage larger than those observed. Furthermore, we provide a regularity condition, Assumption 3, which ensures the well-posedness of estimators based on our characterization of the conditional average structural function. Well-posedness is crucial for deriving transparent convergence rates. Note that our Assumption 3 also replaces Conditions A1, A2 and A3 in Miao *et al.* (2018).

Rokkanen (2015) gives conditions for identification in the setting of regression discontinuity design. Our Assumptions 1.i and 1.ii resemble but are slightly weaker than the analogous Assumptions D.1 and C.2 in Rokkanen (2015). We require only L_2-completeness, compared to the bounded completeness in Assumptions D.2 and C.5 in Rokkanen (2015). Rokkanen (2015) applies the results of Hu & Schennach (2008) and correspondingly his Condition C.4 requires that the mis-measured perfect controls satisfy a normalization like mean- or median-unbiasedness. [7]

2 Estimation

In this section we describe our estimation method. The key step in the procedure corresponds to penalized sieve minimum distance (PSMD) estimation. PSMD estimators and some of their properties are discussed in Chen & Pouzo (2012), Chen & Pouzo (2015), and others. Because the estimation procedure is of the "sieve" type, the practitioner must choose an appropriate sequence of linear sieve spaces.

Let K_n be an increasing sequence of natural numbers. For each n let Φ_n be a

[7] However, we conjecture that a failure of Assumption C.4 does not lead to inconsistency in the estimator suggested by Rokkanen (2015), his work does not consider this possibility.

length-K_n column vector of basis functions defined on the support of (X, S, V). In a first stage the practitioner estimates the vector of regression functions Π_n defined by:

$$\Pi_n(x, s, z) = E[\Phi_n(X, S, V)|X = x, S = s, Z = z]$$

The practitioner also estimates the vector of regression functions α_n defined by:

$$\alpha_n(x_1, x_2, s) = E[\Phi_n(x_1, S, V)|X = x_2, S = s]$$

And finally, the practitioner estimates the function g given by:

$$g(x, s, z) = E[Y|X = x, S = s, Z = z]$$

Denote the estimates of Π_n, α_n and g by $\hat{\Pi}_n$, $\hat{\alpha}_n$ and $\hat{g}$ respectively. The estimation of each of these functions can be carried out using a standard non-parametric regression method like local-linear regression, polynomial-series regression or Nadaraya-Watson. For concreteness, consider the case of series least squares estimation. Let Ψ_n be a column vector of basis functions defined on $\mathcal{X} \times \mathcal{S} \times \mathcal{Z}$. Let χ_n be a column vector of basis functions defined on $\mathcal{X} \times \mathcal{S}$. Define matrices $\hat{Q}_n$ and $\hat{R}_n$ by:

$$\hat{Q}_n = \frac{1}{n} \sum_{i=1}^{n} \Psi_n(X_i, S_i, Z_i) \Psi_n(X_i, S_i, Z_i)'$$

$$\hat{R}_n = \frac{1}{n} \sum_{i=1}^{n} \chi_n(X_i, S_i) \chi_n(X_i, S_i)'$$

Then the sieve estimators for Π_n, α_n and g are given below:

$$\hat{\Pi}_n(x, s, z) = \Psi_n(x, s, z)' \hat{Q}_n^{-1} \frac{1}{n} \sum_{i=1}^{n} \Psi_n(X_i, S_i, Z_i)' \Phi_n(x, s, V_i)$$

$$\hat{\alpha}_n(x_1, x_2, s) = \chi_n(x_2, s)' \hat{R}_n^{-1} \frac{1}{n} \sum_{i=1}^{n} \chi_n(X_i, S_i)' \Phi_n(x_1, s, V_i)$$

$$\hat{g}(x, s, z) = \Psi_n(x, s, z)' \hat{Q}_n^{-1} \frac{1}{n} \sum_{i=1}^{n} \Psi_n(X_i, S_i, Z_i)' Y_i$$

Let P be some penalty function (for example the l_2 penalty). Let λ_n be a positive scalar penalty parameter. In the second stage, the researcher evaluates a vector of coefficients $\hat{\theta} \in \Theta_n \subseteq \mathbb{R}^{K_n}$ that minimize the penalized least squares objective below:

$$\hat{\theta} \in \arg\min_{\theta \in \Theta_n} \left[\frac{1}{n} \sum_{i=1}^{n} \left(\hat{g}(X_i, S_i, Z_i) - \hat{\Pi}_n(X_i, S_i, Z_i)' \theta \right)^2 + \lambda_n P(\theta) \right] \tag{4}$$

The estimate of the conditional average structural function is then given by:

$$\bar{y}(x_1|x_2,s) \approx \hat{\alpha}_n(x_1,x_2,s)'\hat{\theta} \tag{5}$$

Note that for certain choices of the penalty function P and coefficient space Θ_n, the penalized least squares problem that is used to define $\hat{\theta}$ has an analytical solution. This is true, for example, if $\Theta_n = \mathbb{R}^{K_n}$ and the squared l_2 penalty is used (i.e., $P(\theta) = \theta'\theta$). Therefore, depending on the regression method used in the first stage our procedure may not require any kind of numerical optimization.

Consistency and Convergence Rate

Below we provide simple high-level conditions that, when combined with Assumptions 1, 2 and 3 in the previous section, guarantee consistency and a convergence rate for our estimator. We note that the convergence rate does not depend on any 'sieve-measure of ill-posedness' (Chen & Pouzo (2012)). The well-posedness of our problem depends crucially on Assumption 3.

In the assumptions below, 'ess sup' denotes the essential supremum, i.e., the smallest almost sure bound on its argument.

Assumption 4 i. There exists a sequence $\kappa_n \downarrow 0$ so that:

$$\inf_{\theta\in\Theta_n} \operatorname{ess\,sup} E\left[\left(g(X,S,Z) - \Pi_n(X,S,Z)'\theta\right)^2 \middle| X,S\right]^{\frac{1}{2}} = O_p(\kappa_n)$$

ii. There exists a sequence $\eta_n \downarrow 0$ and a constant $c > 0$ so that:

$$\begin{aligned}&\operatorname{ess\,sup} E\left[\left(g(X,S,Z) - \Pi_n(X,S,Z)'\hat{\theta}\right)^2 \middle| X,S\right]^{\frac{1}{2}} \\ &\leq c \inf_{\theta\in\Theta_n} \operatorname{ess\,sup} E\left[\left(g(X,S,Z) - \Pi_n(X,S,Z)'\theta\right)^2 \middle| X,S\right]^{\frac{1}{2}} + O_p(\eta_n)\end{aligned}$$

Where the estimator $\hat{\theta}$ is treated as a constant in the expectation on the LHS above. iii. There exists a sequence $b_n \downarrow 0$ so that uniformly over $x_1, x_2 \in \mathcal{X}$ and $s \in \mathcal{S}$:

$$\left|\left(\hat{\alpha}_n(x_1,x_2,s) - \alpha_n(x_1,x_2,s)\right)'\hat{\theta}\right| \leq O_p(b_n)$$

Theorem 2 below gives a convergence rate for our estimator in terms of the rates given in Assumption 4.

Theorem 2

Suppose Assumptions 1, 2, 3 and 4 hold, then uniformly over F_X-almost all x_1 and x_2 and F_S-almost all s:

$$\left|\bar{y}(x_1|x_2,s) - \hat{\alpha}_n(x_1,x_2,s)'\theta\right| = O_p(\kappa_n + \eta_n + b_n)$$

▲

Panel Models

The analysis in previous sections applies to the model (1) which may apply in both cross-section and panel settings. In the previous sections we are agnostic about the source of the proxy controls W and sub-vectors V and Z. In panel settings, observations from previous and subsequent periods are a natural source of proxy controls. Loosely speaking, if the same factors explain the confounding in each period (there are time-invariant perfect controls), then treatment assignments in other periods are informative about the confounding. Thus we can form vectors V and Z using treatments (and possibly outcomes) from different periods. Then the exclusion restriction in Assumption 1.ii can be understood in terms of the serial dependence of the observables.

In the panel setting, the data have a 'time' dimension and a 'unit' dimension. To apply our analysis in the panel case we rewrite the model (1) with time subscripts:

$$Y_t = y_{0,t}(X_t, U_t)$$

Then for each group there is an associated draw of the random variables $(X_1, ..., X_T, U_1, ..., U_t)$ and a resulting sequence of outcomes $(Y_1, ..., Y_T)$. We assume that the data are iid across groups but not necessarily within groups. More precisely, we assume that draws of $(X_1, ..., X_T, U_1, ..., U_t)$ are independent and identically distributed. However, within each group the random variables (X_t, U_t) may exhibit various forms of serial-dependence.

In the panel setting our goal is to identify and estimate causal objects of the form below for a particular value of t:

$$E[y_{0,t}(x_1, U_t)|X_t = x_2, S = s]$$

The above is the conditional average potential outcome at period t from treatment x_1 conditional on assignment of treatment x_2 at t and additional characteristics s. As we discuss below, under some assumptions regarding the serial dependence structure it may be possible to identify the object above for some values of t and not others.

In this context Assumptions 1.i and 1.ii state that:

$$U_t \perp (X_t, Z)|(W^*, S)$$

$$V \perp (X_t, Z)|(W^*, S)$$

We will form the vectors of proxy controls V and Z using assigned treatments and possibly outcomes from periods other than t. The use of observations from previous and subsequent periods for V and Z is redolent of the use of lagged variables as instruments in Arellano & Bond (1991) and Holtz-Eakin *et al.* (1988) for linear dynamic panel models. The appropriate choices for V and Z depend on the time-dependence structure of the data. A key case is that in which the assigned treatments, and possibly the outcomes, follow a Markov dependence structure.

Markov Treatment Assignments

Fix some period t. Suppose that conditional on some (possibly period t-specific) latent variables $\tilde{W}^*$, the following exclusion restriction holds:

$$U_t \perp (X_1, ..., X_T)|(\tilde{W}^*, S)$$

In words, the assumption above states that any association between the assigned treatments in all periods and heterogeneity in time-t potential outcomes, is explained by some factors $\tilde{W}^*$ and the additional conditioning variables S.

We suppose that conditional on the latent variables $\tilde{W}^*$ and additional conditioning variables S, the regressors satisfy a first-order Markov dependence structure at time $t+1$. Formally:

$$(X_{t+2}, ..., X_T) \perp (X_1, ..., X_t)|(\tilde{W}^*, X_{t+1}, S)$$

That is, conditional on the latent variables $\tilde{W}^*$, additional covariates S, the treatment assignments for period prior to $t+1$ are only only related to treatments after $t+1$ through the treatment at $t+1$. In this case suppose we set V and Z as follows:

$$V = (X_{t+1}, ..., X_T)$$

$$Z = (X_1, ..., X_{t-1}, X_{t+1})$$

Let the perfect controls W^* consist not only of the latent factors $\tilde{W}^*$ but also the treatment assignment at $t+1$. That is:

$$W^* = (\tilde{W}^*, X_{t+1})$$

Note that V, Z and W^* all contain X_{t+1}. With the definitions above, Assumption 1 holds. That is:

$$U_t \perp (X_t, Z)|(W^*, S)$$

$$V \perp (X_t, Z)|(W^*, S)$$

In fact, we can allow for greater order Markov processes. Suppose we make the following k^{th} order Markov assumptions:

$$(X_{t+k+2}, ..., X_T) \perp (X_1, ..., X_t)|(\xi, X_{t+1}, .., X_{t+k+1}, S)$$

Then letting V, Z and W^* be defined as follows, Assumption 1 is satisfied:

$$Z = (X_1, ..., X_{t-1}, X_{t+1}, ..., X_{t+k})$$
$$V = (X_{t+1}, ..., X_T)$$

$$W^* = (\tilde{W}^*, X_{t+1}, ..., X_{t+k})$$

Note that the assumption $U_t \perp X|\tilde{W}^*, S$ is unnecessarily strong. The following weaker, but less intuitive assumption would suffice:

$$U_t \perp (X_1, ..., X_t)|(\tilde{W}^*, X_{t+1}, ..., X_{t+k}, S)$$

Markov Treatment Assignments and Heterogeneity

We now give conditions under which Z and V may be composed not only of treatment assignments from periods other than t, but also the outcomes from other periods. We strengthen the exclusion restriction from the previous subsection:

$$(U_1, ..., U_T) \perp (X_1, ..., X_T)|(\tilde{W}^*, S)$$

The above states that any dependence between treatment assignments in all periods and heterogeneity in potential outcomes in all periods, is explained by the (possibly period-t specific) factors $\tilde{W}^*$ and additional conditioning variables S.

We suppose that conditional on the latent variables $\tilde{W}^*$ and additional conditioning variables S, both the treatment assignments and heterogeneity are first-order Markov process:

$$(X_{t+2}, ..., X_T, U_{t+2}, ..., U_T) \perp (X_1, ..., X_t, U_1, ..., U_t)|(\tilde{W}^*, X_{t+1}, U_{t+1}, S)$$

Recall from the first section that we assume (without loss of generality) that for any x and any $u_1 \neq u_2$ that $y_{0,t}(x, u_1) \neq y_{0,t}(x, u_2)$. Thus the condition above is equivalent to:

$$(X_{t+2}, ..., X_T, Y_{t+2}, ..., Y_T) \perp (X_1, ..., X_t, Y_1, ..., Y_t)|(\tilde{W}^*, X_{t+1}, Y_{t+1}, S)$$

Then set V, Z and W^* as follows:

$$V = (X_{t+1}, ..., X_T, Y_{t+1}, ..., Y_T)$$

$$Z = (X_1, ..., X_{t-1}, X_{t+1}, Y_1, ..., Y_{t-1}, Y_{t+1})$$

$$W^* = (\tilde{W}^*, X_{t+1}, Y_{t+1})$$

Then Assumption 1 holds:

$$U_t \perp (X_t, Z)|(W^*, S)$$

$$V \perp (X_t, Z)|(W^*, S)$$

Again, we can allow for greater order Markov processes. Suppose we make the following k^{th} order Markov assumptions. Define C_t by:

$$C_t = (\tilde{W}^*, X_{t+1}, ..., X_{t+k+1}, U_{t+1}, ..., U_{t+k+1}, S)$$

Then:

$$(X_{t+k+2}, ..., X_T, U_{t+k+2}, ..., U_T) \perp (X_1, ..., X_t, U_1, ..., U_t)|C_t$$

Then letting V, Z and W^* be defined as follows, Assumption 1 is satisfied:

$$Z = (X_1, ..., X_{t-1}, X_{t+1}, ..., X_{t+k}, Y_1, ..., Y_{t-1}, Y_{t+1}, ..., Y_{t+k})$$
$$V = (X_{t+1}, ..., X_T, , Y_{t+1}, ..., Y_T)$$

$$W^* = (\tilde{W}^*, X_{t+1}, ..., X_{t+k}, Y_{t+1}, ..., Y_{t+k})$$

Assumption 2 in the Panel Case

Assumption 2 consists of two L_2-completeness conditions. As discussed in Section 1, these conditions can be understood as instrumental relevance conditions: conditioning on X_t and S, both V and Z must be relevant for the perfect controls W^*.

In the Markov treatment assignment case discussed above, both V and Z are likely to be strongly associated with the perfect controls $W^* = (\tilde{W}^*, X_{t+1})$. Both Z and V contain X_{t+1} and in addition some treatment assignments for periods other than t. Recall that $\tilde{W}^*$ is a vector of perfect controls, i.e., a vector that explains confounding between $(X_1, ..., X_T)$ and U_t. If there is confounding in each period that is explained by the presence of those same variables $\tilde{W}^*$ then each component of Z and V ought to be informative about $\tilde{W}^*$.

Note that the larger the time dimension T, the more numerous are available treatments from different periods from which one may form Z and V. Z and V are then more likely to satisfy Assumption 2. As we discuss in Section 1, it is prudent to require that an order condition hold, i.e., that V and Z each be of a weakly larger dimension than W^*. In the first-order Markov treatment assignment example above when treatments are scalar, Z is of length t and V is of length $T-t$. Therefore, the order condition requires that $\tilde{W}^*$ be of length at most $\min\{t, T-t\}$. In the k^{th} order Markov case we need $\tilde{W}^*$ to be of dimension weakly less than $\min\{t, T-t-k+1\}$.

Proofs

Proof Theorem 1

For each $x \in \mathcal{X}$ and $s \in \mathcal{S}$ define the linear operator $B_{x,s} : L_2(F_{W^*|X=x,S=s}) \to L_2(F_{V|X=x,S=s})$ by:

$$B_{x,s}[\delta](v) = E[\delta(W^*)|V = v, X = x, S = s]$$

For $F_{V|X=x,S=s}$-almost all v. Assumption 2.ii is then equivalent to the statement that the operator $B_{x,s}$ is injective. The adjoint operator $B^*_{x,s}$: $L_2(F_{V|X=x,S=s}) \to L_2(F_{W^*|X=x,S=s})$ is then given by:

$$B^*_{x,s}[\delta](w^*) = E[\delta(Z)|W^* = w^*, X = x, S = s]$$

For $F_{W^*|X=x,S=s}$-almost all w^*. Injectivity of an operator implies that its adjoint has dense range (see e.g., Kress (2014) Theorem 15.8). Therefore, there exists a sequence of functions defined on $\mathcal{X} \times \mathcal{S} \times \mathcal{V}$, $\{\gamma_k\}_{k=1}^{\infty}$, so that for F_X-almost all x and F_S-almost all s, $\gamma_k(x, s, \cdot)$ is in $L_2(F_{V|X=x,S=s})$ and:

$$\lim_{k\to\infty} E\bigg[E[\gamma_k(X, S, V) - Y|X, S, W^*]^2|X = x, S = s\bigg] = 0$$

By iterated expectations and Assumption 1.b:

$$\begin{aligned} E[\gamma_k(X, S, V)|X, S, Z] &= E\big[E[\gamma_k(X, S, V)|X, S, W^*, Z]\big|X, S, Z\big] \\ &= E\big[E[\gamma_k(X, S, V)|X, S, W^*]\big|X, S, Z\big] \end{aligned}$$

And by Assumption 1.a and iterated expectations implies:

$$\begin{aligned} E[Y|X, S, Z] &= [y_0(X, U)|X, S, Z] \\ &= E\big[E[y_0(X, U)|X, S, W^*, Z]\big|X, S, Z\big] \\ &= E\big[E[y_0(X, U)|X, S, W^*]\big|X, S, Z\big] \\ &= E\big[E[Y|X, S, W^*]\big|X, S, Z\big] \end{aligned}$$

Using the above and noting that a conditional expectation operator has operator norm weakly less than unity:

$$\lim_{k\to\infty} E\bigg[E[\gamma_k(X, S, V) - Y|X, S, Z]^2|X = x, S = s\bigg] = 0$$

Recall the decompositions $W^* = (\tilde{W}^*, \bar{W})$ and $Z = (\tilde{Z}, \bar{W})$ and definition of the linear operator $A_{x,s,\bar{w}}$ given in the main body of the paper. Assumption 2.i states that the adjoint operator $A^*_{x,s,\bar{w}}$ is injective and so $A_{x,s,\bar{w}}$ has dense range. Thus using Assumption 2.i and Assumption 3, the function $\tilde{w}^* \mapsto \frac{dF_{W^*|X=x_2,S=s}}{dF_{W^*|X=x_1,S=s}}(\tilde{w}^*, \bar{w})$ satisfies Picard's criterion to be in the range of $A_{x,s,\bar{w}}$ (see Kress (2014) Theorem 15.18). Thus there exists a function φ defined on $\mathcal{X}^2 \times \mathcal{S} \times \mathcal{Z}$ with $\varphi\big(x_1, x_2, s, (\cdot, \bar{w})\big) \in L_2(F_{\tilde{Z}|X=x_1,S=s,\bar{W}=\bar{w}})$ so that for F_X-almost all x_1 and x_2, F_S-almost all s, $F_{\bar{W}}$-almost all $\bar{w}$ and $F_{\tilde{W}^*}$-almost all w^*:

$$A_{x,s,\bar{w}}[\varphi\big(x_1, x_2, s, (\cdot, \bar{w})\big)](\tilde{w}^*) = \frac{dF_{W^*|X=x_2,S=s}}{dF_{W^*|X=x_1,S=s}}(\tilde{w}^*, \bar{w})$$

Equivalently, for F_X-almost all x_1 and x_2, F_S-almost all $s$and F_{W^*}-almost all w^*:

$$E[\varphi(x_1, x_2, s, Z)|X = x_1, S = s, W^* = w^*] = \frac{dF_{W^*|X=x_2,S=s}}{dF_{W^*|X=x_1,S=s}}(W^*) \quad (6)$$

And further, given Assumption 3.iv, at least one function φ that satisfies the above satisfies the following inequality (again see Kress (2014) Theorem 15.18). For F_X-almost all x_1 and x_2, F_S-almost all s and $F_{\bar{W}}$-almost all $\bar{w}$:

$$E\big[\varphi\big(x_1, x_2, s, (\tilde{Z}, \bar{w})\big)^2|X = x_1, S = s, \bar{W} = \bar{w}\big] \leq C$$

This implies that for F_X-almost all x_1 and x_2 and F_S-almost all s:

$$E[\varphi(x_1, x_2, s, Z)^2|X = x_1, S = s] \leq C$$

Now note that:

$$\begin{aligned}
&E[y_0(x_1, U)|X = x_2, S = s] \\
=&E\big[E[y_0(x_1, U)|X = x_2, S = s, W^*]\big|X = x_2, S = s\big] \\
=&E\big[E[y_0(X, U)|X = x_1, S = s, W^*]\big|X = x_2, S = s\big] \\
=&E\big[E[y_0(X, U)|X, S, W^*]\frac{dF_{W^*|X=x_2,S=s}}{dF_{W^*|X=x_1,S=s}}(W^*)|X = x_1, S = s\big] \\
=&E\big[E[y_0(X, U)|X, S, W^*]E[\varphi(x_1, x_2, s, Z)|X, S, W^*]|X = x_1, S = s\big] \\
=&E\big[E[y_0(X, U)|X, S, W^*]\varphi(x_1, x_2, s, Z)|X = x_1, S = s\big] \\
=&E\big[E[y_0(X, U)|X, S, W^*, Z]\varphi(x_1, x_2, s, Z)|X = x_1, S = s\big] \\
=&E\big[E[Y|X, S, Z]\varphi(x_1, x_2, s, Z)|X = x_1, S = s\big]
\end{aligned}$$

Where the first equality follows by iterated expectations, the second by Assumption 1.i, the third by definition of the Radon-Nikodym derivative, the fourth by 6, the fifth by iterated expectations, the sixth by Assumption 1.i, and the final equality by iterated expectations and the definition of Y.

Moreover, following similar steps:

$$\begin{aligned}
&E\big[\gamma_k(x_1, S, V)|X = x_2, S = s\big] \\
=&E\big[E[\gamma_k(x_1, S, V)|X, S, W^*]|X = x_2, S = s\big] \\
=&E\big[E[\gamma_k(X, S, V)|X = x_1, S, W^*]|X = x_2, S = s\big] \\
=&E\big[E[\gamma_k(X, S, V)|X, S, W^*]\frac{dF_{W^*|X=x_2,S=s}}{dF_{W^*|X=x_1,S=s}}(W^*)|X = x_1, S = s\big] \\
=&E\big[E[\gamma_k(X, S, V)|X, S, W^*]E[\varphi(x_1, x_2, s, Z)|X, S, W^*]|X = x_1, S = s\big] \\
=&E\big[E[\gamma_k(X, S, V)|X, S, W^*]\varphi(x_1, x_2, s, Z)|X = x_1, S = s\big] \\
=&E\big[E[\gamma_k(X, S, V)|X, S, W^*, Z]\varphi(x_1, x_2, s, Z)|X = x_1, S = s\big] \\
=&E\big[E[\gamma_k(X, S, V)|X, S, Z]\varphi(x_1, x_2, s, Z)|X = x_1, S = s\big]
\end{aligned}$$

Where the first equality follows by iterated expecations, the second by Assumption 1.ii, the third by the definition of the Radon-Nikodym derivative, the fourth by 6, the fifth by iterated expectations, the sixth by Assumption 1.ii, and the final equality by iterated expectations. Combining we get:

$$\begin{aligned}
&E[y_0(x_1,U)|X=x_2,S=s] - E\big[\gamma_k(x_1,S,V)|X=x_2,S=s\big] \\
&= E\Bigg[\big(E[Y-\gamma_k(X,S,V)|X=x_1,S=s,Z]\big)\varphi(x_1,x_2,s,Z)\Bigg|X=x_1,S=s\Bigg]
\end{aligned}$$

And so, since $E\big[\varphi(x_1,x_2,s,Z)^2|X=x_1,S=s\big] \leq C$, by Cauchy-Schwartz:

$$\begin{aligned}
&\big(E[y_0(x_1,U)|X=x_2,S=s] - E\big[\gamma_k(x_1,S,V)|X=x_2,S=s\big]\big)^2 \\
&\leq CE\Bigg[\big(E[Y-\gamma_k(X,S,V)|X,S,Z]\big)^2\Bigg|X=x_1,S=s\Bigg]
\end{aligned}$$

And we already established that the RHS converges to zero.
□

Proof of Theorem 2

First we show that under Assumptions 1, 2, 3 and 4.iii for F_X-almost all x:

$$\begin{aligned}
&\operatorname{ess\,sup}\big|\bar{y}(x|X,S) - \hat{\alpha}_n(x,X,S)'\hat{\theta}\big| \\
\leq &\operatorname{ess\,sup}\sqrt{C}E\Bigg[\big(g(X,S,Z) - \Pi_n(X,S,Z)'\hat{\theta}\big)^2\Bigg|X,S\Bigg]^{\frac{1}{2}} \\
&+O_p(b_n)
\end{aligned} \tag{7}$$

Where $\hat{\theta}$ is treated as a constant in the expectation on the RHS above. By the triangle inequality:

$$\begin{aligned}
&\big|\bar{y}(x_1|x_2,s) - \hat{\alpha}_n(x_1,x_2,s)'\hat{\theta}\big| \\
\leq &\big|\bar{y}(x_1|x_2,s) - \alpha_n(x_1,x_2,s)'\hat{\theta}\big| \\
&+\big|\big(\alpha_n(x_1,x_2,s) - \hat{\alpha}_n(x_1,x_2,s)\big)'\hat{\theta}\big|
\end{aligned}$$

By Assumption 4.iii:

$$\operatorname{ess\,sup}\big|\big(\alpha_n(x,X,S) - \hat{\alpha}_n(x,X,S)\big)'\hat{\theta}\big| \leq O_p(b_n)$$

Applying Theorem 1 with $\gamma_k(X,S,V) = \Phi(X,S,V)'\hat{\theta}$:

$$
\begin{aligned}
&\left(\bar{y}(x_1|x_2,s) - \alpha_n(x_1,x_2,s)'\hat{\theta}\right)^2 \\
\leq & CE\left[\left(g(X,S,Z) - \Pi_n(X,S,Z)'\hat{\theta}\right)^2 \middle| X = x_1, S = s\right]
\end{aligned}
$$

Combining gives 7.

Next note that by Assumption 4.i and 4.ii:

$$
\begin{aligned}
&E\left[\left(g(X,S,Z) - \Pi_n(X,S,Z)'\hat{\theta}\right)^2 \middle| X = x_1, S = s\right]^{\frac{1}{2}} \\
\leq & c \inf_{\theta \in \Theta_n} \operatorname{ess\,sup}_{x \in \mathcal{X}, s \in \mathcal{S}} E\left[\left(g(X,S,Z) - \Pi_n(X,S,Z)'\theta\right)^2 \middle| X = x, S = s\right]^{\frac{1}{2}} + O_p(\eta_n) \\
= & O_p(\kappa_n + \eta_n)
\end{aligned}
$$

Combining with 7 we get the result.

□

References

Ai, Chunrong, & Chen, Xiaohong. 2003. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, **71**(6), 1795–1843.

Andrews, Donald WK. 2017. Examples of L2-complete and boundedly-complete distributions. *Journal of Econometrics*, **199**(2), 213–220.

Arellano, Manuel, & Bond, Stephen. 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The review of economic studies*, **58**(2), 277–297.

Chen, Xiaohong, & Pouzo, Demian. 2012. Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Generalized Residuals. *Econometrica*.

Chen, Xiaohong, & Pouzo, Demian. 2015. Sieve Wald and QLR Inferences on Semi/Nonparametric Conditional Moment Models. *Econometrica*, **83**, 1013–1079.

Chen, Xiaohong, Chernozhukov, Victor, Lee, Sokbae, & Newey, Whitney K. 2014. Local Identification of Nonparametric and Semiparametric Models. *Econometrica*.

Cunha, Flavio, Heckman, James, & Schennach, Susanne. 2010. Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Econometrica*.

Darolles, Serge, Fan, Yanqin, Florens, Jean-Pierre, & Renault, Eric. 2011. Nonparametric Instrumental Regression. *Econometrica*.

Deaner, Ben. 2019. Nonparametric Instrumental Variables Estimation Under Misspecification.

D'Haultfoeuille, Xavier. 2011. On the completeness condition in nonparametric instrumental problems. *Econometric Theory*, **27**(3), 460–471.

Hausman, Jerry A., & Taylor, William E. 1981. Panel Data and Unobservable Individual Effects. *Econometrica*, **49**, 1377.

Holtz-Eakin, Douglas, Newey, Whitney, & Rosen, Harvey S. 1988. Estimating Vector Autoregressions with Panel Data. *Econometrica*, **56**, 1371.

Hu, Yingyao, & Schennach, Susanne M. 2008. Instrumental Variable Treatment of Nonclassical Measurement Error Models. *E*, **76**, 195–216.

Hu, Yingyao, & Shiu, Ji-Liang. 2018. Nonparametric identification using instrumental variables: sufficient conditions for completeness. *Econometric Theory*, **34**(3), 659–693.

Ichimura, Hidehiko, & Newey, Whitney K. 2017. *The influence function of semiparametric estimators.*

Kress, Rainer. 2014. *Linear Integral Equations.*

Kuroki, Manabu, & Pearl, Judea. 2014. Measurement Bias and Effect Restoration in Causal Inference. *Biometrika.*

Miao, Wang, Geng, Zhi, & Tchetgen, Eric J. Tchetgen. 2018. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, **105**, 987–993.

Newey, Whitney K., & Powell, James L. 2003. Instrumental Variable Estimation of Nonparametric Models. *Econometrica*, **71**, 1565–1578.

Rokkanen, Miikka AT. 2015. *Exam schools, ability, and the effects of affirmative action: Latent factor extrapolation in the regression discontinuity design.*

Severini, Thomas A., & Tripathi, Gautam. 2012. Efficiency bounds for estimating linear functionals of nonparametric regression models with endogenous regressors. *Journal of Econometrics*, **170**, 491–498.

VanderWeele, Tyler J., & Shpitser, Ilya. 2013. On the definition of a confounder. *The Annals of Statistics*, **41**, 196–220.