Automatic Locally Robust Estimation with Generated Regressors[*]

Juan Carlos Escanciano
Universidad Carlos III de Madrid

Telmo J. Pérez-Izquierdo
BCAM - Basque Center for Applied Mathematics

June 23, 2023

Abstract

Many economic and causal parameters depend on generated regressors. Examples include structural parameters in models with endogenous variables estimated by control functions and in models with sample selection, treatment effect estimation with propensity score matching, and marginal treatment effects. 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 Double-Robust estimators for the Counterfactural 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 MICIN/AEI/10.13039/501100011033, grant CEX2021-001181-M, Comunidad de Madrid, grants EPUC3M11 (V PRICIT) and H2019/HUM-5891, and grant 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), Average Partial Effects (APE) in sample selection models (Das et al., 2003), Propensity Score Matching (Heckman et al., 1998), and Marginal Treatment Effects using Local Instrumental Variables (Heckman and Vytlacil, 2005). 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/Debiased estimators of and inference on structural parameters in such models.

We extend Chernozhukov et al. (2022a)'s results to build debiased moment conditions in the presence of nonparametric/semiparametric generated regressors. By applying the chain rule, the debiasing correction term can be decomposed into one accounting for a first step, which is used to generate a regressor, and the term accounting for the second step, where the outcome variable is regressed onto the generated variable (among other covariates). Chernozhukov et al. (2022a) construct debiased moment conditions which account for this second step (with no generated regressor). Our paper provides the additional correction term that accounts for (i) the plug-in of generated regressors in the moment condition and (ii) the effect of generated regressor on estimation of the second step.

Each of the two correction terms depends on additional nuisance parameters. Under a linearization assumption (as in Ichimura and Newey, 2022; Newey, 1994), we show how the additional nuisance parameters in the correction term can be estimated without knowing their specific analytic shape. This process is called automatic estimation (see Chernozhukov et al., 2022b). Automatic estimation is particularly well motivated in the case of generated regressors, where the nuisance parameters in the correction term take complex shapes (see, for instance, Escanciano et al., 2014; Hahn and Ridder, 2013; Mammen et al., 2016).

As an application of our methods we propose novel Automatic Locally Robust 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 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 Double-Robust estimators are, however, not Locally Robust to the generated regressors in general. To construct fully Locally Robust 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 Double-Robust moment conditions are robust to this model being misspecified.

The finite sample performance of the proposed estimator is evaluated through Monte Carlo simulations. We use Lasso with different dictionaries (linear, quadratic, and one including interactions) to fit the first and second step parameters, as well as the nuisance parameters in the first and second step correction terms. Our result confirm that the plug-in estimator is asymptotically biased (see also Chernozhukov et al., 2018, 2022a; Escanciano and Terschuur, 2022). Correcting the moment condition for the second step estimation reduces de bias, but a first step correction must be added to totally remove it.

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). Our results can be readily extended to allow for profiling, as in Mammen et al. (2016) and Hahn et al. (2022). Furthermore, we contribute to this literature by allowing for machine learning generated regressors.

The second literature we build on is the more recent literature on Locally Robust/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 Locally Robust 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.

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 Locally Robust moments with generated regressors. In Section 4, we provide the details for Debiased Locally Robust GMM estimation. A summary of the estimation algorithm is given in Section 4.2. The asymptotic theory for the proposed estimator is developed in Section 5. Monte Carlo simulations are presented in Section 6. Section 7 concludes.

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 $g_0 \in \Delta_1$, for Δ_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

$$\mathbb{E}[\delta_1(Z)(D - g_0(Z))] = 0 \text{ for all } \delta_1 \in \Delta_1. \tag{2.1}$$

This setting covers parametric, semiparametric, and non-parametric first steps. For example, when $\Delta_1 = L_2(Z)$, we have $g_0(Z) = \mathbb{E}[D|Z]$.

Next, 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

$$\mathbb{E}[\delta_2(D, Z)(Y - h_0(X, V))] = 0 \text{ for all } \delta_2 \in \Delta_2(g_0), \tag{2.2}$$

where $\Delta_2(g_0)$ is a linear and closed subspace of $L_2(D, Z)$. The set $\Delta_2(g_0)$ may depend on the fist step parameter g_0. In some settings, $\Delta_2(g_0)$ includes only functions of X and the generated regressor V. That is, $\Delta_2(g_0)$ includes functions with the following shape: $\delta_2(D, Z) = \delta(X, \varphi(D, Z, g_0))$ for $\delta \in \Delta$, a linear and closed subspace of $L_2(X, V)$. For instance, Hahn and Ridder (2013) and Mammen et al. (2016) consider cases where the second step is a non-parametric regression of Y on (X, V). In that case, $\Delta_2(g) = L_2(g)$, with

$$L_2(g) \equiv \{(d, z) \mapsto \delta(x, \varphi(d, z, g)) \colon \delta \in L_2(X, V)\} \subseteq L_2(D, Z).$$

The above set plays a key role, since $h_0(X, \varphi(D, Z, g_0))$, understood as a function of (D, Z), is an element of $L_2(g_0)$

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

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

[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, we omit 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.

Here we assume that θ_0 is identified by these moments, i.e. that θ_0 is the unique solution to $\mathbb{E}[m(W, g_0, h_0, \theta)] = 0$ over $\theta \in \Theta$.

Our result allows for an arbitrary number of parameters $\theta \in \mathbb{R}^{\dim(\theta)}$ and moment conditions $m\colon \mathbb{R}^{\dim(W)} \times L_2(Z) \times L_2(X, V) \times \Theta \to \mathbb{R}^{\dim(m)}$, with $\dim(m) \geq \dim(\theta) \geq 1$. To ease the exposition, most results are derived for the one dimensional case. Extensions are straightforward by requiring each assumption to hold componentwise. Incorporating multiple variables D and Y to our setup is also simple.

We illustrate the notation and concepts with two general running examples. Common to both problems is that the evaluating moment condition $m(w, g, h, \theta)$ at a certain point $w = (y, d, z)$ requires the whole shape of the second step parameter h, and not only its value $h(x, \varphi(d, z, g))$ at (d, z) (cf. Hahn and Ridder, 2013).

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 $\Delta_2(g) = L_2(g)$.

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

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

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

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

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 $\mathbb{E}[\eta(X, \mathrm{ASF}_0(X))]$ for a function η with domain in an Euclidean space. We will propose below a novel Double-Robust 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 $D = 1\left[g_0\left(Z\right) - U \geq 0\right]$, where U is uniformly distributed in $\left[0, 1\right]$. The unobserved errors ε and U, though independent of Z, are correlated with each other (selection on unobservables). In this example, $V = g_0\left(Z\right) = \mathbb{E}(D|Z)$. Then, it can be shown that

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

This setting provides a nonparametric extension of the classical model of Heckman (1979), where $H(X, \varepsilon) = X'\beta_0 + \varepsilon$, $g_0\left(Z\right) = Z'\gamma_0$, 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

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

This parameter is covered by Proposition 5 in Hahn and Ridder (2019). However, the authors do not consider Locally Robust estimation. In Appendix A we propose a novel Locally Robust estimator for the APE which (i) is Double-Robust to the second step and (ii) allows for ML first and second step estimators. ■

Remark 2.1 (Profiling) Our results can be extended to allow for profiling as in Mammen et al. (2016). That is, we may consider that the second step nuisance parameter depends on θ: $h_0(x, v, \theta)$ is the solution in h of $\mathbb{E}[\delta_2(D, Z)(Y - h(X, V, \theta))] = 0$ for all $\delta_2 \in \Delta_2(g_0, \theta)$. Note that then $h_0(\cdot, \theta)$ is the projection of Y onto $\Delta_2(g_0, \theta)$.

For instance, if h_0 is the conditional expectation given some transformation of (D, Z), denoted $T(D, Z, g, \theta)$ as in Mammen et al. (2016), then one would take $\Delta_2(g, \theta) \equiv L_2(T)$. In models with an index restriction, $T(D, Z, g, \theta) = (\theta' X, \varphi(D, Z, g))$, where X is a subvector of (D, Z).

A modification of equation (2.2) also allows to cover partly linear models, as in the examples discussed in Hahn et al. (2022). One may replace Y in (2.2) by an arbitrary transformation $\xi(W, \theta)$ to get that $h_0(\cdot, \theta)$ is the solution in h to

$$\mathbb{E}[\delta_2(D, Z)(\xi(W, \theta) - h(X, V, \theta))] = 0 \text{ for all } \delta_2 \in \Delta_2(g_0, \theta).$$

For partly linear models, one can take $X = \emptyset$ and $\xi(W, \theta) = Y - \theta' \tilde{X}$ for $\tilde{X}$ (another) subvector of (D, Z). In this case, $X = \emptyset$ indicates that only the generated regressor V enters non-linearly in the structural equation. Thus, h_0 is the projection of $Y - \theta' \tilde{X}$ onto $L_2(V) = \Delta_2(g_0, \theta)$.

2.1 Orthogonal Moment Functions with Generated Regressors

We follow Chernozhukov et al. (2022a, henceforth, CEINR) for the construction of Locally Robust-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 $g(F)(z) = \mathbb{E}_F[D|Z = z]$ 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 $\mathbb{E}_F[|D|]$ is finite. We assume that $g(F)$ is identified as the solution in g to

$$\mathbb{E}_F[\delta_1(Z)(D - g(Z))] = 0 \text{ for all } \delta_1 \in \Delta_1.$$

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

$$\mathbb{E}_F\left[\delta_2(D,Z)\{Y-h(X,\varphi(D,Z,g))\}\right]=0 \text{ for all } \delta_2\in\Delta_2(g).$$

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

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

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 Mises (1947), Hampel (1974), and Huber (1981), equation (2.3) 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,*

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

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 (see Figure 1). 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 $h(F_\tau, g_0)$.

$$\begin{tikzcd}
\tau \arrow[r] & F_\tau \arrow[r] \arrow[rr, bend left] & g(F_\tau) \arrow[r, "(E)"] \arrow[dr, "(D)"] & h(F_\tau, g(F_\tau)) \arrow[d, "(2S)"] \\
 & & & \bar{m}(g(F_\tau), h(F_\tau, g(F_\tau)), \theta)
\end{tikzcd}$$

Figure 1 The effect of a deviation F_τ on the moment condition. (2S) represents the second step effect. (D) represents the direct effect of the first step. The path (E)-(2S) represents the estimation effect of the first step.

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, for all θ and H:

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

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

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 Δ_1 and $\Delta_2(g_0)$ are linear, the above equations mean that, for all $\theta \in \Theta$,

$$\begin{aligned} &\frac{d}{d\tau}\mathbb{E}[\psi_1(W,g_0+\tau\delta_1,\alpha_{10},\theta)] = 0 \text{ for all } \delta_1 \in \Delta_1 \text{ and} \\ &\frac{d}{d\tau}\mathbb{E}[\psi_2(W,h_0+\tau\delta_2,\alpha_{20},\theta)] = 0 \text{ for all } \delta_2 \in \Delta_2(g_0). \end{aligned} \tag{2.7}$$

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.

Remark 2.2 (Profiling) The previous results and those of Section 3.1 readily extend to allow for profiling (see Remark 2.1). Indeed, the IFs in equations (2.5)-(2.6) are defined for a "fixed" parameter θ. Both equations are the Gateaux derivative of $\bar{m}(g,h(F_0,g),\theta)$ and $\bar{m}(g_0,h,\theta)$ with respect to g and h, respectively, for each value of θ.

To be more precise, we discuss each IF separately. For the first-step IF, note that the derivative is determined by how $\bar{m}(g,h(F_0,g),\theta)$ depends on g. It is thus straightforward to allow for the second step to depend on θ and study, instead, the derivative of $\bar{m}(g,h(F_0,g,\theta),\theta)$ w.r.t. g. In turn, the second-step IF will also capture that the derivative of $\bar{m}(g_0,h(F_\tau,g_0,\theta),\theta)$ depends on how the paths $\tau \mapsto h(F_\tau,g_0,\theta)$ vary with θ.

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 (B.8) in the Appendix, the examples in Section 3.1, and Hahn and Ridder, 2013). Therefore, the plug-in estimator for $\hat{\alpha}$ may be cumbersome to compute.

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.

This section shows that, under some assumptions, the correction term takes to form:

$$\phi(w, g_0, h_0, \alpha_0, \theta) = \underbrace{\alpha_{01}(z) \cdot [d - g_0(z)]}_{=\phi_1(w,g_0,\alpha_{01},\theta)} + \underbrace{\alpha_{02}(x, \varphi(d, z, g_0)) \cdot [y - h_0(x, \varphi(d, z, g_0))]}_{=\phi_2(w,h_0,\alpha_{02},\theta)}. \tag{3.1}$$

That is, each correction term is build by multiplying the nuisance parameter by each step's prediction error. This is the first step to build the automatic estimators for α_{01} and α_{02}. The other ingredient is a consistent estimator of the linearization of the moment condition with respect to each parameter (g for the first step and h for the second). Section 3.1 provides the formal development.

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 immediately 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 $\alpha_{02} \in \Delta_2(g_0) \cap L_2(g_0)$ 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. 5) 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 compute the analytic shape of the correction term nuisance parameter 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

$$\begin{aligned}\mathbb{E}[D_2(W, h)] &= \int h(x^*, v) f^*(x^*) f_0^v(v) dx^* dv = \int \frac{f^*(x^*) f_0^v(v)}{f_0^{xv}(x^*, v)} h(x^*, v) f_0^{xv}(x^*, v) dx^* dv \\ &= \mathbb{E}[\alpha_{02}(X, V) h(X, V)],\end{aligned}$$

with $\alpha_{02}(x, v) \equiv f^*(x^*) f_0^v(v) / f_0^{xv}(x^*, v)$. Note that, even if we have found the nuisance parameter α_{02}, 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 α_{02} 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: $\Delta_1 = L_2(Z)$ and $\Delta_2(g) = \{\delta(x, g(z)) \colon \delta \in L_2(X, V)\}$ (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 α_{02} 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:

$$\begin{aligned}\frac{d}{d\tau}\bar{m}(g(F_\tau), h(F_0, g(F_\tau)), \theta) = &\frac{d}{d\tau}\bar{m}(g(F_\tau), h_0, \theta) \\ &+ \frac{d}{d\tau}\bar{m}(g_0, h(F_0, g(F_\tau)), \theta).\end{aligned} \tag{3.2}$$

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 $g \in \Delta_1$ and $\delta \in L_2(X, V)$, we have that $\delta(\cdot, \varphi(\cdot, \cdot, g)) \in \Delta_2(g) \Leftrightarrow \delta(\cdot, \varphi(\cdot, \cdot, g_0)) \in \Delta_2(g_0)$.*

(A4) *h_0 and α_{02} are differentiable w.r.t. v. Moreover, the function $\varphi(d, z, g)$, understood as a mapping $g \mapsto \varphi(d, z, g)$ 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*

$$D_1(w,g) \equiv D_{11}(w,g) + \frac{\partial}{\partial v}\left[\alpha_{02}(x,v)(y - h_0(x,v))\right] \cdot D_\varphi g, \tag{3.3}$$

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 $\alpha_{01} \in \Delta_1$ 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 $\Delta_2(g)$ (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 $\Delta_2(g) = \{(d,z) \mapsto \delta(x, \varphi(d,z,g)) \colon \delta \in \Delta\}$, for Δ a linear subspace of $L_2(X,V)$ satisfies the assumption. When $\Delta = L_2(X,V)$ (i.e., $\Delta_2(g) = L_2(g)$), the second step is a non-parametric regression on X and the generated regressor. One can also take $\Delta = \{\beta' x + \eta(v) \colon \beta \in \mathbb{R}^{\dim(X)}, \eta \in L_2(V)\}$ 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 $\Delta_2(g) = L_2(D,Z)$, 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 equation (3.3)). The first term is standard and 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. 12) 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 $\Delta_2(g) = L_2(D,Z)$ 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 (3.3), where the linearization of the first step effect is defined. Recall that $D_\varphi g = -g$. Then, for the CASF, equation (3.3) 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. 12) 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 (3.3) 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 $(\delta_1, \delta_2) \in \Delta_1 \times \Delta_2(g_0)$. 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 $\delta_2 \in \Delta_2(g_0)$,

$$\frac{d}{d\tau}\bar{m}(g_0, h_0 + \tau\delta_2, \theta) + \frac{d}{d\tau}\mathbb{E}[\phi_2(W, h_0 + \tau\delta_2, \alpha_{20}, \theta)] = 0. \tag{3.4}$$

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

$$\frac{d}{d\tau}\bar{m}(g_0, h_0 + \tau\delta_2, \theta) = \frac{d}{d\tau}\mathbb{E}[D_2(W, h_0 + \tau\delta_2)] = \mathbb{E}[D_2(W, \delta_2)], \text{ for any } \delta_2 \in \Delta_2(g_0). \tag{3.5}$$

Moreover, Proposition 3.1 gives us that $\phi_2 = \alpha_{02}(y - h_0)$. Thus, for any $\delta_2 \in \Delta_2(g_0)$,

$$\frac{d}{d\tau}\mathbb{E}[\phi_2(W, h_0 + \tau\delta_2, \alpha_{20}, \theta)] = -\mathbb{E}[\delta_2(D, Z)\alpha_{20}(X, V)]. \tag{3.6}$$

From equations (3.4)-(3.6), for each $\delta_2 \in \Delta_2(g_0)$,

$$\mathbb{E}[D_2(W,\delta_2)] - \mathbb{E}[\delta_2(D,Z)\alpha_{20}(X,V)] = 0, \text{ for each } \delta_2 \in \Delta_2(g_0). \tag{3.7}$$

Since $\mathbb{E}[D_2(W,\delta_2)]$ is a linear functional, we will have a Riesz Representer $r_2 \in Ł_2(D,Z)$ 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 $\Delta_2(g_0)$. However, the usefulness of the conditions in (3.7) is that they do not require finding the Riesz Representer. They are based on a linearization of the moment condition, which is generally easier to find.

We now assume that there is a dictionary $(b_j)_{j=1}^{\infty}$, with $b_j \in \Delta_2(g_0) \cap L_2(g_0)$, whose closed linear span is $\Delta_2(g_0) \cap L_2(g_0)$. That is, any function in $\Delta_2(g_0) \cap L_2(g_0)$ 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 $\boldsymbol{\rho}_J = (\rho_1, ..., \rho_J)'$.[2] We can now plug in $\mathbf{b}_J' \boldsymbol{\rho}_J$ into equation (3.7) 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:

$$\min_{\boldsymbol{\rho}_J \in \mathbb{R}^J} \left\{ -2\mathbb{E}[D_2(W,\mathbf{b}_J)']\boldsymbol{\rho}_J + \boldsymbol{\rho}_J'\mathbb{E}[\mathbf{b}_J(X,V)\mathbf{b}_J(X,V)']\boldsymbol{\rho}_J + \lambda \|\boldsymbol{\rho}_J\|_q^q \right\}. \tag{3.8}$$

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

[2] For a $d_1 \times d_2$ matrix A, A' denotes its transpose. In this respect, vectors are considered $d_1 \times 1$ matrices.

$\bar{m}(g, h(F_0, g), \theta)$ by $D_1(w, g)$ (see equation (3.3)). Again, we assume that there is a dictionary $(c_k)_{k=1}^{\infty}$ that spans Δ_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 $\boldsymbol{\beta}$:

$$\min_{\boldsymbol{\beta}_K \in \mathbb{R}^K} \left\{ -2\mathbb{E}[D_1(W, \mathbf{c}_K)']\boldsymbol{\beta}_K + \boldsymbol{\beta}_K' \mathbb{E}[\mathbf{c}_K(Z)\mathbf{c}_K(Z)']\boldsymbol{\beta}_K + \lambda \|\boldsymbol{\beta}_K\|_q^q \right\}, \tag{3.9}$$

where the tuning 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 (3.3)). 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 (3.3).

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). 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. Its shape is given in equation (3.1) (see also Proposition 3.1 and Theorem 3.1). The full estimation algorithm is summarized in Figure 2 (Section 4.2).

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 Chernozhukov et al., 2018).

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 datasets.

Following CEINR, debiased sample moment functions are

$$\hat{\psi}(\theta) \equiv \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \hat{\psi}_{i\ell}(\theta), \tag{4.1}$$

with,

$$\begin{aligned} \hat{\psi}_{i\ell}(\theta) &\equiv m(W_i, \hat{g}_\ell, \hat{h}_\ell, \theta) + \phi(W_i, \hat{g}_\ell, \hat{h}_\ell, \hat{\alpha}_\ell, \tilde{\theta}_\ell) \\ &= m(W_i, \hat{g}_\ell, \hat{h}_\ell, \theta) + \hat{\alpha}_{1\ell}(Z_i) \cdot (D_i - \hat{g}_\ell(Z_i)) + \hat{\alpha}_{2\ell}(X_i, \hat{V}_{i\ell}) \cdot (Y_i - \hat{h}_\ell(X_i, \hat{V}_{i\ell})), \end{aligned} \tag{4.2}$$

for $\hat{V}_{i\ell} \equiv \varphi(D_i, Z_i, \hat{g}_\ell)$. Note that the original moment condition is evaluated at θ. On the other hand, the initial estimators $\tilde{\theta}_\ell$ are used to construct the correction term ϕ (i.e., to estimate α_{01} and α_{02}).

In the general case where there is more than one moment condition, the correction term for each component of m is constructed following equation (4.2). This means that a different correctiont term must be estimated for each component of m (see Section 4.1 for the details about how to proceed with automatic estimation of each term). We use these debiased moment functions to construct the debiased GMM estimator:

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}}\, \hat{\psi}(\theta)' \hat{\Upsilon} \hat{\psi}(\theta), \tag{4.3}$$

where $\hat{\Upsilon}$ is a positive semi-definite weighting matrix of dimension $\dim(m) \times \dim(m)$. Under some conditions (see Section 5), the above estimator will be asymptotically normal with the usual GGM asymptotic variance. Indeed, as in Chernozhukov et al. (2022a), there is no need to account for estimation of (g_0, h_0) and $(\alpha_{01}, \alpha_{02})$ because of orthogonality of ψ.

To introduce the asymptotic variance, let

$$\begin{aligned} M &\equiv \mathbb{E}\left[\frac{\partial m}{\partial \theta}(W, g_0, h_0, \theta_0)\right] \quad \text{and} \\ \Psi &\equiv \mathbb{E}[\psi(W, g_0, h_0, \alpha_0, \theta_0)\psi(W, g_0, h_0, \alpha_0, \theta_0)'], \end{aligned}$$

where $\partial m / \partial \theta$ is the $\dim(m) \times \dim(\theta)$-dimensional Jacobian matrix. If $\hat{\Upsilon} \xrightarrow{P} \Upsilon$, the asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta_0)$ is $V \equiv (M'\Upsilon M)^{-1} M'\Upsilon'\Psi\Upsilon M(M'\Upsilon M)^{-1}$. A consistent estimator of the asymptotic variance can be build by replacing the terms in V by their sample analogs:

$$\hat{M} \equiv \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \frac{\partial m}{\partial \theta}(W_i, \hat{g}_\ell, \hat{h}_\ell, \tilde{\theta}_\ell) \text{ and} \tag{4.4}$$

$$\hat{\Psi} \equiv \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \hat{\psi}_{i\ell}(\tilde{\theta}_\ell) \hat{\psi}_{i\ell}(\tilde{\theta}_\ell)'. \tag{4.5}$$

As usual in GMM, a choice of $\hat{\Upsilon}$ that minimizes the asymptotic variance of $\hat{\theta}$ is $\hat{\Upsilon} = \hat{\Psi}^{-1}$. With that choice of a weighting matrix, the asymptotic variance can be estimated by $(\hat{M}'\hat{\Psi}^{-1}\hat{M})^{-1}$.

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

Example 1 (continuing from p. 15) 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 debiased sample moments.

We consider that the econometrician can sample from F^*. Let $(X_s^*)_{s=1}^S$ be a 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 (4.3), the debiased estimator for the CASF is

$$\hat{\theta} = \frac{1}{nS} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \sum_{s=1}^{S} \hat{h}_\ell(X_s^*, \hat{V}_{i\ell}) + \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \phi(W_i, \hat{g}_\ell, \hat{h}_\ell, \hat{\alpha}_\ell, \tilde{\theta}_\ell). \tag{4.6}$$

The next section develops an automatic estimator for the correction term ϕ. ■

Remark 4.1 (Profiling) Estimation of the correction term with a profiled-out h_0 is based on the initial estimators $\tilde{\theta}_\ell$. For each ℓ, the $\hat{h}_\ell(\cdot, \theta)$ is estimated for $\theta = \tilde{\theta}_\ell$. Additionally, we note that the estimators $\hat{h}_{\ell\ell'}(\cdot, \theta)$ and $\hat{h}_{\ell\ell'\ell''}(\cdot, \theta)$ (required for automatic estimation), that do not use observations in $I_\ell \cup I_{\ell'}$ or not in $I_\ell \cup I_{\ell'} \cup I_{\ell''}$, respectively, are estimated for initial estimators $\tilde{\theta}_{\ell\ell'}$ and $\tilde{\theta}_{\ell\ell'\ell''}$ satisfying those same properties.

To sum up, in the presence of profiling, the debiased moment functions in equation (4.2) are estimated by:

$$\hat{\psi}_{i\ell}(\theta) \equiv m(W_i, \hat{g}_\ell, \hat{h}_\ell(\cdot, \theta), \theta) + \phi(W_i, \hat{g}_\ell, \hat{h}_\ell, \hat{\alpha}_\ell(\cdot, \tilde{\theta}_\ell), \tilde{\theta}_\ell).$$

Also, the Jacobian of m w.r.t. θ must be extended:

$$M \equiv \mathbb{E}\left[\frac{\partial m}{\partial \theta}(W, g_0, h_0(\cdot, \theta), \theta) \Big|_{\theta = \theta_0} \right],$$

which may be estimated by

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

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 $I_\ell \cup I_{\ell'}$; and estimators $(\hat{g}_{\ell\ell'\ell''}$, $\hat{h}_{\ell\ell'\ell''}, \tilde{\theta}_{\ell\ell'\ell''})$ that use only observations not in $I_\ell \cup I_{\ell'} \cup I_{\ell''}$..

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 (3.8) that uses only observations not in I_ℓ. Recall that we have a dictionary $(b_j)_{j=1}^{\infty}$ that spans $\Delta_2(g_0) \cap L_2(g_0)$. 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

$$\hat{\boldsymbol{\rho}}_{J\ell} = \underset{\boldsymbol{\rho}_J \in \mathbb{R}^J}{\operatorname{argmin}} \left\{ -2\hat{D}_{2\ell}' \boldsymbol{\rho}_J + \boldsymbol{\rho}_J' \hat{B}_\ell \boldsymbol{\rho}_J + \lambda \|\boldsymbol{\rho}_J\|_q^q \right\}. \tag{4.7}$$

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

Example 1 (continuing from p. 20) 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

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

for each atom b_j in the dictionary.

We follow the same strategy as before and approximate D_2 by numerical integration. Let $(X_s^*)_{s=1}^S$ be a sample drawn from F^*. To construct the objective function to estimate $\hat{\boldsymbol{\rho}}_{J\ell}$, we approximate $D_2(W_i, b_j | \hat{g}_{\ell\ell'}, \hat{h}_{\ell\ell'}, \tilde{\theta}_{\ell\ell'})$, for an observation $i \in I_{\ell'}$, by

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

■

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 (3.9) 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 Δ_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

$$\hat{D}_{1\ell} \equiv \frac{1}{n - n_\ell} \sum_{\ell' \neq \ell} \sum_{i \in I_{\ell'}} D_1(W_i, \mathbf{c}_K | \hat{g}_{\ell\ell'}, \hat{h}_{\ell\ell'}, \hat{\alpha}_{2\ell\ell'}, \tilde{\theta}_{\ell\ell'}). \tag{4.8}$$

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. Indeed, we can define

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

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

Thus, $\hat{\boldsymbol{\rho}}_{J\ell\ell'}$ is given by the optimization problem in (4.7) 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 $h_0 \in \Delta_2(g_0) \cap L_2(g_0)$, 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 tuning parameter depending on the sample size. We propose to estimate $\partial h_0/\partial v(x, v)$ by

$$\frac{\partial\hat{h}_{\ell\ell'}}{\partial v}(x, v) \equiv \frac{\hat{h}_{\ell\ell'}(x, v + t_n) - \hat{h}_{\ell\ell'}(x, v)}{t_n}. \tag{4.9}$$

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 (4.8). 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

$$\hat{\boldsymbol{\beta}}_{K\ell} = \underset{\boldsymbol{\beta}_K \in \mathbb{R}^K}{\operatorname{argmin}} \left\{ -2\hat{D}_{1\ell}'\boldsymbol{\beta}_K + \boldsymbol{\beta}_K'\hat{C}_\ell\boldsymbol{\beta}_K + \lambda\|\boldsymbol{\beta}_K\|_q^q \right\}. \tag{4.10}$$

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

Example 1 (continuing from p. 21) 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

$$-\left(\frac{\partial \mathbf{b}_J}{\partial v}(X_i, \hat{V}_{i\ell\ell'})\right)' \hat{\boldsymbol{\rho}}_{J\ell\ell'} \cdot (Y_i - \hat{h}_{\ell\ell'}(X_i, \hat{V}_{i\ell\ell'})) + \mathbf{b}_J(X_i, \hat{V}_{i\ell\ell'})' \hat{\boldsymbol{\rho}}_{J\ell\ell'} \cdot \frac{\partial \hat{h}_{\ell\ell'}}{\partial v}(X_i, \hat{V}_{i\ell\ell'}). \tag{4.11}$$

Therefore, to estimate $D_{1\ell}$ according to equation (4.8), we have that, for $i \in I_{\ell'}$,

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

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 provide an illustration of our estimation algorithm. The inputs to the algorithm are cross-fit estimators of g_0 and h_0. An initial estimator of θ_0 must also be supplied. We note that one must provide a total of L estimators $(\hat{g}_\ell, \hat{h}_\ell, \tilde{\theta}_\ell)$ only using observations not in I_ℓ, $L(L-1)/2$ estimators $(\hat{g}_{\ell\ell'}, \hat{h}_{\ell\ell'}, \tilde{\theta}_{\ell\ell'})$ only using observations not in $I_\ell \cup I_{\ell'}$, and $L(L-1)(L-2)/6$ estimators $(\hat{g}_{\ell\ell'\ell''}, \hat{h}_{\ell\ell'\ell''}, \tilde{\theta}_{\ell\ell'\ell''})$ only using observations not in $I_\ell \cup I_{\ell'} \cup I_{\ell''}$.

Figure 2 provides a diagram showing how to compute the moment condition $\hat{\psi}_{i\ell}$ for an observation $i \in I_\ell$. The debiased moment condition is given in equation (4.2). To this equation, the diagram below adds the discussion in the above section, i.e., how to construct automatic estimators of the nuisance parameters (α_{01} and α_{02}) in the correction term. The arrows in the diagram indicate how to estimate each term.

Once the debiased moment condition $\hat{\psi}_{i\ell}$ is built, Automatic Debiased GMM estimation is conducted with the objective function in equation (4.3).

5 Asymptotic theory

This section gives general conditions for asymptotic normality of the automatic debiased GMM and conditions for consistent estimation of its asymptotic variance. The conditions are based in the mean-square consistency, small interaction of estimation biases and locally robust conditions

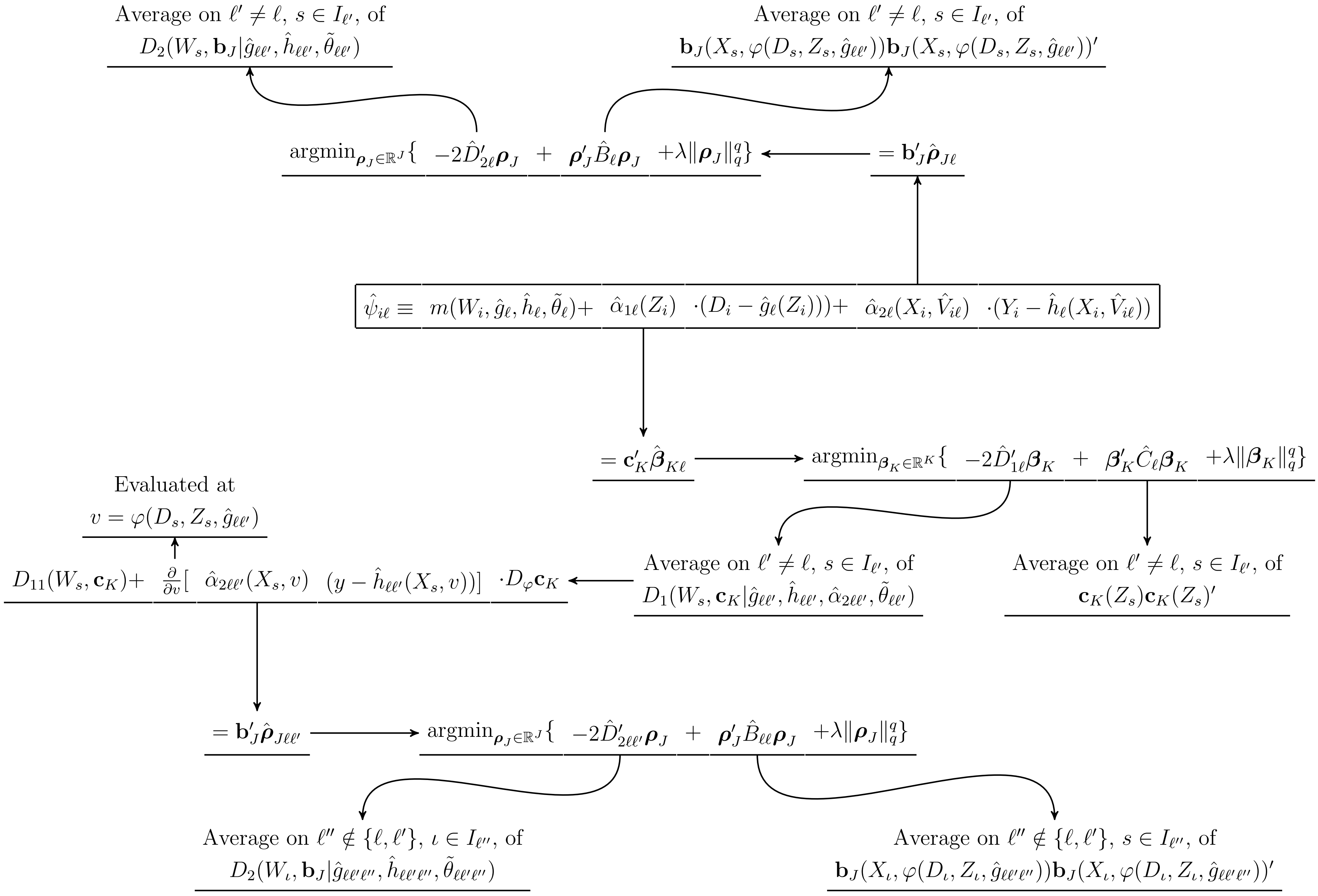
Average on $\ell' \neq \ell$, $s \in I_{\ell'}$, of $D_2(W_s, \mathbf{b}_J | \hat{g}_{\ell\ell'}, \hat{h}_{\ell\ell'}, \tilde{\theta}_{\ell\ell'})$
Average on $\ell' \neq \ell$, $s \in I_{\ell'}$, of $\mathbf{b}_J(X_s, \varphi(D_s, Z_s, \hat{g}_{\ell\ell'}))\mathbf{b}_J(X_s, \varphi(D_s, Z_s, \hat{g}_{\ell\ell'}))'$
$\operatorname{argmin}_{\boldsymbol{\rho}_J \in \mathbb{R}^J}\{ \ -2\hat{D}_{2\ell}'\boldsymbol{\rho}_J \ + \ \boldsymbol{\rho}_J'\hat{B}_\ell\boldsymbol{\rho}_J \ +\lambda\|\boldsymbol{\rho}_J\|_q^q\}$
$= \mathbf{b}_J'\hat{\boldsymbol{\rho}}_{J\ell}$
$\hat{\psi}_{i\ell} \equiv \ m(W_i, \hat{g}_\ell, \hat{h}_\ell, \tilde{\theta}_\ell)+ \ \hat{\alpha}_{1\ell}(Z_i) \ \cdot(D_i - \hat{g}_\ell(Z_i)))+ \ \hat{\alpha}_{2\ell}(X_i, \hat{V}_{i\ell}) \ \cdot(Y_i - \hat{h}_\ell(X_i, \hat{V}_{i\ell}))$
$= \mathbf{c}_K'\hat{\boldsymbol{\beta}}_{K\ell}$
$\operatorname{argmin}_{\boldsymbol{\beta}_K \in \mathbb{R}^K}\{ \ -2\hat{D}_{1\ell}'\boldsymbol{\beta}_K \ + \ \boldsymbol{\beta}_K'\hat{C}_\ell\boldsymbol{\beta}_K \ +\lambda\|\boldsymbol{\beta}_K\|_q^q\}$
Evaluated at $v = \varphi(D_s, Z_s, \hat{g}_{\ell\ell'})$
$D_{11}(W_s, \mathbf{c}_K)+ \ \frac{\partial}{\partial v}[\ \hat{\alpha}_{2\ell\ell'}(X_s, v) \ (y - \hat{h}_{\ell\ell'}(X_s, v))] \ \cdot D_\varphi \mathbf{c}_K$
Average on $\ell' \neq \ell$, $s \in I_{\ell'}$, of $D_1(W_s, \mathbf{c}_K | \hat{g}_{\ell\ell'}, \hat{h}_{\ell\ell'}, \hat{\alpha}_{2\ell\ell'}, \tilde{\theta}_{\ell\ell'})$
Average on $\ell' \neq \ell$, $s \in I_{\ell'}$, of $\mathbf{c}_K(Z_s)\mathbf{c}_K(Z_s)'$
$= \mathbf{b}_J'\hat{\boldsymbol{\rho}}_{J\ell\ell'}$
$\operatorname{argmin}_{\boldsymbol{\rho}_J \in \mathbb{R}^J}\{ \ -2\hat{D}_{2\ell\ell'}'\boldsymbol{\rho}_J \ + \ \boldsymbol{\rho}_J'\hat{B}_{\ell\ell}\boldsymbol{\rho}_J \ +\lambda\|\boldsymbol{\rho}_J\|_q^q\}$
Average on $\ell'' \notin \{\ell, \ell'\}$, $\iota \in I_{\ell''}$, of $D_2(W_\iota, \mathbf{b}_J | \hat{g}_{\ell\ell'\ell''}, \hat{h}_{\ell\ell'\ell''}, \tilde{\theta}_{\ell\ell'\ell''})$
Average on $\ell'' \notin \{\ell, \ell'\}$, $s \in I_{\ell''}$, of $\mathbf{b}_J(X_\iota, \varphi(D_\iota, Z_\iota, \hat{g}_{\ell\ell'\ell''}))\mathbf{b}_J(X_\iota, \varphi(D_\iota, Z_\iota, \hat{g}_{\ell\ell'\ell''}))'$

Figure 2 Illustration of the algorithm to estimate the moment condition $\hat{\psi}_{i\ell}$ for an observation $i \in I_\ell$.

discussed in CEINR. Furthermore, estimation rates for the nuisance parameters $(\alpha_{01}, \alpha_{02})$ require (i) that the dictionaries approximate well the nuisance parameters and (ii) being able to estimate the linear approximations of $\bar{m}(g, h, \theta)$ given by $\mathbb{E}[D_1(W, g)]$ and $\mathbb{E}[D_2(W, h)]$ at a certain rate (see Chernozhukov et al., 2022b).

In the presence of generated regressors, the theory needs to account for the fact that the estimator of the correction term (and probably that of the moment condition) evaluate the estimators $\hat{h}_\ell$ and $\hat{\alpha}_{2\ell}$ in the generated regressor $\hat{V}_{i\ell} \equiv \varphi(D_i, Z_i, \hat{g}_\ell)$ (c.f., equation (4.2)). We modify the expansion of $\hat{\psi}_{i\ell}(\theta_0) - \psi(W_i, g_0, h_0, \alpha_0, \theta_0)$ given by CEINR to deal with this fact. Also, we rely on smoothness conditions on the dictionaries and φ to ensure that evaluating at the generated regressor is not problematic.

We begin with assumptions on the dictionaries. The first formaly states that the dictionaries $(b_n)_{j=1}^{\infty}$ and $(c_k)_{k=1}^{\infty}$ span $\Delta_2(g_0) \cap L_2(g_0)$ and Δ_1, respectivelly.[3]

Assumption 5.1

a. For every j, $b_j \in \Delta_2(g_0) \cap L_2(g_0)$. Also, $\forall \delta_2 \in \Delta_2(g_0) \cap L_2(g_0)$ and for every $\varepsilon > 0$, there exist J and $\boldsymbol{\rho}_J$ such that $\|\delta_2 - \mathbf{b}'_J \boldsymbol{\rho}_J\|_2 < \varepsilon$.

b. For every k, $c_k \in \Delta_1$. Also, $\forall \delta_1 \in \Delta_1$ and for every $\varepsilon > 0$, there exist K and $\boldsymbol{\beta}_K$ such that $\|\delta_1 - \mathbf{c}'_K \boldsymbol{\beta}_K\|_2 < \varepsilon$.

We also assume bounded dictionaries:

Assumption 5.2 $\sup_{j \in \mathbb{N}} |b_j(X, V)| < \infty$ and $\sup_{k \in \mathbb{N}} |c_k(Z)| < \infty$ almost surely.

The assumption translates into consistency of $\hat{B}_\ell$ and $\hat{C}_\ell$. Also, on top of the following assumption, it will guarantee that the correction term nuisance parameters are bounded:

Assumption 5.3 For the real-valued sequences $(\rho_j)_{j=1}^{\infty}$ and $(\beta_k)_{k=1}^{\infty}$ such that $\alpha_{02}(x, v) = \sum_{j=1}^{\infty} \rho_j b_j(x, v)$ and $\alpha_{01}(z) = \sum_{k=1}^{\infty} \beta_k c_k(z)$:

a. $\sum_{j=1}^{\infty} |\rho_j| < \infty$ and $\sum_{k=1}^{\infty} |\beta_k| < \infty$.

b. For a $C > 0$, the atoms b_j and c_k corresponding to the largest $C\sqrt{n}$ values of ρ_j and β_k are included in $\mathbf{b}_J$ and $\mathbf{c}_K$.

This Assumption guarantees the L_1-norm of the coefficient of the Lasso penalized regression to be under control. The result is relevant to estimate the asymptotic variance (see Chernozhukov et al., 2022b) and to evaluate the estimators at the generated regressor.

We require the following estimation rates:

Assumption 5.4 There is $1/3 < r < 1/2$ such that

[3] In this section, for a measurable function $f(w)$, $\|f\|_2 \equiv \sqrt{\mathbb{E}[f(W)^2]}$ denotes its L_2-norm. Also, for a $d_1 \times d_2$ matrix $A = (A_{i,j})_{i=1,j=1}^{d_1,d_2}$, $\|A\|_\infty \equiv \max_{i,j} |A_{ij}|$.

a. $\|\hat{g}_\ell - g_0\|_2 = O_p(n^{-r})$ and $\|\hat{h}_\ell - h_0\|_2 = O_p(n^{-r})$.

b. $\|\hat{D}_{1\ell} - \mathbb{E}[D_1(W, \mathbf{c}_K)]\|_\infty = O_p(n^{-r})$ and $\|\hat{D}_{2\ell} - \mathbb{E}[D_2(W, \mathbf{b}_J)]\|_\infty = O_p(n^{-r})$.

The assumption gives rate conditions on the estimators of the nuisance parameters and on the linearization of the moment condition. Assumption 5.4.b may be derived from Assumption 5.4.a, some regularity conditons on the linearizations (see Chernozhukov et al., 2022b, Ass. 12), and some regularity conditions that allow evaluation at the generated regressor (see Assumption 5.9 below).

We also aks for the following rates for the Lasso penalty and the number of terms in the dictionaries:

Assumption 5.5

a. The Lasso penalty term $\lambda = \lambda(n)$ for estimation of $(\alpha_{01}, \alpha_{02})$ satisfies: $n^{-r} = o(\lambda)$ and $\lambda = o(n^{c-r})$ for every $c > 0$.

b. The number of terms in the dictionaries satisfy $J, K = O(n^\kappa)$ for a constant $\kappa > 0$.

This assumptions asks for the Lasso penalty to go to zero slightly slower than n^{-r}. For instance, a rate of $\log(n)/n^r$ is allowed. Moreover, it requires polynomial rates in the growth of the number of terms in the dictionaries.

The above are general conditions imposed on the dictionaries and the tunning parameters for the Lasso penalized regression. The specific problem at hand only apears in two instances. First, Assumption 5.1 requires that the dictionaries approximate well the correction term nuisance parameters (living in Δ_1 and $\Delta_2(g_o) \cap L_2(g_0)$). Second, Assumption 5.4 requires (i) mean-square rates for the estimators of g_0 and h_0 and (ii) to be able to estimate the linearizations at the same rate. As discussed before, these conditions provide rates of estimators of the nuissance parameters in correction terms α_{01} and α_{02} (see Chernozhukov et al., 2022b). For instance, the convergence rate of $\hat{\alpha}_{1\ell}$ will be fast enough to guarantee that the interaction term satisfies $\|\hat{\alpha}_{1\ell} - \alpha_{01}\|_2 \cdot \|\hat{g}_\ell - g_0\|_2 = o_p(n^{-1/2})$ (c.f. Assumption 2 in CEINR).

We now provide assumptions on the moment condition. The first is a mean-square consistency condition similar to Assumption 1 in CEINR:

Assumption 5.6

a. $\mathbb{E}[m(W, g_0, h_0, \theta_0)^2] < \infty$.

b. $\int [m(w, \hat{g}_\ell, \hat{h}_\ell, \theta_0) - m(w, g_0, h_0, \theta_0)]^2 dF_0(w) \xrightarrow{P} 0$.

c. $\int [m(w, \hat{g}_\ell, \hat{h}_\ell, \tilde{\theta}_\ell) - m(w, \hat{g}_\ell, \hat{h}_\ell, \theta_0)]^2 dF_0(w) \xrightarrow{P} 0$.

d. $\mathbb{E}[(Y - h_0(X, V))^2 | X, V]$ and $\mathbb{E}[(D - g_0(Z))^2 | Z]$ are bounded almost surely.

The first point is necessary for regular estimation of θ_0. Assumptions 5.6.b and 5.6.c are mean-square consistency conditios for the moment condition. Boundedness of the conditional variances (Assumption 5.6.d) easily translates into mean-square consistency conditions for the correction term ϕ. We repeat here that boundedness of the correction term nuisance parameters α_{01} and α_{02} is implied by Assumptions 5.2 and 5.3.a.

We require the linear approximation of $\bar{m}(g, h, \theta_0)$ to be good enoug (in a neighborhood of (g_0, h_0)):

Assumption 5.7 There is a $\varepsilon > 0$ and a $C > 0$ such that, if $\|g - g_0\|_2 < \varepsilon$ and $\|h - h_0\|_2 < \varepsilon$, then

$$\begin{aligned}|\mathbb{E}\left[m(W, g, h, \theta_0) - m(W, g_0, h_0, \theta_0) - D_1(W, g - g_0) - D_2(W, h - h_0)\right]|\\ \leq C\left(\|g - g_0\|_2^2 + \|h - h_0\|_2^2\right).\end{aligned}$$

This Assumption translates in the Locally Robust property in Assumption 3.iii in CEINR. It asks that, once we have removed the first-order effect of estimating (g_0, h_0), the remainder term must be at most quadratic. In many case $\bar{m}(h, g, \theta_0)$ is affine in h, so the above assumption translates into a quadratic bias condition on the effect of first-step estmation: $|\mathbb{E}[m(W, g, h_0, \theta_0) - m(W, g_0, h_0, \theta_0) - D_1(W, g)]| \leq C\|g - g_0\|_2^2$.

The GMM procedure requires consistent estimation of the Jacobian of the moment condition. The following assumption gives sufficient conditions:

Assumption 5.8 There exists a neighborhood $\mathcal{N}$ of θ_0 such that, for small $\|g - g_0\|_2$ and $\|h - h_0\|_2$:

a. $m(W, g, h, \theta)$ is almost surely differentiable in $\mathcal{N}$.

b. There exists a $C > 0$ and a function $d(W, g, h)$, with $\mathbb{E}[d(W, g, h)] < C$, such that for $\theta \in \mathcal{N}$

$$\left\|\frac{\partial m}{\partial \theta}(W, g, h, \theta) - \frac{\partial m}{\partial \theta}(W, g, h, \theta_0)\right\|_\infty \leq d(W, g, h)\|\theta - \theta_0\|_\infty^{1/C} \text{ almost surely.}$$

Moreover, we assume that:

c. M, the expectation of the Jacobian, exists.

d. It holds that

$$\int \left\|\frac{\partial m}{\partial \theta}(w, \hat{g}_\ell, \hat{h}_\ell, \theta_0) - \frac{\partial m}{\partial \theta}(w, g_0, h_0, \theta_0)\right\|_\infty dF_0(w) \xrightarrow{P} 0.$$

We conclude the set of sssumptions with smoothness conditions on the dictionary $(b_j)_{j=1}^\infty$ and the function giving the generated regressor φ. The following assumption allows us to deal with $\hat{h}_\ell$ and $\hat{\alpha}_{2\ell}$ being evaluated at the generated regressor.

Assumption 5.9

- **a.** h_0 also safisfies Assumption 5.3.
- **b.** $\sup_{j\in\mathbb{N}} |\partial b_j(X,V)/\partial v| < \infty$ almost surely.
- **c.** There exists a $\varepsilon > 0$ such that, if $\|g - g_0\|_2 < \varepsilon$, $b_j(\cdot, \varphi(\cdot,\cdot,g)) \in \Delta(g_0)$ for every $j \in \mathbb{N}$.

In Theorem 5.1, we give asymptotic normality of the automatic debiased GMM estimator when $\hat{h}_\ell$ is a Lasso penalized regression of Y onto $\mathbf{b}_J(X,V)$ (see Section 4). For other estimators of the second step nuisance paremeters, Assumption 5.9.a may be replaced by $\|\tilde{h}_\ell - h_0\|_2 = O_p(\|\hat{h}_\ell - h_0\|_2)$, where $\tilde{h}_\ell(w) \equiv \hat{h}_\ell(x, \varphi(d,z,\hat{g}_\ell))$.

Assumption 5.9.b alows us to bound the efect of departures from evaluation at the "true" generated regressor $V \equiv \varphi(D,Z,g_0)$. Assumption 5.9.c simply requires that, for small $\|g-g_0\|_2$, the dictionary evaluated in the generated regressor is in the space where h_0 and α_{02} live. For instance, in the case $\Delta_2(g) = \{(d,z) \mapsto \delta(x, \varphi(d,z,g)) \colon \delta \in \Delta\}$ for a subspace $\Delta \subseteq L_2(X,V)$, the assumption generaly requires that $\mathbb{E}[b_j(X, \varphi(D,Z,g))^2] < \infty$. This will follow from Hadamard differentiability of φ (Assumption (A4)) and 5.9.b.

Assumptions 5.3, 5.4, 5.6, and 5.7 are stated for a single moment condition. In the presence of more than one condition, they must be understood to hold componentwise. The same happens with Assumptions (A1), (A2), and (A4) in Section 3.1. Assumption 5.8, since it refers to a GMM-specific situation, is already formulated in the general case. The remaining assumptions do not depend on the dimension of the moment condition (they depend, on the other hand, on the dimension of Y and D).

Recall that $V \equiv (M'\Upsilon M)^{-1} M'\Upsilon'\Psi\Upsilon M (M'\Upsilon M)^{-1}$ gives the asymptotic variance of the automatic debiased GMM estimator. Define $\hat{V} \equiv (\hat{M}'\hat{\Upsilon}\hat{M})^{-1}\hat{M}'\hat{\Upsilon}'\hat{\Psi}\hat{\Upsilon}\hat{M}(\hat{M}'\hat{\Upsilon}\hat{M})^{-1}$ as the plug-in estimator of the asymptotic variance, where $\hat{M}$ and $\hat{\Psi}$ are given in equations (4.4) and (4.5), respectivelly. The following theorem ensures asymptotic normality of $\sqrt{n}(\hat{\theta} - \theta_0)$:

Theorem 5.1 *Consider that Assumptions (A1)-(A4) and 5.1-5.9 are satisfied,* $\hat{\Upsilon} \xrightarrow{P} \Upsilon$, *$M'\Upsilon M$ is non-singular, and $\hat{h}_\ell$ are Lasso estimators of h_0. Then, the automatic debiased GMM estimator in equation (4.3) satisfies*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0, V).$$

Moreover, the plug-in estimator for the asymptotic variance is consistent: $\hat{V} \xrightarrow{P} V$.

6 Monte Carlo simulation

This section describes the Monte Carlo simulation to evaluate the finite sample properties of the CASF-estimator proposed in this paper. Before presenting the results, we briefly describe the Data Generating Process and the implemented estimators.

6.1 Description

The Data Generating Process is

$$(Z, U, V) \sim N\left(0, \begin{bmatrix} \mathrm{Id}_6 & 0 & 0 \\ 0 & 1 & 1/2 \\ 0 & 1/2 & 1 \end{bmatrix}\right),$$

where Id_6 denotes the 6×6 Identity Matrix. Therefore, Z is a 6-dimensional random vector. The correlation between U and V is $1/2$. Note that the fact that $Z \perp U$ and $Z \perp V$ guarantees that the Control Function Assumption is satisfied.

Both D and Y are generated by the following linear models:

$$Y = \sum_{k=1}^{5} Z_k + 2D + U \text{ and}$$

$$D = \sum_{k=1}^{6} Z_k + V.$$

So Z_6 is excluded from the structural equation (i.e., it does not directly affect Y) and may be used as an instrument.

We estimate the CASF for the following counterfactual distribution F_X^*: (i) the distribution of $(Z_1, \ldots, Z_5)$ remains unchanged and (ii) D is normal with mean 1 (instead of 0) and the same variance as in the DGP. Therefore, the true parameter is $\theta_0 = 2$.

We note here that, even if the model considered is linear, the second-step correction nuisance parameter is highly non-linear. Letting $s \equiv \sum_{k=1}^{5} z_k$, the nuisance parameter is

$$\alpha_{02}(z_1, \ldots, z_5, d, v) = C \cdot \exp\left(-\frac{1}{4} - \frac{s}{2} + \frac{d}{2} + \frac{s^2}{4} + \frac{v^2}{2} + \frac{d^2}{4} - \frac{sd}{2} + sv - dv\right),$$

for a constant C. This nuisance parameter is automatically estimated with dictionaries that do not account for the complexity of the function (see below). Moreover, we would like to highlight that the first-step correction nuisance parameter includes terms as $\mathbb{E}[\alpha_{02}(Z_1, \ldots, Z_5, D, V)|Z]$.

We display results for three different estimators of the CASF:

- The naive plug-in estimator: $\hat{\theta}_{PI} \equiv n^{-1} \sum_{i=1}^{n} m(W_i, \hat{g}, \hat{h})$.

- A cross-fitted debiased estimator that only corrects for the effect of pluging-in $\hat{h}$: $\hat{\theta}_{DB2}$ is as in equation (4.6) but with $\phi(W_i, \hat{g}_\ell, \hat{h}_\ell, \hat{\alpha}_\ell, \tilde{\theta}_\ell)$ replaced by $\hat{\alpha}_{2\ell}(X_i, \hat{V}_{i\ell}) \cdot (Y_i - \hat{h}_\ell(X_i, \hat{V}_{i\ell}))$. That is, the correction term for the first step is omitted.

- The cross-fitted fully debiased estimator: $\hat{\theta}_{DBF}$ as in equation (4.6). That is, it used the debiased moment condition in equation (4.2).

Numerical integration, with a sample of size $S = 10^7$, is used to solve the integrals w.r.t. F_X^*. The estimators for the nuisance parameters g_0 and h_0 are Lasso with three dictionaries: one that includes linear terms, another including linear and quadratic terms, and a last one including linear, quadratic, and interaction terms. The number of splits for cross-fitting is $L = 5$ for every sample size.

To perform inference with each estimator, we present results that parallel common practice. The fully debiased estimator uses the correct asymptotic variance, that does account for first and second step estimation. This is given by equation (4.5). The estimator $\hat{\theta}_{DB2}$ only accounts for the second step when computing the asymptotic variance (as it does for estimation). Its asymptotic variance can be constructed by replacing $\phi(W_i, \hat{g}_\ell, \hat{h}_\ell, \hat{\alpha}_\ell, \tilde{\theta}_\ell)$ by $\hat{\alpha}_{2\ell}(X_i, \hat{V}_{i\ell}) \cdot (Y_i - \hat{h}_\ell(X_i, \hat{V}_{i\ell}))$, the second step IF. To emphasize that plug-in estimation leads to an asymptotic bias problem, confidence intervals for the plug-in estimator are built with correct asymptotic variance (the one in equation (4.5)).

6.2 Results

The next tables report results for a Monte Carlo simulation with $B = 1098$ replications. Each table gives results for a different dictionary: linear, quadratic or the one which also includes interaction terms.

	Mean Absolute Bias			Standard Error			Coverage		
n	PI	DB2	DBF	PI	DB2	DBF	PI	DB2	DBF
100	0.2285	0.1463	0.1482	0.1649	0.1812	0.1733	0.6388	0.8681	0.8626
500	0.1425	0.0594	0.0516	0.0685	0.0692	0.0645	0.3876	0.8954	0.9208
1000	0.1236	0.0435	0.0376	0.049	0.0488	0.0462	0.2266	0.8744	0.9272
5000	0.0852	0.0234	0.0169	0.0227	0.0212	0.0202	0.0227	0.7925	0.9290
10000	0.0746	0.0181	0.0123	0.017	0.0148	0.0142	0.0018	0.7489	0.9163

Table 1 CASF results for the dictionary including linear terms.

Tables 1 and 2 present results for the linear and quadratic dictionaries, respectively. Correcting for the second-step already reduces a large amount of the bias of the plug-in estimator. Adding the first-step correction further decreases bias. It is worth highlighting that the the better performance relative the the bias comes with a smaller standard deviation, particularly for larger samples. As shown in the tables, however, the estimator accounting only for the second-step fails to keep coverage at the nominal 95% level as sample size increases.

The tables highlight that the plug-in estimator suffers from severe asymptotic bias issues: coverage decreases rapidly, even if the confidence interval is constructed with correct standard

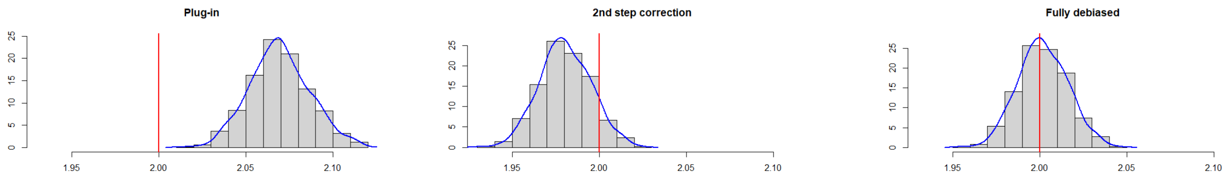

FIGURE 3 Distribution of the CASF estimators using a quadratic dictionary for a sample size of $n = 10000$.

errors. Indeed, Figure 3 shows that, for a sample size of $n = 10000$, the distribution of plug-in estimators has almost zero mass near the true parameter $\theta_0 = 2$.

	Mean Absolute Bias			Standard Error			Coverage		
n	PI	DB2	DBF	PI	DB2	DBF	PI	DB2	DBF
100	0.2764	0.1898	0.1957	0.1766	0.2003	0.2	0.5532	0.7534	0.7561
500	0.1462	0.0603	0.0527	0.0692	0.0709	0.0663	0.3794	0.8963	0.9327
1000	0.1214	0.0446	0.0374	0.0491	0.0488	0.0465	0.2329	0.8717	0.929
5000	0.079	0.0259	0.0161	0.0236	0.0212	0.0202	0.0437	0.737	0.9354
10000	0.0694	0.0209	0.0115	0.0172	0.0148	0.0143	0.0036	0.6533	0.9327

TABLE 2 CASF results for the dictionary including linear and quadratic terms.

Table 3 displays results for the the dictionary that also includes interaction terms. The results are striking as the estimator only account for the second step performs well. It is able to keep coverage at nominal levels, outperforming the fully debiased estimator. We believe that this fact rests on the linear dictionary performing well to estimate α_{02} but notably worst to estimate the more complex $\partial\alpha_{02}/\partial v$, which is needed for the fist-step correction. Nevertheless, the decrease in coverage is small: 1-2% for intermediate sample sizes ($n = 500$ and 1000) and 4-5% for large samples ($n = 5000$ and 10000). This result suggest that the complexity of the dictionary must be increased faster when accounting for the first step.

7 Conclusion

We propose an Automatic Locally Robust estimators for structural parameters in the presence of generated regressors. We show that the debiasing correction term can be decomposed into terms accounting for first step and second step estimation. Each of the first and second step IF depends on an additional nuisance parameter, which can be automatically estimated (i.e., estimated without finding their analytic shape).

	Mean Absolute Bias			Standard Error			Coverage		
n	PI	DB2	DBF	PI	DB2	DBF	PI	DB2	DBF
100	0.3447	0.3583	0.3857	0.179	0.3606	0.4054	0.9016	0.7969	0.806
500	0.1707	0.0997	0.1085	0.0693	0.1213	0.1359	0.7925	0.9481	0.9227
1000	0.1367	0.0635	0.0674	0.0488	0.0776	0.0849	0.5883	0.9372	0.9262
5000	0.0846	0.0259	0.0283	0.0229	0.0312	0.0349	0.1383	0.9372	0.8926
10000	0.0729	0.0173	0.0205	0.017	0.0207	0.0228	0.0337	0.9399	0.8926

Table 3 CASF results for the dictionary including linear, quadratic, and interaction terms.

We apply our results to construct Automatic Locally Robust estimators for the CASF under the control function assumption and Average Partial Effects in sample selection models. The analytic shape of the nuisance parameters in these two cases is particularly complex, as the moment conditions depend on the whole shape of the second step parameter (not only its pointwise value). Therefore, automatic estimation is particularly suited for these problems.

Appendix A Additional examples

Example 2 (continuing from p. 6) 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 tuning parameter s_n. Alternatively, we can take advantage of a differentiable dictionary $(b_j(x, v))_{j=1}^{\infty}$, as described below.

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

$$\left(\frac{\partial^2 \mathbf{b}_J}{\partial x \partial v}(X_i, \hat{g}_{\ell\ell'}(Z_i)) \right)' \boldsymbol{\eta}_J,$$

in case that $h_0(x, v)$ is approximated by $\mathbf{b}_J(x, v)'\boldsymbol{\eta}_J$. To estimate the second term we can use equation (4.11), 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}.

$\blacksquare$

Appendix B 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:

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

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 (B.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 $\Delta_2(g_0)$:

$$\alpha_{20} = \operatorname*{argmin}_{\alpha \in \Delta_2(g_0)} \mathbb{E}[(r_2(X, \varphi(D,Z,g_0)) - \alpha(D,Z))^2]. \tag{B.2}$$

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 $\alpha \in \Delta_2(g_0)$ 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 $\alpha \in \Delta_2(g_0)$,

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

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 $\alpha \in \Delta_2(g_0) \cap L_2(g_0)$. This gives Point (IF). ■

Proof of Theorem 3.1: We compute $d\bar{m}(g(F_\tau), h_0, \theta)/d\tau$ and $d\bar{m}(g_0, h(F_0, g(F_\tau)), \theta)/d\tau$ separately and then add them according to equation (3.2). By Assumptions (A1) and (A2), using the Riesz Representation Theorem, we have that for the differentiable paths $\tau \mapsto g(F_\tau)$ and $\tau \mapsto h(F_0, g(F_\tau))$:

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

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

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

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 $\delta \in \Delta_2(g(F_\tau)) \cap L_2(g_0)$, 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:

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

A final step is needed to connect equation (B.4) with the above result. To perform it, we use Assumption (A3) in two directions. First, since $\alpha_{02} \in \Delta_2(g_0) \cap L_2(g_0)$, we have that $\alpha_{02}(\cdot, \varphi(\cdot, \cdot, g(F_\tau))) \in \Delta_2(g(F_\tau)) \cap L_2(g_0)$. We can then apply equation (B.5) to α_{02}. Moreover, since $h(F_0, g(F_\tau))(\cdot, \varphi(\cdot, \cdot, g(F_\tau))) \in \Delta_2(g(F_\tau))$, we also have that $h(F_0, g(F_\tau))(\cdot, \varphi(\cdot, \cdot, g_0)) \in \Delta_2(g_0)$. This means that, in equation (B.4), we can dismiss the component of r_2 that is orthogonal to $\Delta_2(g_0)$. 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 (B.4). Combining this with equation (B.5):

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

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 (B.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 (B.6):

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

Since D_φ is linear in g, the function inside the expectation in the RHS is linear in g. We now use equation (3.2) to combine the results in equations (B.3) and (B.7). This gives:

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

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 (B.7), noting that $V \equiv \varphi(D,Z,g_0)$ is a function of (D,Z):

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

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 (3.2) to combine this last result with that in equation (B.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:

$$\alpha_{10} = \operatorname*{argmin}_{\alpha\in\Delta_1}\mathbb{E}[(\tilde{\nu}(Z)-\alpha(Z))^2], \tag{B.8}$$

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

The asymptotic normality and consistent estimation of the asymptotic variance result in Theorem 5.1 relies on the following lemma:

Lemma B.1 *Consider Assumptions (A4), 5.3, 5.4.a, and 5.9. Let* $r,\xi>0$. *For* $\tilde{h}_\ell(w)\equiv\hat{h}_\ell(x,\varphi(d,z,\hat{g}_\ell))$ *and* $\tilde{\alpha}_{2\ell}(w)\equiv\hat{\alpha}_{2\ell}(x,\varphi(d,z,\hat{g}_\ell))$*:*

$$\|\hat{h}_\ell-h_0\|_2=O_p(n^{-r})\Rightarrow\|\tilde{h}_\ell-h_0\|_2=O_p(n^{-r}),\ \textit{and}$$
$$\|\hat{\alpha}_{2\ell}-\alpha_{02}\|_2=o_p(n^{\xi}n^{-r/2})\Rightarrow\|\tilde{\alpha}_{2\ell}-\alpha_{02}\|_2=o_p(n^{\xi}n^{-r/2}).$$

Proof of Lemma B.1: Recall that $\hat{h}_\ell(x,v)\equiv\mathbf{b}_J(x,v)'\hat{\boldsymbol{\eta}}_J$ and $\hat{\alpha}_{2\ell}(x,v)\equiv\mathbf{b}_J(x,v)'\hat{\boldsymbol{\rho}}_J$. We show that

$$\|\hat{h}_\ell-\hat{h}_\ell\|_2=O_p(n^{-r}).$$

The conclusion for $\tilde{\alpha}_{2\ell}$ follows the same reasoning.

Let $\tilde{b}_j(w) \equiv b_j(w, \varphi(d, z, \hat{g}_\ell))$. By the triangle inequality $\|\hat{h}_\ell - \hat{h}_\ell\|_2 \leq \sum_{j=1}^J |\hat{\eta}_j| \|\tilde{b}_j - b_j\|_2$. Moreover, by the Mean Value Theorem and Assumption 5.9.b,

$$\|\tilde{b}_j - b_j\|_2^2 = \int (b_j(x, \varphi(d, z, \hat{g}_\ell)) - b_j(x, v))^2 dF_0(w) \leq \kappa^2 \|\varphi(\cdot, \cdot, \hat{g}_\ell) - \varphi(\cdot, \cdot, g_0)\|_2^2.$$

Then, $\sup_{j \leq J} \|\tilde{b}_j - b_j\|_2 \leq \kappa \|\varphi(\cdot, \cdot, \hat{g}_\ell) - \varphi(\cdot, \cdot, g_0)\|_2$. Also, since φ is Hadamard differentiable (Assumption (A4)): $\|\varphi(\cdot, \cdot, \hat{g}_\ell) - \varphi(\cdot, \cdot, g_0) - D_\varphi(\hat{g}_\ell - g_0)\|_2 = o_p(\|\hat{g}_\ell - g_0\|)$ and $\|D_\varphi(\hat{g}_\ell - g_0)\|_2 \leq C\|\hat{g}_\ell - g_0\|_2$ (see Yamamuro, 1974, Result 1.2.6). Thus, by the triangle inequality:

$$\sup_{j \leq J} \|\tilde{b}_j - b_j\|_2 \leq \kappa C \|\hat{g}_\ell - g_0\|_2 + o_p(\|\hat{g}_\ell - g_0\|_2).$$

Now, Assumption 5.3 allows us to apply Lemma A9 in Chernozhukov et al. (2022b) to get $\sum_{j=1}^J |\hat{\eta}_j| = O_p(1)$. Therefore, if Assumption 5.4.a holds,

$$\begin{aligned} \|\hat{h}_\ell - \hat{h}_\ell\|_2 &\leq \sum_{j=1}^J |\hat{\eta}_j| \|\tilde{b}_j - b_j\|_2 \leq \left(\sum_{j=1}^J |\hat{\eta}_j|\right) (\kappa C \|\hat{g}_\ell - g_0\|_2 + o_p(\|\hat{g}_\ell - g_0\|_2)) \\ &= O_p(1) \cdot [O_p(n^{-r}) + o_p(n^{-r})] = O_p(n^{-r}). \end{aligned} \tag{B.9}$$

The conclusion for $\|\tilde{h}_\ell - h_0\|_2$ follows directly from equation (B.9) and the triangle inequality. The conclusion for $\|\tilde{\alpha}_{2\ell} - \alpha_{02}\|_2$ follows equally, taking into account that $O_p(n^{-r}) = o_p(n^\xi n^{-r/2})$ for $\xi > 0$. ■

Proof of Theorem 5.1: We start with asymptotic normality of $\sqrt{n}(\hat{\theta} - \theta_0)$. The proof follows standard GMM techniques and Theorem 9 in Chernozhukov et al. (2022b). A relevant deviation from the previous results is that the estimators are evaluated at the generated regressor. Lemma B.1 allows to deal with that situation.

The cornerstone of the result is Lemma 8 in Chernozhukov et al. (2022a), CEINR in what follows, which states that:

$$\sqrt{n}\hat{\psi}(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(W_i, g_0, h_0, \alpha_0, \theta_0) + o_p(1) \tag{B.10}$$

under some conditions. We will apply Lemma 8 to a modified expansion of the difference between $\hat{\psi}(\theta_0)$ and $n^{-1/2} \sum_{i=1}^n \psi(W_i, g_0, h_0, \alpha_0, \theta_0)$ that allows to deal with estimators evaluated at the generated regressor.

Consider first that equation (B.10) holds for each component of $\hat{\psi}$. Consistency of $\hat{\theta}$ follows under standard conditions that guarantee uniform convergence of $\hat{\psi}(\theta)'\hat{\Upsilon}\hat{\psi}(\theta)$ in Θ (c.f. Wooldridge, 2010, Th. 14.1). These conditions will follow from Assumption 5.8 if Θ is compact. Moreover, by Assumptions 5.4.a and 5.8.a we can apply the Mean Value Theorem to get

$$\sqrt{n}\left(\hat{\psi}(\hat{\theta}) - \hat{\psi}\,(\theta_0)\right) = \sqrt{n}\frac{\partial \hat{\psi}}{\partial \theta}(\bar{\theta}) \cdot (\hat{\theta} - \theta_0)$$

for $\bar{\theta}$ a point between θ_0 and $\hat{\theta}$ (that is $\bar{\theta} \xrightarrow{P} \theta_0$). Then, if equation (B.10) holds:

$$\sqrt{n}\frac{\partial \hat{\psi}}{\partial \theta}(\bar{\theta}) \cdot (\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(W_i, g_0, h_0, \alpha_0, \theta_0) + \sqrt{n}\hat{\psi}(\hat{\theta}) + o_p(1). \tag{B.11}$$

Now, note that

$$\begin{aligned}\frac{\partial \hat{\psi}}{\partial \theta}(\theta) &= \frac{\partial}{\partial \theta}\left(\frac{1}{n}\sum_{\ell=1}^{L}\sum_{i \in I_\ell}\left[m(W_i, \hat{g}_\ell, \hat{h}_\ell, \theta) + \phi(W_i, \hat{g}_\ell, \hat{h}_\ell, \hat{\alpha}_\ell, \tilde{\theta}_\ell)\right]\right)\\ &= \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i \in I_\ell}\frac{\partial m}{\partial \theta}(W_i, \hat{g}_\ell, \hat{h}_\ell, \theta)\end{aligned}$$

Then, since Assumptions 5.4.a and 5.8, on top of $\hat{\theta} \xrightarrow{P} \theta_0$, guarantee that we can apply Lemma E2 in CEINR, we have that $\partial\hat{\psi}(\hat{\theta})/\partial\theta \xrightarrow{P} M$, so it is bounded in probability. Therefore, since $\hat{\Upsilon}$ is also $O_p(1)$, equation (B.11) implies

$$\begin{aligned}\frac{\partial \hat{\psi}}{\partial \theta}(\hat{\theta})'\hat{\Upsilon}\frac{\partial \hat{\psi}}{\partial \theta}(\bar{\theta}) \cdot \sqrt{n}(\hat{\theta} - \theta_0) = {} & \frac{\partial \hat{\psi}}{\partial \theta}(\hat{\theta})'\hat{\Upsilon} \cdot \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(W_i, g_0, h_0, \alpha_0, \theta_0)\\ & + \sqrt{n}\frac{\partial \hat{\psi}}{\partial \theta}(\hat{\theta})'\hat{\Upsilon}\hat{\psi}(\hat{\theta}) + o_p(1).\end{aligned}$$

Thus, since $(\partial\hat{\psi}(\hat{\theta})/\partial\theta)'\hat{\Upsilon}\hat{\psi}(\hat{\theta}) = 0$ is the first-order condition for the minimization problem in equation (4.3), $\partial\hat{\psi}(\bar{\theta})/\partial\theta \xrightarrow{P} M$ (by Lemma E2 in CEINR), and $M'\Upsilon M$ is non-sigular:

$$\sqrt{n}(\hat{\theta} - \theta_0) = (M'\Upsilon M)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(W_i, g_0, h_0, \alpha_0, \theta_0) + o_p(1).$$

Then, the asymptotic normality result follows from $n^{-1/2}\sum_{i=1}^{n}\psi(W_i, g_0, h_0, \alpha_0, \theta_0) \xrightarrow{D} N(0, \Psi)$.

It remains to verify the assumptions for Lemma 8 in CEINR, so that equation (B.10) holds for each component of $\hat{\psi}$. First, to handle estimators evaluated at the generated regressor, we provide a modified expansion of the difference between $\hat{\psi}(\theta_0)$ and $n^{-1/2}\sum_{i=1}^{n}\psi(W_i, g_0, h_0, \alpha_0, \theta_0)$. Let $\bar{\phi}(w, \bar{v}, g, h, \alpha) \equiv \alpha_1(z) \cdot (d - g(z)) + \alpha_2(x, \bar{v}) \cdot (y - h(x, \varphi(d, z, g)))$, which makes explicity that α_2 is evaluated at $(x, \bar{v})$. Then, being $\hat{\psi}_{i\ell}(\theta)$ given by equation (4.2), we have that $\hat{\psi}_{i\ell}(\theta_0) - \psi(W_i, g_0, h_0, \alpha_0, \theta_0) = \hat{R}_{1i\ell} + \hat{R}_{2i\ell} + \hat{R}_{3i\ell} + \hat{\Delta}_{i\ell}$, where

$$\begin{aligned}\hat{R}_{1i\ell} &\equiv m(W_i, \hat{g}_\ell, \hat{h}_\ell, \theta_0) - m(W_i, g_0, h_0, \theta_0),\\ \hat{R}_{2i\ell} &\equiv \bar{\phi}(W_i, \varphi(D_i, Z_i, g_0), \hat{g}_\ell, \hat{h}_\ell, \alpha_0) - \phi(W_i, g_0, h_0, \alpha_0, \theta_0),\\ \hat{R}_{3i\ell} &\equiv \bar{\phi}(W_i, \varphi(D_i, Z_i, \hat{g}_\ell), g_0, h_0, \hat{\alpha}_\ell) - \phi(W_i, g_0, h_0, \alpha_0, \theta_0),\\ \hat{\Delta}_{i\ell} &\equiv \phi(W_i, \hat{g}_\ell, \hat{h}_\ell, \hat{\alpha}_\ell, \tilde{\theta}_\ell) - \bar{\phi}(W_i, \varphi(D_i, Z_i, g_0), \hat{g}_\ell, \hat{h}_\ell, \alpha_0), \text{ and}\\ &\quad - \bar{\phi}(W_i, \varphi(D_i, Z_i, \hat{g}_\ell), g_0, h_0, \hat{\alpha}_\ell) + \phi(W_i, g_0, h_0, \alpha_0, \theta_0).\end{aligned} \tag{B.12}$$

We will apply Lemma 8 in CEINR to this expansion.

Following Chernozhukov et al. (2022b), we begin by providing rates for estimation of α_{01} and α_{02}. Assumption 5.2 allows us to apply Lemma A10 in Chernozhukov et al. (2022b) to get $\|\hat{B}_\ell - \mathbb{E}[\mathbf{b}_J(X,V)\mathbf{b}_J(X,V)']\|_\infty = O_p(\sqrt{\log(J)/n})$ and $\|\hat{C}_\ell - \mathbb{E}[\mathbf{c}_K(Z)\mathbf{c}_K(Z)']\|_\infty = O_p(\sqrt{\log(K)/n})$. The fact that Assumption 5.5.b imposes a polynomial rate on J and K then implies that $O_p(\sqrt{\log(J)/n}) = O_p(n^{-r})$ and $O_p(\sqrt{\log(K)/n}) = O_p(n^{-r})$, since $r < 1/2$ (Assumption 5.4). This, on top of Assumptions 5.1, 5.3, 5.4.b, and 5.5, means that we can apply Theorem 2 in Chernozhukov et al. (2022b) to get:

$$\|\hat{\alpha}_{1\ell} - \alpha_{01}\|_2 = o_p(n^\xi n^{-r/2}) \text{ and } \|\hat{\alpha}_{2\ell} - \alpha_{02}\|_2 = o_p(n^\xi n^{-r/2})$$

for any $\xi > 0$. We choose a ξ satisfying $0 < \xi < (3r-1)/2 < r/2$, which is possible since, by Assumption 5.4, $r \in (1/3, 1/2)$. This guarantees that

$$\|\hat{\alpha}_{1\ell} - \alpha_{01}\|_2 = o_p(1) \text{ and } \|\hat{\alpha}_{2\ell} - \alpha_{02}\|_2 = o_p(1); \text{ and} \tag{B.13}$$

$$\sqrt{n}\|\hat{\alpha}_{1\ell} - \alpha_{01}\|_2\|\hat{g} - g_0\|_2 = o_p(1) \text{ and } \sqrt{n}\|\hat{\alpha}_{2\ell} - \alpha_{02}\|_2\|\hat{h}_\ell - h_0\| = o_p(1), \tag{B.14}$$

where the last line follows from Assumption 5.4.a.

We now check Assumption 1 in CEINR. Assumption 1.i is identical to Assumption 5.6.a and 5.6.b. To show the remaining points, define $\tilde{h}_\ell(w) \equiv \hat{h}_\ell(x, \varphi(d, z, \hat{g}_\ell))$ and $\tilde{\alpha}_{2\ell}(w) \equiv \hat{\alpha}_{2\ell}(x, \varphi(d, z, \hat{g}_\ell))$. Note also that by Assumptions 5.2 and 5.3.a:

$$|\alpha_{01}(Z)| \leq \sum_{k=1}^{\infty} |\beta_k||c_k(Z)| \leq \sup_{k\in\mathbb{N}} |c_k(Z)| \cdot \sum_{k=1}^{\infty} |\beta_k| \equiv \kappa_1 < \infty \text{ and}$$
$$|\alpha_{02}(X,V)| \leq \sum_{j=1}^{\infty} |\rho_j||b_j(X,V)| \leq \sup_{j\in\mathbb{N}} |b_j(X,V)| \cdot \sum_{j=1}^{\infty} |\rho_j| \equiv \kappa_2 < \infty.$$

Thus, by the triangle inequality, being $v \equiv \varphi(d, z, g_0)$:

$$\begin{aligned}\int \left[\bar{\phi}(w, \varphi(d,z,g_0), \hat{g}_\ell, \hat{h}_\ell, \alpha_0) - \phi(w, g_0, h_0, \alpha_0, \theta_0)\right]^2 dF_0(w) &\leq \int \alpha_{01}(z)^2 \left[\hat{g}_\ell(z) - g_0(z)\right]^2 dF_0(w) \\ &+ \int \alpha_{02}(x,v)^2 \left[\tilde{h}_\ell(w) - h_0(x,v)\right]^2 dF_0(w) \\ &\leq \kappa_1^2 \|\hat{g}_\ell - g_0\|_2^2 + \kappa_2^2 \|\tilde{h}_\ell - h_0\|_2^2.\end{aligned}$$

Assumption 1.ii in CEINR follows from the above display, Assumption 5.4.a, and Lemma B.1.

Also, calling $\kappa_3, \kappa_4 < \infty$ to the bounds given by Assumption 5.6.d:

$$\begin{aligned}\int \left[\bar{\phi}(w, \varphi(d,z,\hat{g}_\ell), g_0, h_0, \hat{\alpha}_\ell) - \phi(w, g_0, h_0, \alpha_0, \theta_0)\right]^2 dF_0(w) &\leq \int \left[d - g_0(z)\right]^2 \left[\hat{\alpha}_{1\ell}(z) - \alpha_{01}(z)\right]^2 dF_0(w) \\ &+ \int \left[y - h_0(x,v)\right]^2 \left[\tilde{\alpha}_{2\ell}(w) - \alpha_{02}(x,v)\right]^2 dF_0(w) \\ &\leq \kappa_3^2 \|\hat{\alpha}_{1\ell} - \alpha_{01}\|_2^2 + \kappa_4^2 \|\tilde{\alpha}_{2\ell} - \alpha_{02}\|_2^2,\end{aligned}$$

Thus, Assumption 1.iii in CEINR follows from the above display, equation (B.13), and Lemma B.1.

We now move to check Assumption 2 in CEINR. In particular, we show that 2.iii holds. We have that

$$\hat{\Delta}_{i\ell} = [\hat{\alpha}_{1\ell}(Z_i) - \alpha_{01}(Z_i)] \cdot [\hat{g}_\ell(Z_i) - g_0(Z_i)] + [\tilde{\alpha}_{2\ell}(W_i) - \alpha_{02}(X_i, V_i)] \cdot [\tilde{h}_\ell(W_i) - h_0(X_i, V_i)].$$

Thus, as in Chernozhukov et al. (2022b, proof of Th. 9), an application of the Cauchy-Schwarz, conditional Markov, and triangle inequalities leads to:

$$\left| \frac{1}{\sqrt{n}} \sum_{i \in I_\ell} \hat{\Delta}_{i\ell} \right| = O_p(\sqrt{n} \|\hat{\alpha}_{1\ell} - \alpha_{01}\|_2 \|\hat{g}_\ell - g_0\|_2) + O_p(\sqrt{n} \|\tilde{\alpha}_{2\ell} - \alpha_{01}\|_2 \|\tilde{h}_\ell - h_0\|_2).$$

Thus, by equation (B.14) and Lemma B.1, Assumption 2.iii in CEINR is satisfied.

To see that Assumption 3.iii in CEINR holds, note that by Assumptions (A1)-(A4): $\mathbb{E}[D_1(W, g))] = \mathbb{E}[\alpha_{01}(Z) g(Z)]$ and $\mathbb{E}[D_2(W, h(\cdot, \varphi(\cdot, \cdot, g)))] = \mathbb{E}[\alpha_{02}(X, V) h(X, \varphi(D, Z, g))]$. Moreover, $D - g_0(Z)$ and $Y - h_0(X, V)$ are orthogonal to Δ_1 and $\Delta_2(g_0)$, respectively. Then,

$$\begin{aligned} \mathbb{E}[\bar{\phi}(W, \varphi(D, Z, g_0), g, h, \alpha_0)] &= \mathbb{E}[\alpha_{01}(Z)(g_0(Z) - g(Z))] + \mathbb{E}[\alpha_{02}(X, V)(h_0(X, V) - h(X, \varphi(D, Z, g)))] \\ &= -\mathbb{E}[D_1(W, g - g_0)] - \mathbb{E}[D_2(W, h(\cdot, \varphi(\cdot, \cdot, g)) - h_0)]. \end{aligned}$$

Thus, by Assumption 5.7, for $\|g - g_0\|_2 < \varepsilon$ and $\|h - h_0\|_2 < \varepsilon$:

$$\begin{aligned} &\left| \mathbb{E}[m(W, g, h, \alpha_0, \theta_0) + \bar{\phi}(W, \varphi(D, Z, g_0), g, h, \alpha_0)] \right| \\ &= \left| \mathbb{E}\left[m(W, g, h, \theta_0) - m(W, g_0, h_0, \theta_0) - D_1(W, g - g_0) - D_2(W, h(\cdot, \varphi(\cdot, \cdot, g)) - h_0) \right] \right| \\ &\leq C \left(\|g - g_0\|_2^2 + \|h(\cdot, \varphi(\cdot, \cdot, g)) - h_0\|_2^2 \right). \end{aligned}$$

The above display, on top of Assumption 5.4.a and Lemma B.1, gives Assumption 3.iii in CEINR for the functional $(g, h) \mapsto \mathbb{E}[m(W, g, h, \alpha_0, \theta_0) + \bar{\phi}(W, \varphi(D, Z, g_0), g, h, \alpha_0)]$.

To conclude, we verify that Lemma 8 in CEINR can be applied to our modified expansion. Being I_ℓ^c all observations not in I_ℓ, note that

$$\begin{aligned} &\mathbb{E}[\hat{R}_{1i\ell} + \hat{R}_{2i\ell} | I_\ell^c] = \mathbb{E}[m(W, \hat{g}_\ell, \hat{h}_\ell, \alpha_0, \theta_0) + \bar{\phi}(W, \varphi(D, Z, g_0), \hat{g}_\ell, \hat{h}_\ell, \alpha_0) | I_\ell^c] \text{ and} \\ &\mathbb{E}[\hat{R}_{3i\ell} | I_\ell^c] = 0. \end{aligned}$$

The last equation follows from orthogonality of $D - g_0(Z)$ and $Y - h_0(X, V)$ to Δ_1 and $\Delta_2(g_0)$, respectively, and Assumption 5.9.c. This means that the strategy of Lemma 8 can be applied to our expansion (for more details, we refer to the proof of the lemma in Chernozhukov et al., 2022a).

We conclude the proof of Theorem 5.1 by providing consistency of $\hat{V}$. Call $\psi_i \equiv \psi(W_i, g_0, h_0, \alpha_0, \theta_0)$ and $\bar{\Psi} \equiv n^{-1} \sum_{i=1}^n \psi_i \psi_i'$. We have that

$$\|\hat{\Psi} - \bar{\Psi}\|_\infty \leq \sum_{\ell=1}^L \frac{1}{n} \sum_{i \in I_\ell} \left(\|\hat{\psi}_{i\ell} - \psi_i\|_\infty^2 + 2 \|\hat{\psi}_{i\ell} - \psi_i\|_\infty \|\psi_i\|_\infty \right)$$

We now expand $\hat{\psi}_{i\ell}(\tilde{\theta}_\ell) - \psi(W_i, g_0, h_0, \alpha_0, \theta_0) = \hat{R}_{1i\ell} + \hat{R}_{2i\ell} + \hat{R}_{3i\ell} + \hat{R}_{4i\ell} + \hat{\Delta}_{i\ell}$, with

$$\hat{R}_{4i\ell} \equiv m(W_i, \hat{g}_\ell, \hat{h}_\ell, \tilde{\theta}_\ell) - m(W_i, \hat{g}_\ell, \hat{h}_\ell, \theta_0)$$

and the remaining terms are given in equation (B.12). Then

$$\frac{1}{n}\sum_{i\in I_\ell}\|\hat{\psi}_{i\ell}-\psi_i\|_\infty^2 \leq C\frac{1}{n}\sum_{i\in I_\ell}\left(\|\hat{R}_{1i\ell}\|_\infty^2+\|\hat{R}_{2i\ell}\|_\infty^2+\|\hat{R}_{3i\ell}\|_\infty^2+\|\hat{R}_{4i\ell}\|_\infty^2+\|\hat{\Delta}_{i\ell}\|_\infty^2\right)$$

by the triangle inequality. The constant C comes from the presence of the interation terms: e.g., $2\|\hat{R}_{1i\ell}\|_\infty\|\hat{R}_{2i\ell}\|_\infty \leq 2\max\{\|\hat{R}_{1i\ell}\|_\infty,\|\hat{R}_{2i\ell}\|_\infty\}$.

Applying Assumptions 5.6.b and 5.6.c to each component of $\hat{R}_{1i\ell}$ and $\hat{R}_{4i\ell}$, respectively, yields $\mathbb{E}[\|\hat{R}_{1i\ell}\|_\infty^2|I_\ell^c] \xrightarrow{P} 0$ and $\mathbb{E}[\|\hat{R}_{4i\ell}\|_\infty^2|I_\ell^c] \xrightarrow{P} 0$. Moreover, by the argument we have followed to show that Assumption 1.ii and 1.ii in CEINR are satisfied: $\mathbb{E}[\|\hat{R}_{2i\ell}\|_\infty^2|I_\ell^c] \xrightarrow{P} 0$ and $\mathbb{E}[\|\hat{R}_{3i\ell}\|_\infty^2|I_\ell^c] \xrightarrow{P} 0$. Also, by the Cauchy-Schwarz inequality, equation (B.14) and Lemma B.1 (applied to each component):

$$\mathbb{E}[\|\hat{\Delta}_{i\ell}\|_\infty^2|I_\ell^c] \leq 3\left(\|\hat{\alpha}_{1\ell}-\alpha_{01}\|_2\|\hat{g}_\ell-g_0\|_2+\|\tilde{\alpha}_{2\ell}-\alpha_{02}\|_2\|\tilde{h}_\ell-h_0\|_2\right)=o_p(1).$$

Thus, collecting the above results:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i\in I_\ell}\|\hat{\psi}_{i\ell}-\psi_i\|_\infty^2\middle| I_\ell^c\right] \leq C\mathbb{E}\left[\|\hat{R}_{1i\ell}\|_\infty^2+\|\hat{R}_{2i\ell}\|_\infty^2+\|\hat{R}_{3i\ell}\|_\infty^2+\|\hat{R}_{4i\ell}\|_\infty^2+\|\hat{\Delta}_{i\ell}\|_\infty^2\middle| I_\ell^c\right]=o_p(1).$$

An application of the conditional Markov inequality gives then $n^{-1}\sum_{i\in I_\ell}\|\hat{\psi}_{i\ell}-\psi_i\|_\infty^2=o_p(1)$. Also, by Assumptions 5.2, 5.3.a, 5.6.a, and 5.6.d: $\mathbb{E}[\psi_i\psi_i'] < \infty$. So, by the Law of Large Numbers, $\bar{\Psi} \xrightarrow{P} \mathbb{E}[\psi_i\psi_i']$. Therefore, by Cauchy-Schwarz:

$$\begin{aligned}\|\hat{\Psi}-\bar{\Psi}\|_\infty &\leq \sum_{\ell=1}^{L}\left[\frac{1}{n}\sum_{i\in I_\ell}\|\hat{\psi}_{i\ell}-\psi_i\|_\infty^2+2\sqrt{\frac{1}{n}\sum_{i\in I_\ell}\|\hat{\psi}_{i\ell}-\psi_i\|_\infty^2}\sqrt{\frac{1}{n}\sum_{i\in I_\ell}\|\psi_i\|_\infty^2}\right]\\ &=o_p(1)+o_p(1)\cdot O_p(1)=o_p(1).\end{aligned}$$

This leads to $\hat{\Psi}=\bar{\Psi}+o_p(1)\xrightarrow{P}\mathbb{E}[\psi_i\psi_i']$. ■

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.

Escanciano, J. C. and Terschuur, J. R. (2022). Debiased semiparametric u-statistics: Machine learning inference on inequality of opportunity. *arXiv preprint arXiv:2206.05235*.

Hahn, J., Liao, Z., Ridder, G., Shi, R., et al. (2022). The influence function of semiparametric two-step estimators with estimated control variables. Unpublished manuscript.

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.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393.

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.

Heckman, J. J. and Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation 1. *Econometrica*, 73(3):669–738.

Huber, P. (1981). *Robust statistics.* New York: Wiley.

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.

Mises, R. v. (1947). On the asymptotic distribution of differentiable statistical functions. *The annals of mathematical statistics*, 18(3):309–348.

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.

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.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. The MIT Press.

Yamamuro, S. (1974). *Differential calculus in topological linear spaces*, volume 374. Springer.