Automatic Locally Robust Estimation with Generated Regressors*

Juan Carlos Escanciano
Universidad Carlos III de Madrid

Telmo J. Pérez-Izquierdo
Universidad Carlos III de Madrid

November 21, 2022

Abstract

Many economic and causal parameters of interest depend on generated regressors, including structural parameters in models with endogenous variables estimated by control functions and in models with sample selection. Inference with generated regressors is complicated by the very complex expression for influence functions and asymptotic variances. To address this problem, we propose automatic Locally Robust/debiased GMM estimators in a general setting with generated regressors. Importantly, we allow for the generated regressors to be generated from machine learners, such as Random Forest, Neural Nets, Boosting, and many others. We use our results to construct novel Doubly Robust estimators for the Counterfactual Average Structural Function and Average Partial Effects in models with endogeneity and sample selection, respectively.

Keywords: Local robustness, orthogonal moments, double robustness, semiparametric estimation, bias, GMM. JEL Classification: C13; C14; C21; D24

*Research supported by Spanish grants PGC 2018-096732-B-100 and PID2021-127794NB-I00 (MCI/AEI/FEDER, UE).

1 Introduction

Many economic and causal parameters of interest depend on generated regressors. Leading examples include the Counterfactural Average Structural Function (CASF) in models with endogenous variables estimated by control functions (cf. Blundell and Powell, 2004; Stock, 1989, 1991), and Average Partial Effects (APE) in sample selection models (Das et al., 2003). There are currently no econometric methods for inference on these parameters allowing for generated regressors obtained by machine learning. The goal of this paper is to propose Automatic Locally Robust(ALR)/Debiased estimators of and inference on structural parameters in such models.

The paper builds on two different literatures. The first literature is the classical literature on semiparametric estimators with generated regressors, see Ahn and Powell (1993); Heckman et al. (1998); Ichimura and Lee (1991); Imbens and Newey (2009); Newey et al. (1999); Rothe (2009), among others. The asymptotic properties of several estimators within this class is given by Hahn and Ridder (2013, 2019) and Mammen et al. (2012, 2016). With respect to these papers, we allow the second step to be semiparametric or parametric (on top of fully non-parametric). Furthermore, we contribute to this literature by proposing LR automatic estimators. Automatic estimation is very well motivated in this setting because the form of the influence function and asymptotic variances is very complex.

The second literature we build on is the more recent literature on LR/Debiased estimators, see Chernozhukov et al. (2018, 2022a). With the only exception of Sasaki and Ura (2021), this literature has not considered models with generated regressors. Our results complement the analysis of the Policy Relevant Treatment Effect (PRTE) in Sasaki and Ura (2021) by providing automatic estimation of the influence function. Relative to the automatic LR literature (e.g. Chernozhukov et al., 2022b) we innovate in considering a nonlinear setting with an implicit functional (the generated regressor as a conditioning argument) for which an analytic derivative is not available for general machine learners.

As an application of our methods we propose novel Automatic Locally Robust (ALR) estimators for the CASF parameter of Blundell and Powell (2004) and for the APE in a sample selection model with a flexible selection equation estimated by machine learning. All these examples are characterized by being linear functionals of a second step function satisfying orthogonality conditions involving generated regressors (the control function or the propensity score) from a first step. We show that it is straightforward to construct automatic Double-Robust (DR) estimators that are robust to functional form assumptions for the second step. For instance, a practical approach could be to fit a partially linear specification for the second step, like in Robinson (1988) but with a non-parametric function of the generated regressors. Our results cover this case, in which the second step is semiparametric.

The DR estimators are, however, not LR to the generated regressors in general. To construct fully LR estimators we use numerical derivatives to account for the presence of generated

regressors. Fortunately, our automatic approach is amenable to any machine learning method for which predictions out of sample are available. Another approach could be to specify a model for the second step for which analytical derivatives are available. We note that the DR moment conditions are robust to this model being misspecified.

The rest of the paper is organized as follows. Section 2 introduces the setting and the examples. Section 2.1 finds the influence function of parameters identified by moments with generated regressors. Section 3 gives the general construction of automatic LR moments with generated regressors. In Section 4, we provide the details for Debiased LR GMM estimation. A summary of the estimation algorithm is given in Section 4.2. Section 5 develops some examples.

2 Setting and Examples

We observe data $W = (Y, D, Z)$ with cumulative distribution function (cdf) F_0. For simplicity, we consider that Y and D are one-dimensional. In our setting, there is a first step linking D with Z. The first step results in a one-dimensional generated regressor

$$V \equiv \varphi(D, Z, g_0),$$

where φ is a known function of observed variables (D, Z) and an unknown parameter $g_0 \in \Gamma_1$, for Γ_1 a linear and closed subspace of the Hilbert space $L_2(Z)$ of square-integrable functions of Z.[1] The unknown parameter g_0 solves the orthogonal moments

$$\mathbb{E}[\delta_1(Z)(D - g_0(Z))] = 0 \text{ for all } \delta_1 \in \Gamma_1. \tag{2.1}$$

This setting covers parametric, semiparametric, and non-parametric first steps. For example, when $\Gamma_1 = L_2(Z)$, we have $g_0(Z) = \mathbb{E}[D|Z]$.

Then, there is a second step linking Y with a component of (D, Z), denoted by X, and the generated regressor V, through the moment restrictions

$$\mathbb{E}[\delta_2(D, Z)(Y - h_0(X, V))] = 0 \text{ for all } \delta_2 \in \Gamma_2(g_0), \tag{2.2}$$

where $\Gamma_2(g_0)$ is a linear and closed subspace of $L_2(D, Z)$. We have that $h_0(X, \varphi(D, Z, g_0))$, understood as a function of (D, Z) is an element of $\Gamma_2(g_0)$. The set $\Gamma_2(g_0)$ may depend on the fist step parameter g_0. In some settings, $\Gamma_2(g_0)$ includes only functions of X and the generated regressor V. That is, $\Gamma_2(g_0)$ includes functions with the following shape: $\delta_2(D, Z) = \delta(X, \varphi(D, Z, g_0))$ for

[1] *Notation:* For a (measurable) function $f(w)$, $\mathbb{E}[f(W)] \equiv \int f(w) dF_0(w)$ denotes expectation w.r.t. the distribution F_0. For simplicity of notation, omit that the measure when referring to the L_2 Hilbert spaces of measurable functions with finite second moments. This measure is the marginal distribution that F_0 induces on some of the components of W.

$\delta \in \Gamma$, a linear and closed subspace of $L_2(X, V)$. For instance, Hahn and Ridder (2013) and Mammen et al. (2016) consider where the second step is a non-parametric regression of Y on (X, V). In that case, $\Gamma_2(g) = L_2(g)$, with $L_2(g) \equiv \{(d, z) \mapsto \delta(x, \varphi(d, z, g)) \colon \delta \in L_2(X, V)\} \subseteq L_2(D, Z)$.

Let $\Theta \subseteq \mathbb{R}$ denote the space where the structural parameter of interest lies. We have the moment function $m \colon \mathbb{R}^{\dim(W)} \times L_2(Z) \times L_2(X, V) \times \Theta \to \mathbb{R}$. The parameter of interest θ_0 is identified in a third step by a GMM moment condition

$$\mathbb{E}[m(W, g_0, h_0, \theta_0)] = 0.$$

Here we assume that θ_0 is identified by these moments, i.e. that θ_0 is the unique solution to $E[m(W, g_0, h_0, \theta)] = 0$ over $\theta \in \Theta$. Extensions of our setting to a larger number of moment conditions, structural parameters, and multiple variables D and Y are straightforward.

We illustrate the notation and concepts with two general running examples.

Example 1 (Control Function Approach) We observe $W = (Y, D, Z)$ satisfying the model $Y = H(X, U)$, for an unknown function H. The main feature of this model is that D, a component of X, may be an endogenous regressor. We assume that the endogenous regressor satisfies $D = g_0(Z) + V$, with U and V being unobserved correlated error terms. The function g_0 could be identified by a conditional mean restriction, as in equation (2.1). We assume a Control Function approach: where $U|X, V \sim U|V$, where $\sim$ denotes equally distributed. Thus, the corresponding φ is

$$V \equiv \varphi(X, Z, g_0) \equiv D - g_0(Z).$$

As in Blundell and Powell (2004), the Control Function assumption implies

$$\begin{aligned}\mathbb{E}[Y|X = x, V = v] &= \mathbb{E}[H(X, U)|X = x, V = v] = \mathbb{E}[H(x, U)|X = x, V = v] \\ &= \mathbb{E}[H(x, U)|V = v] \equiv h_0(x, v).\end{aligned}$$

This defines the second step. In this example, we have that $\Gamma_2(g) = L_2(g)$.

The Control Function assumption allows us to identify the Average Structural Function (ASF) at a point $x \in \mathbb{R}^{\dim(X)}$:

$$\text{ASF}_0(x) \equiv \mathbb{E}[H(x, U)] = \mathbb{E}[\mathbb{E}[H(x, U)|V]] = \mathbb{E}[h_0(x, V)].$$

Some conditions on the support of the random vectors are needed for the above equation to hold (see Blundell and Powell, 2004; Imbens and Newey, 2009).

In this setup, a parameter of interest is the Counterfactual Average Structural Function (CASF) given by

$$\theta_0 = \int \text{ASF}(x^*) dF^*(x^*),$$

for an counterfactual distribution F^*. When F^* is implied by a certain policy, the CASF may be used to measure the effect of the policy (see Blundell and Powell, 2004; Stock, 1989, 1991). By Fubini's Theorem, the CASF can be written as a function of (g_0, h_0):

$$\theta_0 = \int \mathbb{E}[h_0(x^*, \varphi(D, Z, g_0))] dF^*(x^*) = \mathbb{E}\left[\int h_0(x^*, \varphi(D, Z, g_0)) dF^*(x^*)\right].$$

Hence, the moment function that identifies the CASF is:

$$m(w, g, h, \theta) = \int h(x^*, \varphi(d, z, g)) dF^*(x^*) - \theta.$$

We note here that the CASF is not covered by the work of Hahn and Ridder (2013, 2019). The key difference is that the functional defining the CASF cannot be written as $\mathbb{E}[\eta(X, \text{ASF}_0(X))]$ for a function η with domain in an Euclidean space. We will propose below a novel Doubly Robust (DR) estimator for the CASF. ■

Example 2 (Sample Selection Models) We observe $W = (Y, D, Z)$ following the model $Y = Y^*D \equiv H(X, \varepsilon)D$, where X is a component of Z, and we do not observe Y^* when $D = 0$. This is a very general setting for sample selection models. We do not know much about the selection, so this is given by $D = 1[g_0(Z) - U \geq 0]$, where U is uniformly distributed in $[0, 1]$. The unobserved errors ε and U, though independent of Z, are correlated with each other (selection on unobservables). In this example, $V = g_0(Z) = \mathbb{E}(D|Z)$. Then, it can be shown that

$$\begin{aligned}\mathbb{E}(Y|Z) &= \mathbb{E}(H(X, \varepsilon) 1[g_0(Z) - U \geq 0] | Z) \\ &= h_0(X, V).\end{aligned}$$

This setting provides a nonparametric extension of the classical model of Heckman (1979), where $H(X, \varepsilon) = X'\beta_0 + \varepsilon$, $g_0(Z) = Z'\gamma_0$, and the joint distribution of (ε, U) is bivariate Gaussian.

As a parameter of interest consider the Average Partial Effects (APE) given, for simplicity of presentation for a one-dimensional continuous regressor, by

$$\theta_0 = \mathbb{E}\left[\frac{\partial h_0}{\partial x}(X, V)\right].$$

The moment function identifying the APE is

$$m(w, g, h, \theta) = \frac{\partial}{\partial s} h(s, g(z))\Big|_{s=x} - \theta.$$

This parameter is covered by Proposition 5 in Hahn and Ridder (2019). However, the authors do not allow for ML estimators in the first and second steps. Bellow we prose a novel DR estimator for the APE which allows for ML first and second step estimators. ■

2.1 Orthogonal Moment Functions with Generated Regressors

We follow Chernozhukov et al. (2022a, henceforth, CEINR) for the construction of LR-Debiased-Orthogonal moment functions. Furthermore, we show that the effect of the first and second step estimation can be studied separately. This will allow us to construct separate automatic estimators of the nuisance parameters in first and second step Influence Functions (IF).

We begin by introducing some additional concepts and notation. Let F denote a possible cdf for a data observation W. We denote by $g(F)$ the probability limit an estimator $\hat{g}$ of the first step when the true distribution of W is F, i.e., under general misspecification (see Newey, 1994). Here, F is unrestricted except for regularity conditions such as existence of $g(F)$ or the expectation of certain functions of the data. For example, if $\hat{g}(z)$ is a nonparametric estimator of $\mathbb{E}[D|Z=z]$ then $g(F)(z) = E_F[D|Z=z]$ is the conditional expectation function when F is the true distribution of W, denoted by E_F, which is well defined under the regularity condition that $E_F[|D|]$ is finite. We assume that $g(F)$ is identified as the solution in g to

$$E_F[\delta_1(Z)(D - g(Z))] = 0 \text{ for all } \delta_1 \in \Gamma_1.$$

Hence, we have that $g(F_0) = g_0$, consistent with g_0 being the probability limit of $\hat{g}$ when F_0 is the cdf of W.

To study the effect of the second step, suppose that W is distributed according to F. However, the first step parameter is independently fixed to g. Let $h(F, g)$ be the solution in h to

$$E_F\left[\delta_2(D, Z)\{Y - h(X, \varphi(D, Z, g))\}\right] = 0 \text{ for all } \delta_2 \in \Gamma_2(g).$$

The solution of the above equation is a function of (x, v): $h(F, g)(x, v)$. We have that $h(F_0, g_0) = h_0$. We may think of the mapping $h(F, g)$ as the probability limit of an estimator of h_0 under the following conditions: (i) the true distribution of W is F and (ii) the estimator is built with the first step parameter fixed to g. A feasible estimator $\hat{h}$ of h_0 will, however, rely on the estimator $\hat{g}$. Therefore, we assume that the probability limit of $\hat{h}$ under general misspecification is $h(F, g(F))$.

To introduce orthogonal moments, let H be some alternative distribution that is unrestricted except for regularity conditions, and $F_\tau \equiv (1-\tau)F_0 + \tau H$ for $\tau \in [0, 1]$. We assume that H is chosen so that $g(F_\tau)$ and $h(F_\tau, g(F_\tau))$ exist for τ small enough, and possibly other regularity conditions are satisfied. The IF that corrects for *both first and second step estimation*, as introduced in CEINR, is the function $\phi(w, g, h, \alpha, \theta)$ such that

$$\begin{aligned} &\frac{d}{d\tau}\mathbb{E}[m(W, g(F_\tau), h(F_\tau, g(F_\tau)), \theta)] = \int \phi(w, g_0, h_0, \alpha_0, \theta) dH(w), \\ &E[\phi(W, g_0, h_0, \alpha_0, \theta)] = 0, \text{ and } E[\phi(W, g_0, h_0, \alpha_0, \theta)^2] < \infty, \end{aligned} \tag{2.3}$$

for all H and all θ. Here α is an unknown function, additional to (g, h), on which only the IF depends. The "true parameter" α_0 is the α such that equation (2.3) is satisfied. Throughout

the paper, $d/d\tau$ is the derivative from the right (i.e. for non-negative values of τ) at $\tau = 0$. As in the work of von Mises (1947), Hampel (1974), and Huber (1981), this equation is the Gateaux derivative characterization of the IF of the functional $\bar{m}(g(F), h(F, g(F)), \theta)$, with

$$\bar{m}(g, h, \theta) \equiv \mathbb{E}[m(W, g, h, \theta)].$$

Orthogonal moment functions can be constructed by adding this IF to the original identifying moment functions to obtain

$$\psi(w, g, h, \alpha, \theta) \equiv m(w, g, h, \theta) + \phi(w, g, h, \alpha, \theta). \tag{2.4}$$

This vector of moment functions has two key orthogonality properties. First, we have that varying (g, h) away from (g_0, h_0) has no effect, locally, on $\mathbb{E}[\psi(W, g, h, \alpha_0, \theta)]$. The second property is that varying α will have no effect, globally, on $\mathbb{E}[\psi(W, g_0, h_0, \alpha, \theta)]$. These properties are shown in great generality in CEINR.

The IF in equation (2.3) measures the effect that the first step (estimation of g_0) and the second step (estimation of h_0) will have on the moment condition. We can show that these effects can be studied separately. The following lemma gives the result:

Lemma 2.1 *Assume that the chain rule can be applied. Then,*

$$\begin{aligned}\frac{d}{d\tau}\bar{m}(g(F_\tau), h(F_\tau, g(F_\tau)), \theta)] = &\frac{d}{d\tau}\bar{m}(g(F_\tau), h(F_0, g(F_\tau)), \theta)] \\ &+ \frac{d}{d\tau}\bar{m}(g_0, h(F_\tau, g_0), \theta)].\end{aligned}$$

The first derivative in the RHS accounts for the first step. As in Hahn and Ridder (2013), the first step affects the moment condition in two ways. We have a *direct impact* on $\bar{m}$, which includes the *effect of evaluating* h on the generated regressor. We also have an *indirect effect* on the moment that comes from g affecting estimation of h_0 in the second step (through conditioning). This is present in the term $h(F_0, g(F_\tau))$. The second derivative accounts for the effect of the second step. This effect is independent from the first step and, as such, considers that g_0 is known. This is captured by $h(F_\tau, g_0)$.

We may then find an IF corresponding to each step: $\phi_1(w, g, \alpha_1, \theta)$ and $\phi_2(w, h, \alpha_2, \theta)$, respectively. The IFs satisfy:

$$\frac{d}{d\tau}\bar{m}(g(F_\tau), h(F_0, g(F_\tau)), \theta)] = \int \phi_1(w, g_0, \alpha_{10}, \theta) dH(w) \text{ and} \tag{2.5}$$

$$\frac{d}{d\tau}\bar{m}(g_0, h(F_\tau, g_0), \theta)] = \int \phi_2(w, h_0, \alpha_{20}, \theta) dH(w), \tag{2.6}$$

on top of the zero mean and square integrability conditions (see equation 2.3). We therefore have that the IF accounting for both the first and second step is $\phi(w, g, h, \alpha, \theta) = \phi_1(w, g, \alpha_1, \theta) + \phi_2(w, h, \alpha_2, \theta)$, with $\alpha = (\alpha_1, \alpha_2)$.

We now provide the orthogonality conditions that will serve as a basis for the automatic estimation of the nuisance parameters α_{01} and α_{02}. Define the following moment conditions: $\psi_1(w, g, \alpha_1, \theta) \equiv m(w, g, h(F_0, g), \theta) + \phi_1(w, g, \alpha_1, \theta)$ for the first step, and $\psi_2(w, h, \alpha_2, \theta) \equiv m(w, g_0, h, \theta) + \phi_2(w, h, \alpha_2, \theta)$ for the second step. We note here that, in general, $\psi \neq \psi_1 + \psi_2$. Applying separately Theorem 1 in CEINR to ψ_1 and ψ_2 one gets

$$\frac{d}{d\tau}\mathbb{E}[\psi_1(W, g(F_\tau), \alpha_1(F_\tau), \theta)] = 0 \text{ and } \frac{d}{d\tau}\mathbb{E}[\psi_2(W, h(F_\tau, g_0), \alpha_2(F_\tau), \theta)] = 0.$$

Since Γ_1 and $\Gamma_2(g_0)$ are linear, the above equations then mean that, for all $\theta \in \Theta$,

$$\begin{aligned} &\frac{d}{d\tau}\mathbb{E}[\psi_1(W, g_0 + \tau\delta_1, \alpha_{10}, \theta)] = 0 \text{ for all } \delta_1 \in \Gamma_1 \text{ and} \\ &\frac{d}{d\tau}\mathbb{E}[\psi_2(W, h_0 + \tau\delta_2, \alpha_{20}, \theta)] = 0 \text{ for all } \delta_2 \in \Gamma_2(g_0). \end{aligned} \tag{2.7}$$

This result comes from applying Theorem 3 in CEINR. Here δ_1 represents a possible direction of deviation of $g(F)$ from g_0. In turn, δ_2 represents a possible deviation of $h(F, g_0)$ from h_0. The parameter τ is the size of a deviation. The innovation with respect to CEINR is that we can compute the IF ϕ by separately studying ψ_1 and ψ_2, corresponding to the first and second steps, respectively.

3 Automatic estimation of the nuisance parameters

The debiased moments require a consistent estimator $\hat{\alpha}$ of the nuisance parameters $\alpha_0 \equiv (\alpha_{01}, \alpha_{02})$. When the form of α_0 is known, one can plug-in nonparametric estimators of the unknown components of α_0 to form $\hat{\alpha}$. In the generated regressors setup, however, the nuisance parameters (specially α_{01}) have a complex analytical shape (see the result in equation (A.8) in the Appendix, the examples in Section 3.1, and Hahn and Ridder, 2013). Therefore, the plug-in estimator for $\hat{\alpha}$ may behave badly.

We propose an alternative approach which uses the orthogonality of ψ_1 and ψ_2 with respect to g and h, respectively, to construct estimators of $(\alpha_{10}, \alpha_{20})$. This approach does not require to know the form of α_0, it is "automatic" in only requiring the orthogonal moment functions and data for construction of $\hat{\alpha}$. Moreover, an automatic estimator can be constructed separately for each step. For more details, we refer to Section 3.2.

To build the automatic estimator, we need two ingredients: (i) a consistent estimator of the linearization of the moment condition and (ii) the shape of the first and second step IFs (up to α_1 and α_2, respectively). We therefore start by deriving these two ingredients.

3.1 First and Second Step Linearization

We start with the linearization of the second step effect. This result is well established in the literature and will follow inmediately if $\bar{m}(g_0, h, \theta)$ can be linearized in h (as in Newey, 1994, Equation 4.1). The shape of the influence function can be found by applying the results in Ichimura and Newey (2022).

Before introducing the result, we note that throughout this section (i) $\tau \mapsto h_\tau$ denotes a differentiable path, i.e., $0 \mapsto h_0$ and $dh_\tau/d\tau$ exists (equivalently for g_τ) and (ii) H is regular in the sense that, for $F_\tau \equiv (1-\tau)F_0 + \tau H$, $g(F_\tau)$ is a differentiable path in $L_2(Z)$, and $h(F_\tau, g_0)$ and $h(F_0, g(F_\tau))$ are differentiable paths in $L_2(X, V)$.

Proposition 3.1 *Under the following assumption:*

(A1) *There exists a function $D_2(w, h)$, linear and continuous in h, such that $d\bar{m}(g_0, h_\tau, \theta)/d\tau = d\mathbb{E}[D_2(W, h_\tau)]/d\tau$, for every $\theta \in \Theta$.*

We have that:

(Lin) *We can linearize the effect of the second step estimation:*

$$\frac{d}{d\tau}\bar{m}(g_0, h(F_\tau, g_0), \theta)] = \frac{d}{d\tau}\mathbb{E}[D_2(W, h(F_\tau, g_0))].$$

(IF) *There exists an $\alpha_{02} \in \Gamma_2(g_0) \cap L_2(g_0)$ such that the function*

$$\phi_2(w, h_0, \alpha_{02}, \theta) = \alpha_{02}(x, \varphi(d, z, g_0)) \cdot \{y - h_0(x, \varphi(d, z, g_0))\},$$

satisfies equation (2.5) and is thus the Second Step IF.

We note that, since $\bar{m}$ is linearized at (g_0, h_0, θ), D_2 (and also α_{02}) may also depend on (g_0, h_0, θ). This is omitted for notational simplicity, but will became relevant to construct feasible automatic estimators (see Section 4). We now find the linearization of $\bar{m}(g_0, h, \theta)$ in some examples:

Example 1 (continuing from p. 4) Assumption (A1) is easy to check for the CASF. Since $m(w, g_0, h, \theta)$ is already linear, we have that

$$D_2(w, h) = \int h(x^*, \varphi(d, z, g_0))dF^*(x^*).$$

In this case, we can also compute r_2, the Riesz Representer of $\mathbb{E}[D_2(W, h)]$. In the present example, since $\Gamma_2(g_0) = L_2(g_0)$ (non-parametric regression) and $r_2 \in L_2(g_0)$, we have that $\alpha_{02} = r_2$ (see equation (A.2) in the Appendix for the definition of α_{02}). To find it, we follow Pérez-Izquierdo (2022) and assume the existence of densities f^*, f_0^v and, f_0^{xv} for F^*, F_0^v and F_0^{xv},

respectively. Here F_0^v and F_0^{xv} denote the distribution under F_0 of V and (X,V), respectively. We then have that

$$\begin{aligned}\mathbb{E}[D_2(W,h)] &= \int h(x^*,v)f^*(x^*)f_0^v(v)dx^*dv = \int \frac{f^*(x^*)f_0^v(v)}{f_0^{xv}(x^*,v)}h(x^*,v)f_0^{xv}(x^*,v)dx^*dv \\ &= \mathbb{E}[r_2(X,V)h(X,V)],\end{aligned}$$

with $r_2(x,v) \equiv f^*(x^*)f_0^v(v)/f_0^{xv}(x^*,v)$. Note that, even if we have found the nuisance parameter $\alpha_{02} = r_2$, it has a rather complex shape. It depends on the density of the generated regressor V and on the joint density of (X,V). These objects are generally hard to estimate and may cause the plug-in estimator for r_2 to behave poorly. We advocate automatic estimation (Section 3.2) as a potential solution to this issue. ■

Example 3 (Hahn and Ridder (2013)' Setup) This example discusses the non-parametric setup in Hahn and Ridder (2013, Th. 5). Our theory generalizes their results in two ways: (i) we will allow for a wider range of generated regressors $\varphi(D,Z,g_0)$ and (ii) we consider a larger class of moment conditions. The authors focus on the case where there is a function $\eta\colon \mathbb{R}^{\dim(W)+1} \to \mathbb{R}$ such that

$$m(w,g,h,\theta) = \eta(w,h(x,g(z))) - \theta.$$

That is, in Hahn and Ridder (2013)'s setup, (g,h) enters the moment condition by the values that the "link" function η, with domain in an Euclidean space, takes at $(w,h(x,g(z)))$. Note that they fix $\varphi(d,z,g) = g(z)$ and that their Theorem 5 covers the fully non-parametric case: $\Gamma_1 = L_2(Z)$ and $\Gamma_2(g) = \{\delta(x,g(z))\colon \delta \in L_2(X,V)\}$ (other results in Hahn and Ridder, 2013, cover parametric first steps, but not the semiparametric case as in equation 2.1).

We start by linearizing the moment condition in h. To do it, we assume that η is differentiable w.r.t. y. In that case, as long as we can interchange differentiation and integration:

$$\begin{aligned}\frac{d}{d\tau}\bar{m}(g_0,h_\tau,\theta) &= \mathbb{E}\left[\frac{d}{d\tau}\eta(W,g_\tau(X,g_0(Z)))\right] \\ &= \mathbb{E}\left[\frac{\partial\eta}{\partial y}(W,h_0(X,g_0(Z)))\frac{d}{d\tau}h_\tau(X,g_0(Z))\right] \\ &= \frac{d}{d\tau}\mathbb{E}\left[\frac{\partial\eta}{\partial y}(W,h_0(X,g_0(Z)))h_\tau(X,g_0(Z))\right],\end{aligned}$$

so that $D_2(w,h) = \partial\eta/\partial y(w,h_0(x,g_0(z))) \cdot h(x,g_0(z))$. In the fully non-parametric case, the second step nuisance parameter is the Riesz Representer of $\mathbb{E}[D_2(W,h)]$. This is given by the expectation of $\partial\eta/\partial y(W,h_0(X,g_0(Z)))$ conditional on (X,V). ■

We now move to linearize the first step effect. Note that if the chain rule can be applied:

$$\begin{aligned}\frac{d}{d\tau}\bar{m}(g(F_\tau),h(F_0,g(F_\tau)),\theta) = &\frac{d}{d\tau}\bar{m}(g(F_\tau),h_0,\theta) \\ &+ \frac{d}{d\tau}\bar{m}(g_0,h(F_0,g(F_\tau)),\theta).\end{aligned} \tag{3.1}$$

The first derivative in the RHS can be easily analyzed if we linearize $\bar{m}(g, h_0, \theta)$ in g (see Assumption (A2) in Theorem 3.1 below).

To study $d\bar{m}(g_0, h(F_0, g(F_\tau)), \theta)/d\tau$ we proceed as in Lemma 1 in Hahn and Ridder (2013). Our extension of the lemma to allow for semiparametric second steps is based on Assumption (A3) in Theorem 3.1. The assumption is discussed below. Under Assumption (A3), we have that

$$\begin{aligned}\frac{d}{d\tau}\bar{m}(g_0, h(F_0, g(F_\tau)), \theta) = &-\frac{d}{d\tau}\mathbb{E}[\alpha_{02}(X, V)h_0(X, \varphi(D, Z, g(F_\tau)))] \\ &+ \frac{d}{d\tau}\mathbb{E}\left[\alpha_{02}(X, \varphi(D, Z, g(F_\tau))) \cdot (Y - h_0(X, V))\right].\end{aligned}$$

Therefore, the remaining step to linearize the moment condition in g is to linearize the terms $h_0(X, \varphi(D, Z, g(F_\tau)))$ and $\alpha_{02}(X, \varphi(D, Z, g(F_\tau)))$. To achieve this, we require h_0, α_0, and φ to be differentiable in the appropriate sense (see Assumption (A4) bellow).

Theorem 3.1 *Consider that Assumption (A1) holds and:*

(A2) *There exists a function $D_{11}(w, g)$, linear and continuous in g, such that $d\bar{m}(g_\tau, h_0, \theta)/d\tau = d\mathbb{E}[D_{11}(W, g_\tau)]/d\tau$, for every $\theta \in \Theta$.*

(A3) *For every $g \in \Gamma_1$ and $\delta \in L_2(X, V)$, we have that $\delta(\cdot, \varphi(\cdot, \cdot, g)) \in \Gamma_2(g) \Leftrightarrow \delta(\cdot, \varphi(\cdot, \cdot, g_0)) \in \Gamma_2(g_0)$.*

(A4) *h_0 and α_{02} are differentiable w.r.t. v. Moreover, the function $\varphi(d, z, g)$, understood as a mapping from $L_2(Z)$ to $L_2(D, Z)$, is Hadamard differentiable at g_0, with derivative D_φ.*

Then, we have that:

(Lin) *The function*

$$D_1(w, g) \equiv D_{11}(w, g) + \frac{\partial}{\partial v}\left[\alpha_{02}(x, v)(y - h_0(x, v))\right] \cdot D_\varphi g. \tag{3.2}$$

where the derivative is evaluated at $v = \varphi(d, z, g_0)$, satisfies

$$\frac{d}{d\tau}\bar{m}(g(F_\tau), h(F_0, g(F_\tau)), \theta)] = \frac{d}{d\tau}\mathbb{E}[D_1(W, g(F_\tau))].$$

(IF) *There exists an $\alpha_{01} \in \Gamma_1$ such that the function*

$$\phi_1(w, g_0, \alpha_{01}, \theta) = \alpha_{01}(z) \cdot \{d - g_0(z)\},$$

satisfies equation (2.6) and is thus the First Step IF.

Some comments are in order. Assumption (A3) simply means that the functions in $\Gamma_2(g)$ (at least those that only depend on (X, V)) have the same shape. It does not rule out any relevant case, up to our knowledge. For instance, the general case in which $\Gamma_2(g) = \{(d, z) \mapsto \delta(x, \varphi(d, z, g)) \colon \delta \in \Gamma\}$, for Γ a linear subspace of $L_2(X, V)$ satisfies the assumption. When $\Gamma = L_2(X, V)$ (i.e., $\Gamma_2(g) = L_2(g)$), the second step is a non-parametric regression on X and the generated regressor. One can also take $\Gamma = \{\beta' x + \eta(v) \colon \beta \in \mathbb{R}^{\dim(X)}, \eta \in L_2(V)\}$ to specify a partly linear model for the second step (Robinson, 1988). What Assumption (A3) rules out is to specify a partly linear model for some g's and a non-parametric regression for others. We also note that Assumption (A3) also covers the case in which $\Gamma_2(g) = L_2(D, Z)$, as in Escanciano et al. (2016, 2014).

Regarding Assumption (A4), the Haddamard derivative of φ is a linear and continuous map $D_\varphi \colon L_2(Z) \to L_2(D, Z)$ such that

$$\frac{d}{d\tau}\varphi(d, z, g_\tau) = \frac{d}{d\tau} D_\varphi g_\tau.$$

Usually, either $\varphi(d, z, g) = g(z)$ (first step prediction) or $\varphi(d, z, g) = d - g(z)$ (first step residual). In those cases, $D_\varphi g = g$ or $D_\varphi g = -g$, respectively.

The linearization of the first step effect is a rather complex function (see its definition in 3.2). The first term corresponds to the linearization of the *direct* effect of g. It is given by D_{11}, the linearization of $d\bar{m}(g, h_0, \theta)/\tau$. The second term corresponds to the *indirect* effect. Consistent estimation of the second term requires estimators for (i) g_0, (ii) h_0, (iii) $\partial h_0/\partial v$, (iv) α_{02}, and (v) $\partial \alpha_{02}/\partial v$. In Section 3.2, we propose an automatic estimator of the second step nuisance parameter, α_{02}. We can then plug-in to construct an automatic estimator of the first step nuisance parameter. An estimator for $\partial h_0/\partial v$ is discussed in Section 4.

We conclude the section by finding D_1 for several examples:

Example 1 (continuing from p. 9) The Control Function setup introduced in this paper satisfies Assumption (A3). In addition, as discussed above, our result also covers the setup in which it is assumed that $U|X, Z \sim U|X, V \sim U|V$ (see Blundell and Powell, 2003, 2004). In that case, since $h_0(X, \varphi(D, Z, g_0)) = \mathbb{E}[Y|D, Z]$, we would have that $\Gamma_2(g) = L_2(D, Z)$ for every g.

Moreover, the Control Function approach we follow here uses the residual of the first step to control for potential endogeneity. Thus, $\varphi(d, z, g) = d - g(z)$ and its linearization is $D_\varphi g = -g$. Provided that h_0 is differentiable w.r.t. v (Assumption (A4)), this allows us to linearize, w.r.t.

g, the moment condition defining the CASF. We have that:

$$\begin{aligned}\frac{d}{d\tau}\bar{m}(g_\tau, h_0, \theta) &= \frac{d}{d\tau}\mathbb{E}\left[\int h_0(x^*, \varphi(D, Z, g_\tau))dF^*(x^*) - \theta\right] \\ &= \mathbb{E}\left[\int \frac{d}{d\tau}h_0(x^*, \varphi(D, Z, g_\tau))dF^*(x^*)\right] \\ &= \mathbb{E}\left[\int \frac{\partial h_0}{\partial v}(x^*, \varphi(D, Z, g_0))\frac{d}{d\tau}\varphi(D, Z, g_\tau)dF^*(x^*)\right] \\ &= \frac{d}{d\tau}\mathbb{E}\left[-\int \frac{\partial h_0}{\partial v}(x^*, \varphi(D, Z, g_0))dF^*(x^*)g_\tau(Z)\right].\end{aligned}$$

This means that the linearization of the moment condition w.r.t. g is $D_{11}(w, g) = D_{11}(d, z)g(z)$, with

$$D_{11}(d, z) \equiv -\int \frac{\partial h_0}{\partial v}(x^*, d - g_0(z))dF^*(x^*).$$

We can now plug in the expression for D_{11} into equation (3.2), where the linearization of the first step effect is defined. Recall that $D_\varphi g = -g$. Then, for the CASF, equation (3.2) becomes

$$D_1(w, g) \equiv \left\{D_{11}(d, z) + \frac{\partial}{\partial v}\left[\alpha_{02}(x, v)(y - h_0(x, v))\right]\right\}g(z).$$

As discussed above, the linearization depends on h_0 and α_{02} and the derivatives of these functions w.r.t. v. It also depends on g_0, as $v \equiv d - g_0(z)$. Section 3.2 discusses how to construct an automatic estimator for the first step nuisance parameter α_{02}, which we can latter use to compute its derivative. Finding an estimator of the derivative of h_0 will depend on the estimator at hand. In Section 4 we propose a numerical derivative approach that works for a variety of second step estimators, such as Random Forest. ■

Example 3 (continuing from p. 10) Theorem 3.1 generalizes Theorem 5 in Hahn and Ridder (2013) to allow for (i) generated regressors given by arbitrary Hadamard differentiable functions φ and (ii) arbitrary functionals $\bar{m}(g, h, \theta)$ that are Hadamard differentiable w.r.t. g and h. We show how the expression for D_1 simplifies to that in Hahn and Ridder (2013, Th. 5).

We start by linearizing $\bar{m}(g, h_0, \theta)$ w.r.t. g. Note that Hahn and Ridder (2013), in the non-parametric case, fix $\varphi(d, z, g) = g(z)$. Then, $D_\varphi g = g$. On top of η being differentiable w.r.t. y, we require h_0 to be differentiable w.r.t. v (Assumption (A4)). Then:

$$\begin{aligned}\frac{d}{d\tau}\bar{m}(g_\tau, h_0, \theta) &= \mathbb{E}\left[\frac{\partial \eta}{\partial y}(W, h_0(X, g_0(Z)))\frac{d}{d\tau}h_0(X, g_\tau(Z))\right] \\ &= \mathbb{E}\left[\frac{\partial \eta}{\partial y}(W, h_0(X, g_0(Z)))\frac{\partial h_0}{\partial v}(X, g_0(Z))\frac{d}{d\tau}g_\tau(Z)\right] \\ &= \frac{d}{d\tau}\mathbb{E}\left[\frac{\partial \eta}{\partial y}(W, h_0(X, g_0(Z)))\frac{\partial h_0}{\partial v}(X, g_0(Z))g_\tau(Z)\right],\end{aligned}$$

and therefore $D_{11}(w,g) = \partial\eta/\partial y(w, h_0(x, g_0(z))) \cdot \partial h_0/\partial v(x, g_0(z)) \cdot g(z)$.

Recall from the previous discussion that the Second Step nuisance parameter satisfies:

$$\alpha_{02}(x,v) = \mathbb{E}\left[\frac{\partial\eta}{\partial y}(W, h_0(X, g_0(Z)))\middle| X = x, g_0(Z) = v\right].$$

So, if we denote $\xi(w) \equiv \partial\eta/\partial y(w, h_0(x, g_0(z)))$, equation (3.2) becomes:

$$D_1(w,g) \equiv \left\{(y - h_0(x,v)) \cdot \frac{\partial\alpha_{02}}{\partial v}(x,v) + (\xi(w) - \alpha_{02}(x,v)) \cdot \frac{\partial h_0}{\partial v}(x,v)\right\} g(z),$$

where $v \equiv g_0(z)$. This is the result in Hahn and Ridder (2013, Th. 5).

Moreover, note that $\alpha_{02}(x,v) = \mathbb{E}[\xi(W)|X = x, V = v]$. Then, if ξ is only a function of (x,v), the second term in the above equation cancels out. This is the case in Theorem 2 in Hahn and Ridder (2013). There, $\eta\colon \mathbb{R} \to \mathbb{R}$, and therefore, $\xi(w) = \partial\eta/\partial y(h_0(x,v))$ is a function of (x,v). ■

3.2 Building the automatic estimators

Equations (2.7) can be thought of as a population moment condition for $(\alpha_{01}, \alpha_{02})$ for each $(\delta_1, \delta_2) \in \Gamma_1 \times \Gamma_2(g_0)$. We start with the procedure to automatically estimate α_{02}, the nuisance parameter of the Second Step IF. We want to stress, nevertheless, that the procedure is quite general. Indeed, we will also apply it, *mutatis mutandis*, to the estimation of the nuisance parameter in the First Step IF.

The starting point is to expand the second equation in (2.7). For $\delta_2 \in \Gamma_2(g_0)$,

$$\frac{d}{d\tau}\bar{m}(g_0, h_0 + \tau\delta_2, \theta) + \frac{d}{d\tau}\mathbb{E}[\phi_2(W, h_0 + \tau\delta_2, \alpha_{20}, \theta)] = 0.$$

We will now combine the above equation with Proposition 3.1. By continuity and linearity of D_2, we have that

$$\frac{d}{d\tau}\bar{m}(g_0, h_0 + \tau\delta_2, \theta) = \frac{d}{d\tau}\mathbb{E}[D_2(W, h_0 + \tau\delta_2)] = \mathbb{E}[D_2(W, \delta_2)], \text{ for any } \delta_2 \in \Gamma_2(g_0).$$

Moreover, Proposition 3.1 gives us that $\phi_2 = \alpha_{02}(y - h_0)$. Thus, for any $\delta_2 \in \Gamma_2(g_0)$,

$$\frac{d}{d\tau}\mathbb{E}[\phi_2(W, h_0 + \tau\delta_2, \alpha_{20}, \theta)] = -\mathbb{E}[\delta_2(D,Z)\alpha_{20}(X,V)].$$

We note here that a sufficient condition to compute the derivative of ϕ_2 w.r.t. τ is that ϕ_2 is affine and continuous in h.

This means that, for each $\delta_2 \in \Gamma_2(g_0)$, the second equation in (2.7) leads to a moment condition for α_{02}:

$$\mathbb{E}[D_2(W, \delta_2)] - \mathbb{E}[\delta_2(D,Z)\alpha_{20}(X,V)] = 0, \text{ for each } \delta_2 \in \Gamma_2(g_0). \tag{3.3}$$

Since $\mathbb{E}[D_2(W,\delta_2)]$ is a linear functional, we will have a Riesz Representer $r_2 \in Ł_2(X,V)$ that expresses the first term above as the L_2 scalar product. This means that the above conditions are projection moment conditions. Indeed, they embed the notion that α_{02} is the projection of r_2 onto $\Gamma_2(g_0)$. However, the usefulness of the conditions in (3.3) is that they do not require finding the Riesz Representer. They are based in a linearization of the moment condition, which is generaly easier to find.

We now assume that there is a dictionary $(b_j)_{j=1}^{\infty}$, with $b_j \in \Gamma_2(g_0) \cap L_2(g_0)$, whose closed linear span is $\Gamma_2(g_0) \cap L_2(g_0)$. That is, any function in $\Gamma_2(g_0) \cap L_2(g_0)$ is can be approximated, in the L_2 sense, by a linear combination of b_j's. Then, there exists a sequence of real numbers $(\rho_j)_{j=1}^{\infty}$ such that $\alpha_{02} = \sum_{j=1}^{\infty} \rho_j b_j$. Thus, α_{02} can be approximated by $\mathbf{b}_J'\boldsymbol{\rho}_J$, where $\mathbf{b}_J = (b_1, ..., b_J)'$ and $\boldsymbol{\rho}_J = (\rho_1, ..., \rho_J)'$. We can now plug in $\mathbf{b}_J'\boldsymbol{\rho}_J$ into equation (3.3) for $\delta_2 = b_j$, $j = 1, ..., J$. This gives the following J moment conditions:

$$\mathbb{E}[\mathbf{b}_J(X,V)\mathbf{b}_J(X,V)']\boldsymbol{\rho}_J = \mathbb{E}[D_2(W,\mathbf{b}_J)],$$

where $D_2(w,\mathbf{b}_J) \equiv (D_2(w,b_1), ..., D_2(w,b_J))'$.

The above moment conditions can be used to construct an OLS-like estimator of $\boldsymbol{\rho}$. Note, however, that in high dimensional settings $\mathbb{E}[\mathbf{b}_J(X,V)\mathbf{b}_J(X,V)']$ may be near singular. Therefore, we rather focus on a regularized estimator for $\boldsymbol{\rho}$. Note that the moment conditions are the first order conditions of the minimization problem:

$$\min_{\boldsymbol{\rho}_J \in \mathbb{R}^J} \left\{ -2\mathbb{E}[D_2(W,\mathbf{b}_J)']\boldsymbol{\rho}_J + \boldsymbol{\rho}_J'\mathbb{E}[\mathbf{b}_J(X,V)\mathbf{b}_J(X,V)']\boldsymbol{\rho}_J \right\}.$$

We can regularize the problem by adding a penalty to the above objective function. Let $\|\boldsymbol{\rho}_J\|_q \equiv (\sum_{j=1}^{J} |\rho_j|^q)^{1/q}$ for $q \geq 1$. For a tunning parameter $\lambda \geq 0$, we can estimate $\boldsymbol{\rho}$ by minimizing:

$$\min_{\boldsymbol{\rho}_J \in \mathbb{R}^J} \left\{ -2\mathbb{E}[D_2(W,\mathbf{b}_J)']\boldsymbol{\rho}_J + \boldsymbol{\rho}_J'\mathbb{E}[\mathbf{b}_J(X,V)\mathbf{b}_J(X,V)']\boldsymbol{\rho}_J + \lambda\|\boldsymbol{\rho}_J\|_q^q \right\}. \tag{3.4}$$

For $q = 1$, the above is the Lasso objective function, while $q = 2$ corresponds to Ridge Regression. Additionally, we could consider elasic net type penalties, where $\lambda(\xi\|\boldsymbol{\rho}_J\|_2^2 + (1-\xi)\|\boldsymbol{\rho}_J\|_1)$, for $\xi \in [0,1]$, is addeed to the objective function.

We propose now an automatic estimator of α_{01}, the nuisance parameter of the First Step IF. The procedure is parallel to that proposed above. By Theorem 3.1, we can linearize $\bar{m}(g, h(F_0, g), \theta)$ by $D_1(w,g)$ (see equation 3.2). Again, we assume that there is a dictionary $(c_k)_{k=1}^{\infty}$ that spans Γ_1. Thus, $\alpha_{01} = \sum_{k=1}^{\infty} \beta_k c_k$ for a sequence of real numbers $(\beta_k)_{k=1}^{\infty}$. We can therefore construct K moment conditions

$$\mathbb{E}[\mathbf{c}_K(Z)\mathbf{c}_K(Z)']\boldsymbol{\beta}_K = \mathbb{E}[D_1(W,\mathbf{c}_K)],$$

where $\mathbf{c}_K = (c_1, ..., c_K)'$, $\boldsymbol{\beta}_K = (\beta_1, ..., \beta_K)'$, and $D_1(w,\mathbf{c}_K) \equiv (D_1(w,c_1), ..., D_1(w,c_K))'$. We use these conditions as a basis to construct the objective function to estimate $\boldsymbol{\beta}$:

$$\min_{\boldsymbol{\beta}_K \in \mathbb{R}^K} \left\{ -2\mathbb{E}[D_1(W,\mathbf{c}_K)']\boldsymbol{\beta}_K + \boldsymbol{\beta}_K'\mathbb{E}[\mathbf{c}_K(Z)\mathbf{c}_K(Z)']\boldsymbol{\beta}_K + \lambda\|\boldsymbol{\beta}_K\|_q^q \right\}, \tag{3.5}$$

where the tunning parameter λ may be different from that of the second step.

From the above discussion we conclude that automatic estimation of the first and second step nuisance parameters reduces to finding a consistent estimator of $\mathbb{E}[D_2(W, \mathbf{b}_J)']$ and $\mathbb{E}[D_1(W, \mathbf{c}_K)]$. We note that, in general, both D_2 and D_1 depend on (g_0, h_0, θ). In the sample moment conditions, these are replaced by cross-fit estimators (Section 4.1).

Furthermore, D_1 may additionally depend on $\partial h_0/\partial v$, the nuisance parameter of the Second Step α_{02} and its derivative $\partial \alpha_{02}/\partial v$ (see equation 3.2). Estimation of $\partial h_0/\partial v$ is discussed in Section 4.1. Here, we sketch a parsimonious approach to estimate the derivative of α_{02}. Recall that the Second Step nuisance parameter can be approximated by $\mathbf{b}_J'\boldsymbol{\rho}_J$. We may assume that the atoms $b_j(x, v)$ are differentiable w.r.t. v. We can then replace the nuisance parameter by its approximation $\mathbf{b}_J'\boldsymbol{\rho}$ and its derivative by $(\partial \mathbf{b}_J/\partial v)'\boldsymbol{\rho}_J$ in equation (3.2).

4 Estimation

In this section, we build debiased sample moment conditions for GMM estimation of θ. Debiased sample moments are based in the orthogonal moment function ψ in equation (2.4). Note that the IF ϕ that corrects for both the first and second step estimation is $\phi = \phi_1 + \phi_2$, the sum of the First and Second Step IFs. The shape of these functions is given in Proposition 3.1 and Theorem 3.1, respectively.

We propose to construct the sample moment conditions using cross-fitting. That is, we split the sample so that $\psi(W_i, g, h, \alpha, \theta)$ is averaged over observations i that are not used to estimate (g, h, α, θ). Cross-fitting (i) eliminates the "own observation bias", helping remainders to converge faster to zero, and (ii) eliminates the need for Donsker conditions for the estimators of (g, h, α), which is important for first and second step ML estimators (see CEINR; Chernozhukov et al., 2018; Newey and Robins, 2017).

We partition the sample $(W_i)_{i=1}^n$ into L groups I_ℓ, for $\ell = 1, ..., L$. For each group, we have estimators $\hat{g}_\ell$, $\hat{h}_\ell$ and $\hat{\alpha}_\ell = (\hat{\alpha}_{1\ell}, \hat{\alpha}_{2\ell})$ that use observations that are not in I_ℓ. We construct automatic estimators of α_0 satisfying this property in Section 4.1. Moreover, for each group, we consider that there is an initial estimator of θ_0, namely $\tilde{\theta}_\ell$, which does not use the observations in I_ℓ. CEINR propose to chose $L = 5$ for medium size datasets and $L = 10$ for small dataset.

Following CEINR, debiased sample moment functions are

$$\hat{\psi}(\theta) \equiv \hat{m}(\theta) + \hat{\phi}, \tag{4.1}$$

with

$$\hat{m}(\theta) \equiv \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell} m(W_i, \hat{g}_\ell, \hat{h}_\ell, \theta) \text{ and } \hat{\phi} \equiv \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell} \phi(W_i, \hat{g}_\ell, \hat{h}_\ell, \hat{\alpha}_\ell, \tilde{\theta}_\ell).$$

We use these moment functions to construct the debiased GMM estimator:

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \hat{\psi}(\theta)' \hat{\Upsilon} \hat{\psi}(\theta), \tag{4.2}$$

where $\hat{\Upsilon}$ is a positive semi-definite weighting matrix. As usual in GMM, a choice of $\hat{\Upsilon}$ that minimizes the asymptotic variance of $\hat{\theta}$ is $\hat{\Upsilon} = \hat{\Psi}^{-1}$, for

$$\hat{\Psi} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \hat{\psi}_{i\ell} \hat{\psi}_{i\ell}', \text{ with } \hat{\psi}_{i\ell} \equiv m(W_i, \hat{g}_\ell, \hat{g}_\ell, \tilde{\theta}_\ell) + \phi(W_i, \hat{g}_\ell, \hat{h}_\ell, \hat{\alpha}_\ell, \tilde{\theta}_\ell).$$

We illustrate the theory with the construction of debiased GMM estimator for the CASF:

Example 1 (continuing from p. 12) We focus on the construction of $\hat{m}(\theta)$. Note that $\phi_1 = \alpha_{01}(d - g_0)$ and $\phi_2 = \alpha_{02}(y - h_0)$ (see Theorem 3.1 and Proposition 3.1, respectively). Thus, finding $\hat{\phi}$ is straightforward once we have cross-fit estimators for the nuisance parameters (see Section 4.1 for the construction of $\hat{\alpha}_{1\ell}$ and $\hat{\alpha}_{2\ell}$).

Recall that the moment function defining the CASF is

$$m(w, g, h, \theta) = \int h(x^*, \varphi(d, z, g)) dF^*(x^*) - \theta.$$

We take as given that the econometrician has computed cross-fit estimators for the first and second steps: $\hat{g}_\ell$ and $\hat{h}_\ell$. Since the counterfactual distribution F^* is fixed by the econometrician, we propose a numerical integration approach to obtain the debias sample moments.

We consider that the econometrician can sample from F^*. Let $(X_s^*)_{s=1}^S$ be sa sample of size S from F^*. For and observation $i \in I_\ell$, let $\hat{V}_{i\ell} \equiv \varphi(D_i, Z_i, \hat{g}_\ell)$. We approximate the value of the moment function $m(W_i, \hat{g}_\ell, \hat{h}_\ell, \theta)$ by

$$\frac{1}{S} \sum_{s=1}^{S} \hat{h}_\ell(X_s^*, \hat{V}_{i\ell}) - \theta.$$

Note that S may be arbitrarily large (increasing the computational cost), so that the above term is close to $m(W_i, \hat{g}_\ell, \hat{h}_\ell, \theta)$.

Following equations (4.1) and (4.2), the debiased estimator for the CASF is

$$\hat{\theta} = \frac{1}{nS} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \sum_{s=1}^{S} \hat{h}_\ell(X_s^*, \hat{V}_{i\ell}) + \hat{\phi}.$$

■

4.1 Automatic estimation with cross-fitting

Debiased sample moment function require estimators of the nuisance parameters $(\hat{\alpha}_{1\ell}, \hat{\alpha}_{2\ell})$ for each group I_ℓ. These estimators must use only observations not in I_ℓ. This section is devoted to the construction of automatic estimators satisfying this property. Through the section, we consider that the econometrician has at her disposal first and second step estimators, $\hat{g}_{\ell\ell'}$ and $\hat{h}_{\ell\ell'}$, and an initial estimator, $\tilde{\theta}_{\ell\ell'}$, that use only observations not in $I_\ell \cup I_{\ell'}$.

The key to automatic estimation of the Second Step nuisance parameter is to find a consistent estimator of the linearization of the moment condition. In this section, we will write $D_2(w, h|g_0, h_0, \theta)$ to make explicit that the linearization may depend on (h_0, g_0, θ) (see Examples 1 and 3). For the linearization of the effect of first step estimation, we will write $D_1(w, g|g_0, h_0, \alpha_{02}, \theta)$, to emphasize that it may also depend on the Second Step nuisance parameter. D_1 generally depends also on the derivatives $\partial h_0/\partial v$ and $\partial \alpha_{02}/\partial v$. We do not make this explicit, but we will address the issue in this section.

We start with the automatic estimator for the Second Step nuisance parameter. For each ℓ, we provide a sample version of the objective function in (3.4) that uses only observations not in I_ℓ. Recall that we have a dictionary $(b_j)_{j=1}^\infty$ that spans $\Gamma_2(g_0) \cap L_2(g_0)$. We estimate $\mathbb{E}[D_2(W, \mathbf{b}_J)]$ by

$$\hat{D}_{2\ell} \equiv \frac{1}{n - n_\ell} \sum_{\ell' \neq \ell} \sum_{i \in I_{\ell'}} D_2(W_i, \mathbf{b}_J | \hat{g}_{\ell\ell'}, \hat{h}_{\ell\ell'}, \tilde{\theta}_{\ell\ell'}),$$

where n_ℓ is the number of observations in I_ℓ. In turn, $\mathbb{E}[\mathbf{b}_J(X, \varphi(D, Z, g_0))\mathbf{b}_J(X, \varphi(D, Z, g_0))']$ is estimated by

$$\hat{B}_\ell \equiv \frac{1}{n - n_\ell} \sum_{\ell' \neq \ell} \sum_{i \in I_{\ell'}} \mathbf{b}_J(X_i, \varphi(D_i, Z_i, \hat{g}_{\ell\ell'}))\mathbf{b}_J(X_i, \varphi(D_i, Z_i, \hat{g}_{\ell\ell'}))'.$$

With this, we can build an automatic estimator of the Second Step nuisance parameter that only uses observations not in I_ℓ. It is given by $\hat{\alpha}_{2\ell} = \mathbf{b}_J' \hat{\boldsymbol{\rho}}_{J\ell}$, where

$$\hat{\boldsymbol{\rho}}_{J\ell} = \arg \min_{\boldsymbol{\rho}_J \in \mathbb{R}^J} \left\{ -2\hat{D}_{2\ell}' \boldsymbol{\rho}_J + \boldsymbol{\rho}_J' \hat{B}_\ell \boldsymbol{\rho}_J + \lambda \|\boldsymbol{\rho}_J\|_q^q \right\}. \tag{4.3}$$

The tunning parameter λ can be chosen by cross-validation.

Example 1 (continuing from p. 17) We provide the ingredients to conduct automatic estimator of α_{02} for the CASF. Recall that the moment condition for the CASF was already linear in h and hence

$$D_2(w, h|g_0, h_0, \theta) = \int h(x^*, \varphi(d, z, g_0)) dF^*(x^*).$$

We follow the same strategy as before and approximate D_2 by numerical integration. For a sample $(X_s^*)_{s=1}^S$ drawn from F^*, we approximate $D_2(W_i, b_j|\hat{g}_{\ell\ell'}, \hat{h}_{\ell\ell'}, \tilde{\theta}_{\ell\ell'})$ by

$$\frac{1}{S}\sum_{s=1}^{S} b_j(X_s^*, \varphi(D_i, Z_i, \hat{g}_{\ell\ell'})),$$

for $j = 1, ..., J$. This can then be used to construct the objective function to estimate $\hat{\boldsymbol{\rho}}_{J\ell}$. ■

We now discuss automatic estimation of the First Step nuisance parameter. Again, for each ℓ, the goal is to build a sample version of the objective function in (3.5) that uses only observations not in I_ℓ. The constructions is almost similar to the one above. We will focus in the main differences.

For a dictionary $(c_j)_{j=1}^\infty$ that spans Γ_1, we can estimate $\mathbb{E}[\mathbf{c}_K(Z)\mathbf{c}_K(Z)']$ by

$$\hat{C}_\ell \equiv \frac{1}{n - n_\ell}\sum_{\ell' \neq \ell}\sum_{i \in I_{\ell'}} \mathbf{c}_K(Z_i)\mathbf{c}_K(Z_i)',$$

and $\mathbb{E}[D_1(W, \mathbf{c}_K)]$ by

$$\hat{D}_{1\ell} \equiv \frac{1}{n - n_\ell}\sum_{\ell' \neq \ell}\sum_{i \in I_{\ell'}} D_1(W_i, \mathbf{b}_J|\hat{g}_{\ell\ell'}, \hat{h}_{\ell\ell'}, \hat{\alpha}_{2\ell\ell'}, \tilde{\theta}_{\ell\ell'}). \tag{4.4}$$

The first difference is that D_1 depends on α_{02} on top of (g_0, h_0, θ). We therefore need to plug-in an estimator $\alpha_{2\ell\ell'}$ that only uses observations not in $I_\ell \cup I_{\ell'}$. This estimator can be constructed using the methodology above. The only adjustment needed is that one needs to replace I_ℓ by $I_\ell \cup I_{\ell'}$ to define $\hat{D}_{2\ell\ell'}$ and $\hat{B}_{\ell\ell'}$. For instance, to construct $\alpha_{2\ell\ell'} = \mathbf{b}_J'\hat{\boldsymbol{\rho}}_{J\ell\ell'}$, it is simple to define the optimization problem that $\hat{\boldsymbol{\rho}}_{J\ell\ell'}$ solves. Define $\bar{L} \equiv \{\ell, \ell'\}$ and let $\hat{g}_{\bar{L}\ell''}$, $\hat{h}_{\bar{L}\ell''}$, and $\tilde{\theta}_{\bar{L}\ell''}$ be estimators that only use observations not in $I_\ell \cup I_{\ell'} \cup I_{\ell''}$. Then, we can define

$$\hat{D}_{2\ell\ell'} \equiv \frac{1}{n - n_\ell - n_{\ell'}}\sum_{\ell'' \notin \bar{L}}\sum_{i \in I_{\ell''}} D_2(W_i, \mathbf{b}_J|\hat{g}_{\bar{L}\ell''}, \hat{h}_{\bar{L}\ell''}, \tilde{\theta}_{\bar{L}\ell''}) \text{ and}$$

$$\hat{B}_{\ell\ell'} \equiv \frac{1}{n - n_\ell - n_{\ell'}}\sum_{\ell'' \notin \bar{L}}\sum_{i \in I_{\ell''}} \mathbf{b}_J(X_i, \varphi(D_i, Z_i, \hat{g}_{\bar{L}\ell''}))\mathbf{b}_J(X_i, \varphi(D_i, Z_i, \hat{g}_{\bar{L}\ell''}))'.$$

Thus, $\hat{\boldsymbol{\rho}}_{J\ell\ell'}$ is given by the optimization problem in (4.3) with $\hat{D}_{2\ell\ell'}$ and $\hat{B}_{\ell\ell'}$ replacing $\hat{D}_{2\ell}$ and $\hat{B}_\ell$, respectively.

The most important difference is that D_1 generally depends also on the derivatives $\partial h_0/\partial v$ and $\partial\alpha_{02}/\partial v$. In Section 3.2, we have presented a parsimonious approach to estimate the derivative of α_{02}. Indeed, it is simple to construct an estimator $\partial\hat{\alpha}_{2\ell\ell'}/\partial v$ of the derivative of α_{02} that uses only observations not in $I_\ell \cup I_{\ell'}$. Since we have already estimated $\hat{\alpha}_{2\ell\ell} = \mathbf{b}_J'\hat{\boldsymbol{\rho}}_{J\ell\ell'}$, if each b_j is differentiable w.r.t. v, we have that $\partial\hat{\alpha}_{2\ell\ell'}/\partial v \equiv (\partial\mathbf{b}_J/\partial v)'\hat{\boldsymbol{\rho}}_{J\ell\ell'}$.

Estimation of $\partial h_0/\partial v$ may be more tricky. It will depend on the shape of the estimator $\hat{h}_{\ell\ell}$. Note that, since $h_0 \in \Gamma_2(g_0) \cap L_2(g_0)$, we may use the dictionary $(b_j)_{j=1}^{\infty}$ to approximate the parameter. In this case, $\hat{h}_{\ell\ell}$ will be a Lasso or Ridge Regression estimator and we can estimate the derivative of h_0 as we have estimated the derivative of α_{02}. Moreover, estimating h_0 is usually a low dimensional problem. Hence, when $\hat{h}_{\ell\ell}$ is a Kernel or a Local Linear Regression estimator, the derivatives of h_0 can be estimated by finding the analytical expression of the derivatives of the Kernel Function.

For a general ML estimator $\hat{h}_{\ell\ell'}$ (e.g., Random Forest), we propose a numerical derivative approach to estimate $\partial h_0/\partial v$. Let t_n be a tunning parameter depending on the sample size. We propose to estimate $\partial h_0/\partial v(x,v)$ by

$$\frac{\partial \hat{h}_{\ell\ell'}}{\partial v}(x,v) \equiv \frac{\hat{h}_{\ell\ell'}(x,v+t_n) - \hat{h}_{\ell\ell'}(x,v)}{t_n}. \tag{4.5}$$

Note that, usually, we need to compute the derivative evaluated at $(X_i, \varphi(D_i, Z_i, \hat{g}_{\ell\ell'}))$.

We have now seen all the difficulties in estimating $D_{1\ell}$ in equation (4.4). With these solved, we can proceed to construct and automatic estimator of the First Step nuisance parameter. The estimator is given by $\hat{\alpha}_{1\ell} = \mathbf{c}_K'\hat{\boldsymbol{\beta}}_{K\ell}$, where

$$\hat{\boldsymbol{\beta}}_{K\ell} = \arg\min_{\boldsymbol{\beta}_K \in \mathbb{R}^K} \left\{ -2\hat{D}_{1\ell}'\boldsymbol{\beta}_K + \boldsymbol{\beta}_K'\hat{C}_\ell\boldsymbol{\beta}_K + \lambda\|\boldsymbol{\beta}_K\|_q^q \right\}. \tag{4.6}$$

We illustrate this procedure by constructing an automatic estimator of the First Step nuisance parameter for the CASF:

Example 1 (continuing from p. 18) From the previous discussion, we have that:

$$D_1(w,g) = \left\{ D_{11}(d,z) + \frac{\partial}{\partial v}\left[\alpha_{02}(x,v)(y - h_0(x,v))\right] \right\} g(z), \text{ with}$$
$$D_{11}(d,z) \equiv -\int \frac{\partial h_0}{\partial v}(x^*, d - g_0(z)) dF^*(x^*).$$

We approximate D_{11} by numerical integration. Let $(X_s^*)_{s=1}^S$ be a sample from F^*. To estimate $D_{1\ell}$, we approximate $D_{11}(D_i, Z_i)$, with $i \in I_{\ell'}$, by

$$-\frac{1}{S}\sum_{s=1}^{S} \frac{\partial \hat{h}_{\ell\ell'}}{\partial v}(X_s^*, D_i - \hat{g}_{\ell\ell'}(Z_i)).$$

To estimate $D_{1\ell}$, it remains to show how to estimate the second term in the brackets, for an observation $i \in I_{\ell'}$. Define $V_{i\ell\ell'} \equiv \varphi(D_i, Z_i, \hat{g}_{\ell\ell'}) = D_i - \hat{g}_{\ell\ell'}(Z_i)$. Following the chain rule, we can estimate the second term by

$$\left(\frac{\partial \mathbf{b}_J}{\partial v}(X_i, \hat{V}_{i\ell\ell'})\right)' \hat{\boldsymbol{\rho}}_{J\ell\ell'} \cdot (Y_i - \hat{h}_{\ell\ell'}(X_i, \hat{V}_{i\ell\ell'})) - \mathbf{b}_J(X_i, \hat{V}_{i\ell\ell'})'\hat{\boldsymbol{\rho}}_{J\ell\ell'} \cdot \frac{\partial \hat{h}_{\ell\ell'}}{\partial v}(X_i, \hat{V}_{i\ell\ell'}). \tag{4.7}$$

Therefore, to estimate $D_{1\ell}$ according to equation (4.4), we have that, for $i \in I_{\ell'}$,

$$D_1(W_i, c_k|\hat{g}_{\ell\ell'}, \hat{h}_{\ell\ell'}, \hat{\alpha}_{2\ell\ell'}, \tilde{\theta}_{\ell\ell'}) = c_k(Z_i)\cdot \Bigg\{ -\frac{1}{S}\sum_{s=1}^{S} \frac{\partial \hat{h}_{\ell\ell'}}{\partial v}(X_s^*, V_{i\ell\ell'}) + \left(\frac{\partial \mathbf{b}_J}{\partial v}(X_i, \hat{V}_{i\ell\ell'})\right)' \hat{\boldsymbol{\rho}}_{J\ell\ell'} \cdot (Y_i - \hat{h}_{\ell\ell'}(X_i, \hat{V}_{i\ell\ell'})) - \mathbf{b}_J(X_i, \hat{V}_{i\ell\ell'})'\hat{\boldsymbol{\rho}}_{J\ell\ell'} \cdot \frac{\partial \hat{h}_{\ell\ell'}}{\partial v}(X_i, \hat{V}_{i\ell\ell'}) \Bigg\},$$

for each $k = 1, ..., K$. This can then be use to construct the objective function to estimate $\hat{\boldsymbol{\beta}}_{K\ell}$. ■

4.2 Estimation Algorithm

Here we provided an overview of our estimation algorithm. The imputs to the algorithm are cross-fit estimators of g_0, h_0 and θ_0. We note that one must provide a total of L estimators only using observations no in I_ℓ, $L(L-1)/2$ estimators only using observations not in $I_\ell \cup I_{\ell'}$, and $L(L-1)(L-2)/6$ estimators only using observations not in $I_\ell \cup I_{\ell'} \cup I_{\ell''}$. With these estimators at hand, we follow this algorithm:

1. Estimate $\hat{\alpha}_{2\ell\ell'} = \mathbf{b}_J'\hat{\boldsymbol{\rho}}_{J\ell\ell'}$, with $\hat{\boldsymbol{\rho}}_{J\ell\ell'}$ satisfying the optimization problem in (4.3), with $\hat{D}_{2\ell\ell'}$ and $\hat{B}_{\ell\ell'}$ replacing $\hat{D}_{2\ell}$ and $\hat{B}_\ell$, respectively.

2. Estimate $\partial\hat{h}_{\ell\ell'}/\partial v$ by equation (4.5).

3. Use $\hat{g}_{\ell\ell'}$, $\hat{h}_{\ell\ell'}$, $\partial\hat{h}_{\ell\ell'}/\partial v$, $\hat{\alpha}_{2\ell\ell'}$, $\partial\hat{\alpha}_{2\ell\ell'}/\partial v$, and $\tilde{\theta}_{\ell\ell'}$ to construct $\hat{D}_{1\ell}$ (see equation 3.2). Estimate $\hat{\alpha}_{1\ell} = \mathbf{c}_K'\hat{\boldsymbol{\beta}}_{K\ell}$, with $\hat{\boldsymbol{\beta}}_{K\ell}$ satisfying equation (4.6).

4. Use $\hat{g}_{\ell\ell'}$, $\hat{h}_{\ell\ell'}$, and $\tilde{\theta}_{\ell\ell'}$ to construct $\hat{D}_{2\ell}$. Estimate $\hat{\alpha}_{2\ell} = \mathbf{b}_J'\hat{\boldsymbol{\rho}}_{J\ell}$, with $\hat{\boldsymbol{\rho}}_{J\ell}$ satisfying equation (4.3).

5. Compute the bias correction term

$$\hat{\phi} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\left[\hat{\alpha}_{1\ell}(Z_i)\cdot(D_i - \hat{g}_\ell(Z_i)) + \hat{\alpha}_{2\ell}(X_i, \hat{V}_{i\ell})\cdot(Y_i - \hat{h}_\ell(X_i, \hat{V}_{i\ell}))\right],$$

   with $\hat{V}_{i\ell} \equiv \varphi(D_i, Z_i, \hat{g}_\ell)$.

6. Compute the debiased sample moment function in equation (4.1).

7. Construct the LR debiased GMM estimator of θ_0 by solving (4.2).

5 Examples

Example 2 (continuing from p. 5) Let $\partial h/\partial x(x,v)$ denote the derivative of $h(x,v)$ w.r.t. its first argument at (x,v). Let $\partial^2 h/\partial x\partial v(x,v)$ denote the derivative w.r.t. both arguments at (x,v). For the APE, we have that the moment function is linear in h. Thus:

$$D_2(w,h|g_0,h_0,\theta) = \frac{\partial h}{\partial x}(x,g_0(z)),$$

where we have already make explicit the dependence of D_2 on (g_0,h_0,θ). We can also linearize the moment condition in g to obtain:

$$D_1(w,g|g_0,h_0,\alpha_0,\theta) = \left\{ D_{11}(d,z) + \frac{\partial}{\partial v}\left[\alpha_{02}(x,v)(y-h_0(x,v))\right] \right\} g(z), \text{ with}$$
$$D_{11}(d,z) \equiv \frac{\partial^2 h_0}{\partial x \partial v}(x,g_0(z)).$$

The debias estimator for the APE is

$$\hat{\theta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell} \frac{\partial \hat{h}_\ell}{\partial x}(X_i,\hat{g}_\ell(Z_i)) + \hat{\phi},$$

where, for the estimator $\hat{h}_\ell$, we can estimate its derivative w.r.t. x by

$$\frac{\hat{h}_\ell}{\partial x}(x,v) \equiv \frac{\hat{h}_\ell(x+s_n,v) - \hat{h}_\ell(x,v)}{s_n},$$

for a tunning parameter s_n. To construct $\hat{\phi}$, we need to estimate α_{01} and α_{02}. We propose automatic estimators for these nuisance parameters.

We assume that $\partial h_0/\partial x$ and $\partial h_0/\partial v$ are differentiable, so we can interchange the order of differentiation.

Consider a dictionary $(b_j)_{j=1}^{\infty}$ that is differentiable w.r.t. both x and v. To Estimate $\hat{D}_{2\ell}$, we can compute $D_2(W_i,b_j|\hat{g}_{\ell\ell'},\hat{h}_{\ell\ell'},\tilde{\theta}_{\ell\ell'})$ for an observation $i\in I_{\ell'}$ by

$$\frac{\partial b_j}{\partial x}(X_i,\hat{g}_{\ell\ell'}(Z_i)).$$

This derivative can be found analytically for each atom. We can use this to obtain an automatic estimator of α_{02}.

To construct $\hat{D}_{1\ell}$, we need to estimate $D_1(W_i,c_k|\hat{g}_{\ell\ell'},\hat{h}_{\ell\ell'},\hat{\alpha}_{2\ell\ell'},\tilde{\theta}_{\ell\ell'})$ for an observation $i\in I_{\ell'}$ and an arbitrary atom c_k in a dictionary. The first term, $D_{11}(D_i,Z_i)$, can be estimated by

$$\frac{\hat{h}_{\ell\ell'}(X_i+s_n,\hat{g}_{\ell\ell'}(Z_i)+t_n) - \hat{h}_{\ell\ell'}(X_i,\hat{g}_{\ell\ell'}(Z_i))}{t_n s_n}.$$

To estimate the second term we can use equation (4.7), replacing $\hat{V}_{i\ell\ell'}$ by $\hat{g}_{\ell\ell'}(Z_i)$. These are the ingredients to build an automatic estimator for α_{01}.

Estimation of the APE greatly simplifies in a partly linear model. We can propose it taking advantage of the DR property w.r.t. the second step. ■

Appendix A Proofs of the results

Proof of Lemma 2.1: Applying the chain rule several times to $d\bar{m}(g(F_\tau), h(F_\tau, g(F_\tau)), \theta)/d\tau$, we have that:

$$\frac{d}{d\tau}\bar{m}(g(F_\tau), h(F_\tau, g(F_\tau)), \theta) = \frac{d}{d\tau}\bar{m}(g(F_\tau), h_0, \theta) + \frac{d}{d\tau}\bar{m}(g_0, h(F_\tau, g(F_\tau)), \theta).$$

Then, using the chain rule again:

$$\begin{aligned}\frac{d}{d\tau}\bar{m}(g_0, h(F_\tau, g(F_\tau)), \theta) = &\frac{d}{d\tau}\bar{m}(g_0, h(F_0, g(F_\tau)), \theta) \\ &+ \frac{d}{d\tau}\bar{m}(g_0, h(F_\tau, g_0), \theta).\end{aligned}$$

Combining the above equations leads to:

$$\begin{aligned}\frac{d}{d\tau}\bar{m}(g(F_\tau), h(F_\tau, g(F_\tau)), \theta) = &\frac{d}{d\tau}\bar{m}(g(F_\tau), h_0, \theta) \\ &+ \frac{d}{d\tau}\bar{m}(g_0, h(F_0, g(F_\tau)), \theta) \\ &+ \frac{d}{d\tau}\bar{m}(g_0, h(F_\tau, g_0), \theta).\end{aligned} \tag{A.1}$$

Now, note that by the chain rule we have that:

$$\begin{aligned}&\frac{d}{d\tau}\bar{m}(g(F_\tau), h_0, \theta) + \frac{d}{d\tau}\bar{m}(g_0, h(F_0, g(F_\tau)), \theta) \\ &\qquad = \frac{d}{d\tau}\bar{m}(g(F_\tau), h(F_0, g(F_\tau)), \theta).\end{aligned}$$

Hence the first two terms in equation (A.1) equal the derivative of $\bar{m}(g(F_\tau), h(F_0, g(F_\tau)), \theta)$. ■

Proof of Proposition 3.1: For the (differentiable) path $\tau \mapsto h(F_\tau, g_0)$, Assumption (A1) implies

$$\frac{d}{d\tau}\bar{m}(g_0, h(F_\tau, g_0), \theta) = \frac{d}{d\tau}\mathbb{E}[D_2(W, h(F_\tau, g_0)].$$

This gives the linearization (LIN).

To find the shape of the IF, note that $\mathbb{E}[D_2(W, h)]$ is a linear and continuous functional in $L_2(X, V)$, a Hilbert space of square-integrable functions. Thus, by the Riesz Representation Theorem, there exists a r_2 such that $\mathbb{E}[D_2(W, h)] = \mathbb{E}[r_2(X, V)h(X, V)]$, with $V \equiv \varphi(D, Z, g_0)$. Therefore:

$$\frac{d}{d\tau}\bar{m}(g_0, h(F_\tau, g_0), \theta) = \frac{d}{d\tau}\mathbb{E}[r_2(X, V)h(F_\tau, g_0)(X, V)],$$

where $h(F, g)(x, v)$ denotes $h(F, g)$ evaluated at (x, v). This is Assumption 1 in Ichimura and Newey (2022). Since Assumption 2 in that paper is satisfied in our setup, Proposition 1 in

Ichimura and Newey (2022) gives: $\phi_2(w, h_0, \alpha_{02}, \theta) = \alpha_{02}(d,z)\{y - h_0(x, \varphi(d,z,g_0))\}$. The parameter α_{20} is the L_2-projection of r_2 onto $\Gamma_2(g_0)$:

$$\alpha_{20} = \arg \min_{\alpha \in \Gamma_2(g_0)} \mathbb{E}[(r_2(X, \varphi(D,Z,g_0)) - \alpha(D,Z))^2]. \tag{A.2}$$

We now show that, necessarily, $\alpha_{02} \in L_2(g_0) \equiv \{(d,z) \mapsto \delta(x, \varphi(d,z,g)) \colon \delta \in L_2(X,V)\}$. Note that $r_2 \in L_2(g_0)$. Moreover, since $L_2(g_0)$ is a linear and closed subspace of $L_2(D,Z)$, by Luenberger (1997, Th. 1 in Sec. 3.4), for every $\alpha \in \Gamma_2(g_0)$ we have the decomposition $\alpha = m + m^\perp$, with $m \in L_2(g_0)$ and $m^\perp \in L_2(g_0)^\perp$, the orthogonal complement of $L_2(g_0)$. Therefore, for every $\alpha \in \Gamma_2(g_0)$,

$$\|r_2 - \alpha\|^2 = \|r_2 - m - m^\perp\|^2 = \|r_2 - m\|^2 + \|m^\perp\|^2 \geq \|r_2 - m\|^2.$$

Note that $\|\delta\|^2 = \mathbb{E}[\delta(D,Z)^2]$ for every $\delta \in L_2(D,Z)$. The above result uses that $r_2 - m \in L_2(g_0)$ and Pitagoras' Theorem (Luenberger, 1997, Lemma 1 in Sec. 3.3). Since equality is achieved when $m^\perp = 0$, we have that $\|r_2 - \alpha\|^2$ is minimized for an $\alpha \in \Gamma_2(g_0) \cap L_2(g_0)$. This gives Point (IF). ■

Proof of Theorem 3.1: We compute $d\bar{m}(g(F_\tau), h_0, \theta)/d\tau$ and $d\bar{m}(g_0, h(F_0, g(F_\tau)), \theta)/d\tau$ separatelly and then add them according to equation (3.1). By Assumptions (A1) and (A2), using the Riesz Representation Theorem, we have that for the differentiable paths $\tau \mapsto g(F_\tau)$ and $\tau \mapsto h(F_0, g(F_\tau))$:

$$\frac{d}{d\tau}\bar{m}(g(F_\tau), h_0, \theta) = \frac{d}{d\tau}\mathbb{E}[D_{11}(W, g(F_\tau))] = \frac{d}{d\tau}\mathbb{E}[r_1(Z) g(F_\tau)(Z)] \tag{A.3}$$

and, being $V \equiv \varphi(D, Z, g_0)$,

$$\frac{d}{d\tau}\bar{m}(g_0, h(F_0, g(F_\tau)), \theta) = \frac{d}{d\tau}\mathbb{E}[D_2(W, h(F_0, g(F_\tau)))] = \frac{d}{d\tau}\mathbb{E}[r_2(X,V) h(F_0, g(F_\tau))(X,V)]. \tag{A.4}$$

In these equations, $g(F)(z)$ means $g(F)$ evaluated at z, and $h(F,g)(x,v)$ means $h(F,g)$ evaluated at (x,v).

We now proceed as in Hahn and Ridder (2013, Lma. 1). For any function $\delta \in \Gamma_2(g(F_\tau)) \cap L_2(g_0)$, we have that

$$\mathbb{E}[\delta(X, \varphi(D, Z, g(F_\tau))) \cdot \{Y - h(F_0, g(F_\tau))(X, \varphi(D, Z, g(F_\tau)))\}] = 0.$$

This is the orthogonality condition that defines $h(F_0, g(F_\tau))$, as equation (2.2) defines h_0. Taking derivatives in the above equation leads to:

$$\begin{aligned} \frac{d}{d\tau}\mathbb{E}[\delta_2(X,V) h(F_0, g(F_\tau))(X,V)] = &-\frac{d}{d\tau}\mathbb{E}[\delta_2(X,V) h_0(X, \varphi(D, Z, g(F_\tau)))] \\ &+ \frac{d}{d\tau}\mathbb{E}\left[\delta_2(X, \varphi(D, Z, g(F_\tau))) \cdot (Y - h_0(X,V))\right]. \end{aligned} \tag{A.5}$$

A final step is needed to connect equation (A.4) with the above result. To perform it, we use Assumption (A3) in two directions. First, since $\alpha_{02} \in \Gamma_2(g_0) \cap L_2(g_0)$, we have that $\alpha_{02}(\cdot, \varphi(\cdot, \cdot, g(F_\tau))) \in \Gamma_2(g(F_\tau)) \cap L_2(g_0)$. We can then apply equation (A.5) to α_{02}. Moreover, since $h(F_0, g(F_\tau))(\cdot, \varphi(\cdot, \cdot, g(F_\tau))) \in \Gamma_2(g(F_\tau))$, we also have that $h(F_0, g(F_\tau))(\cdot, \varphi(\cdot, \cdot, g_0)) \in \Gamma_2(g_0)$. This means that, in equation (A.4), we can dismiss the component of r_2 that is orthogonal to $\Gamma_2(g_0)$. Then, we can write $d\mathbb{E}[\alpha_{02}(X,V)h(F_0, g(F_\tau))(X,V)]/d\tau$ as RHS in equation (A.4). Combining this with equation (A.5):

$$\begin{aligned}\frac{d}{d\tau}\bar{m}(g_0, h(F_0, g(F_\tau)), \theta) &= \frac{d}{d\tau}\mathbb{E}[\alpha_{02}(X,V)h(F_0, g(F_\tau))(X,V)] \\ &= -\frac{d}{d\tau}\mathbb{E}[\alpha_{02}(X,V)h_0(X, \varphi(D, Z, g(F_\tau)))] \\ &+ \frac{d}{d\tau}\mathbb{E}\left[\alpha_{02}(X, \varphi(D, Z, g(F_\tau))) \cdot (Y - h_0(X,V))\right].\end{aligned} \tag{A.6}$$

Under Assumption (A4), the term in the second row can be linearized in $g(F_\tau)$ as

$$\begin{aligned}\frac{d}{d\tau}\mathbb{E}[\alpha_{02}(X,V)h_0(X, \varphi(D, Z, g(F_\tau)))] &= \mathbb{E}\left[\frac{d}{d\tau}\{\alpha_{02}(X,V)h_0(X, \varphi(D, Z, g(F_\tau)))\}\right] \\ &= \mathbb{E}\left[\alpha_{02}(X,V)\frac{\partial h_0}{\partial v}(X,V)\frac{d}{d\tau}\varphi(D, Z, g(F_\tau))\right] \\ &= \mathbb{E}\left[\alpha_{02}(X,V)\frac{\partial h_0}{\partial v}(X,V)\frac{d}{d\tau}D_\varphi g(F_\tau)(D,Z)\right] \\ &= \frac{d}{d\tau}\mathbb{E}\left[\alpha_{02}(X,V)\frac{\partial h_0}{\partial v}(X,V)D_\varphi g(F_\tau)(D,Z)\right],\end{aligned}$$

where $D_\varphi g(d,z)$ denotes $D_\varphi g$ evaluated at (d,z). We have assumed that derivatives and expectations can be interchanged (we may impose some regularity conditions on H such that this is possible). We can equivalently linearize the term in the third row of equation (A.6) to get

$$\frac{d}{d\tau}\mathbb{E}\left[\alpha_{02}(X, \varphi(D, Z, g(F_\tau))) \cdot (Y - h_0(X,V))\right] = \frac{d}{d\tau}\mathbb{E}\left[(Y - h_0(X,V))\frac{\partial \alpha_{02}}{\partial v}(X,V)D_\varphi g(F_\tau)(D,Z)\right].$$

Pluging in these results back in equation (A.6):

$$\begin{aligned}\frac{d}{d\tau}\bar{m}(g_0, h(F_0, g(F_\tau)), \theta) = \frac{d}{d\tau}\mathbb{E}\Bigg[\Bigg\{&-\alpha_{02}(X,V)\frac{\partial h_0}{\partial v}(X,V) \\ &+(Y - h_0(X,V))\frac{\partial \alpha_{02}}{\partial v}(X,V)\Bigg\} D_\varphi g(F_\tau)(D,Z)\Bigg] \\ = \mathbb{E}\left[\frac{\partial}{\partial v}\left\{\alpha_{02}(X,v) \cdot (Y - h_0(X,v))\right\}\Big|_{v=V} D_\varphi g(F_\tau)(D,Z)\right].&\end{aligned} \tag{A.7}$$

Since D_φ is linear in g, the function inside the expectation in the RHS is linear in g. We now

use equation (3.1) to combine the results in equations (A.3) and (A.7). This gives:

$$\frac{d}{d\tau}\bar{m}(g(F_\tau), h(F_0, g(F_\tau)), \theta) = \frac{d}{d\tau}\mathbb{E}\Bigg[D_{11}(W, g(F_\tau)) + \frac{\partial}{\partial v}\left\{\alpha_{02}(X, v)\cdot(Y - h_0(X, v))\right\}\Big|_{v=V} D_\varphi g(F_\tau)(D, Z)\Bigg],$$

which gives the linearization result of the Theorem (LIN).

To find the shape of the IF, note that the adjoint D_φ^* of D_φ is defined by the equation $\mathbb{E}[\delta(D, Z)D_\varphi g(D, Z)] = \mathbb{E}[D_\varphi^*\delta(Z)g(Z)]$. Therefore, by the Law of Iterated Expectations in equation (A.7), noting that $V \equiv \varphi(D, Z, g_0)$ is a function of (D, Z):

$$\begin{aligned}\frac{d}{d\tau}\bar{m}(g_0, h(F_0, g(F_\tau)), \theta) &= \frac{d}{d\tau}\mathbb{E}\left[\mathbb{E}\left[\frac{\partial}{\partial v}\left\{\alpha_{02}(X, v)\cdot(Y - h_0(X, v))\right\}\Big|_{v=V} D_\varphi g(F_\tau)(D, Z)\,\middle|\, D, Z\right]\right] \\ &= \mathbb{E}[\nu(D, Z)D_\varphi g(F_\tau)(D, Z)] = \mathbb{E}[D_\varphi^*\nu(Z)g(F_\tau)(Z)],\end{aligned}$$

with

$$\nu(d, z) \equiv \frac{\partial}{\partial v}\left\{\alpha_{02}(x, v)\cdot(\mathbb{E}[Y|D = d, Z = z] - h_0(x, v))\right\}\Big|_{v=\varphi(d,z,g_0)}.$$

Again, we can use equation (3.1) to combine this last result with that in equation (A.3):

$$\frac{d}{d\tau}\bar{m}(g(F_\tau), h(F_0, g(F_\tau)), \theta) = \frac{d}{d\tau}\mathbb{E}[\{r_1(Z) + D_\varphi^*\nu(Z)\}g(F_\tau)(Z)].$$

This is Assumption 1 in Ichimura and Newey (2022). Since Assumption 2 in that paper is satisfied in our setup, Proposition 1 in Ichimura and Newey (2022) gives the shape of the IF: $\phi_1(w, g_0, \alpha_{01}, \theta) = \alpha_{01}(z)\cdot\{d - g_0(z)\}$. The parameter α_{10} is the L_2-projection:

$$\alpha_{10} = \arg\min_{\alpha\in\Gamma_1}\mathbb{E}[(\tilde{\nu}(Z) - \alpha(Z))^2], \tag{A.8}$$

where $\tilde{\nu} = r_1 + D_\varphi^*\nu$. ■

References

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

Blundell, R. and Powell, J. L. (2003). Endogeneity in nonparametric and semiparametric regression models. *Econometric society monographs*, 36:312–357.

Blundell, R. W. and Powell, J. L. (2004). Endogeneity in semiparametric binary response models. *The Review of Economic Studies*, 71(3):655–679.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2022a). Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535.

Chernozhukov, V., Newey, W. K., and Singh, R. (2022b). Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027.

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

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

Escanciano, J. C., Jacho-Chávez, D. T., and Lewbel, A. (2014). Uniform convergence of weighted sums of non and semiparametric residuals for estimation and testing. *Journal of Econometrics*, 178:426–443.

Hahn, J. and Ridder, G. (2013). Asymptotic variance of semiparametric estimators with generated regressors. *Econometrica*, 81(1):315–340.

Hahn, J. and Ridder, G. (2019). Three-stage semi-parametric inference: Control variables and differentiability. *Journal of econometrics*, 211(1):262–293.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161.

Heckman, J. J., Ichimura, H., and Todd, P. (1998). Matching as an econometric evaluation estimator. *The review of economic studies*, 65(2):261–294.

Ichimura, H. and Lee, L.-F. (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*, pages 3–49.

Ichimura, H. and Newey, W. K. (2022). The influence function of semiparametric estimators. *Quantitative Economics*, 13(1):29–61.

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

Luenberger, D. G. (1997). *Optimization by vector space methods*. John Wiley & Sons.

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

Mammen, E., Rothe, C., and Schienle, M. (2016). Semiparametric estimation with generated covariates. *Econometric Theory*, 32(5):1140–1177.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6):1349–1382.

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

Newey, W. K. and Robins, J. M. (2017). Cross-fitting and fast remainder rates for semiparametric estimation. *CEMMAP Working paper WP41/17*.

Pérez-Izquierdo, T. J. (2022). The determinants of counterfactual identification in the binary choice model with endogenous regressors. Unpublished manuscript.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954.

Rothe, C. (2009). Semiparametric estimation of binary response models with endogenous regressors. *Journal of Econometrics*, 153(1):51–64.

Sasaki, Y. and Ura, T. (2021). Estimation and inference for policy relevant treatment effects. *Journal of Econometrics*.

Stock, J. H. (1989). Nonparametric policy analysis. *Journal of the American Statistical Association*, 84(406):567–575.

Stock, J. H. (1991). Nonparametric policy analysis: an application to estimating hazardous waste cleanup benefits. *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, pages 77–98.