uc3m | Universidad **Carlos III** de Madrid
Departamento de Economía

CERTIFICATE OF ENROLMENT

To whom it may concern,

This is to certify that based on the records field in this University, **FACUNDO ARGAÑARAZ** is officially enrolled in the fourth year of the Ph.D. Program in Economic at Universidad Carlos III de Madrid.

Date of issue: June 14, 2024

Signed: **Juan Carlos Escanciano**

Director of the Ph.D. in Economics – UC3M

UNIVERSIDAD CARLOS III DE MADRID
- DEPARTAMENTO DE ECONOMÍA -

Debiasing Structural Parameters with General Conditional Moments and High-Dimensional First Stages

Facundo Argañaraz[*]
Universidad Carlos III de Madrid

June 14, 2024

Working Paper

[Click here for the latest version]

Abstract

This paper proposes a method to conduct inference on a finite-dimensional parameter in models defined by a finite number of conditional moment restrictions (CMRs), with possibly different conditioning variables and endogenous regressors. Those conditional moments are allowed to depend on non-parametric components, which might be modeled flexibly using Machine Learning tools. Inference is based on locally robust/orthogonal/debiased moments, extended to the case with CMRs. These moments are less affected by regularization bias, which is relevant to machine learning first steps and typically invalidates standard inference. Under weak smoothness conditions, we exploit the CMRs implied by the model in a general way. Thus, our strategy can be applied uniformly in various contexts where the construction of orthogonal moments has not been explored, such as non-linear GMM settings, models with missing data, production functions at the firm level, dynamic discrete choice models, non-linear simultaneous equations models, and many others. Our approach converts a given function of the conditioning variables into a valid instrument that yields a debiased moment, justifying their use over other "ad-hoc" choices of instruments often used in applied work. We argue that this will necessarily require solving functional equations involving unknown terms directly linked to the particular model at hand. However, by imposing an approximate sparsity condition, our method automatically finds the solutions to those equations using a Lasso-type program and thus can be implemented straightforwardly in the same way, regardless of the particular model. Based on this, we introduce a GMM estimator of a finite-dimensional parameter in a Two-Step setting. We derive theoretical guarantees for our construction of orthogonal moments and show the asymptotic normality of the introduced estimator.

Keywords: Conditional Moment Restrictions; Debiased Inference; Machine Learning.

JEL classification: C14; C31; C36.

[*] Department of Economics. E-mail: farganar@eco.uc3m.es. Web-page: https://argafacu.github.io. The author would like to thank Juan Carlos Escanciano for crucial comments and suggestions. Helpful discussions with and precise comments by Christian Hansen are gratefully acknowledged. All errors that remain are the author's own, likely despite warnings by the aforementioned.

1 Introduction

Models defined by conditional moment restrictions (CMRs) are ubiquitous in economics and statistics. They appear in a variety of settings such as regressions, quantile models, dynamic discrete choice models, non-linear simultaneous equations models, missing data, and production functions, among many others (see Chen and Qiu, 2016). Often, to make the model more plausible or less affected by parametric assumptions, researchers consider semiparametric specifications by allowing the CMRs to depend on an unknown potentially infinite-dimensional parameter (e.g., a flexible conditional expectation), apart from the finite-dimensional structural parameter (e.g., Chamberlain, 1992b). This paper presents a *general* method to construct Locally Robust (LR)/Orthogonal/Debiased Moments (Chernozhukov et al., 2022a; Neyman, 1959) in such contexts. Particularly, we introduce a method to conduct inference on a finite-dimensional parameter based on a general approach to obtain debiased moments in settings with semiparametric CMRs. Therefore, this paper paves the way for the application of debiased moments in new and unexplored scenarios relevant in applied work.

To be concrete, we consider models defined by a finite number of CMRs, with possibly different conditioning variables and endogenous regressors (in the sense that the nuisance parameter may be a function depending on variables other than those used for conditioning). In our context, the only information that the researcher has is a number of CMRs that depend on a finite-dimensional parameter of interest and other parameters that belong to a high-dimensional space. A leading case is a Two-Step setting, where certain CMRs identify conditional expectations. Since this nuisance parameter is unknown, it needs to be estimated using those moments in a first stage; to conduct this estimation, the researcher might use recent machine learning methods such as Lasso, Random Forest, Boosting, Neural Networks, and many others, known for their flexibility. As these deal with high-dimensional objects, regularization is necessary to maintain the variance under control, causing a bias that could, in turn, lead to bias in the estimation of the parameter of interest. As this bias typically fails to decay at a rate faster than $\sqrt{n}$, it would invalidate standard inference.

To alleviate this problem, we introduce a new estimator, the *Debiased CMRs Estimator* (D-CMRs), of a finite-dimensional parameter, that exploits debiased moments extended to the case with CMRs. These moments are suitable in this context, as estimation based on them is less affected by regularization bias present in the first stage, compared to a standard GMM procedure based on non-orthogonal moments. Characterizing debiased moments in our general setting is simple. One can show that an orthogonal moment can be obtained as a sum of the products of the residual functions (present in the CMRs) and some functions that we denote Orthogonal Instrumental Variables (OR-IVs), which depend only on the conditioning variables (Argañaraz and Escanciano, 2023). Those OR-IVs belong to a special subclass of Instrumental Variables (IVs), i.e., functions of the conditioning variables. Moreover, as we will argue, they are necessarily solutions to functional equations.

We aim to design an algorithm that allows researchers to estimate these OR-IVs without having to solve those equations explicitly and that can be applied uniformly in various settings with general CMRs. Our motivation stems from the fact that CMRs appear in a variety of applications in distinct ways. For instance, consider a missing data setting, as studied by Graham (2011), and a semiparametric model of production functions estimation at the firm level, as considered by Ackerberg et al. (2014). The only feature these two examples have in common is that they can be reduced to a set of CMRs. Hence, if one seeks to find debiased moments for them, one would need to solve functional equations directly linked to the particular CMRs implied by each model. By nature of the model, these CMRs depend on nuisance parameters, and thus, the functional equations would depend on unknown objects. Consequently, the potential solution will depend on particular unknown terms, and thus estimation needs to be conducted to obtain the OR-IVs. For instance, in the missing data example, they depend on the unknown conditional probability of observing the data, an object that is absent in the production function model. Hence, the estimation strategy that one might propose would again be subject to the model. In contrast, this paper introduces an algorithm that exploits the CMRs in a general way and thus can be applied to any setting, under some smoothness conditions that we specify. Our approach can be regarded as automatic in the sense that it implicitly estimates a solution to these equations. Thus, the researcher does not have to characterize it, while properly dealing with unknown terms.

As discussed by Argañaraz and Escanciano (2023), the key idea is based on constructing a linear operator derived from the CMRs (Carrasco et al., 2007; Luenberger, 1997). The range of this operator collects all the possible Gateaux derivatives of the initial CMRs with respect to the high-dimensional parameter. Then, what we require for a valid OR-IV is that it has to be orthogonal to the range of the aforementioned operator, under a suitable notion of orthogonality. The current work builds on the observation that this orthogonalization can be accomplished by "residualizing" a given known function from the orthogonal projection onto the range of the operator. The challenge stems from the fact that such an operator is unknown, and thus its orthogonal projection is unknown. In this work, we show how such projection can be approximated in practice by exploiting the CMRs implied by the model following the same recipe, regardless of whether the setting is missing data, production functions, or any other context.

Assuming that such orthogonal projection is approximately sparse, we argue that we can obtain an estimator of the solutions of the aforementioned functional equations, i.e., the OR-IVs, using a Lasso-type program. More precisely, under an approximate sparsity assumption for the orthogonal projection, we show how it can be approximated by a linear combination of a number of known basis functions and a sparse finite-dimensional parameter, which is the solution to the Lasso program. Then, the obtained coefficients are used to approximate the OR-IVs. These make the OR-IVs orthogonal to all the possible deviations of the CMRs with respect to the non-parametric component, which is sufficient to obtain a debiased moment. Such an orthogonality condition can be seen as the minimization of a mean squared error. The dimension of the regressors involved in such minimization is allowed to be greater than the sample size, by adding a ℓ_1- norm penalization term. The resulting objective function is then familiar to any Lasso problem. Nonetheless, this Lasso program is special, in the sense that it is based on unknown regressors even though the starting basis functions are given. This is a by-product of not knowing the linear operator that we mentioned in the previous paragraph. Hence, they need to be estimated before minimizing the corresponding Lasso objective function. Conveniently, those regressors take the form of conditional expectations, given the conditioning variables of the model. Consequently, they can be estimated by suitable machine learning tools. In addition, for theoretical and practical reasons, we use cross-fitting. Once these regressors are estimated, the problem is a Lasso regression and as such, it can be implemented straightforwardly by well-known algorithms for each fold in the sample. As these features are particular to our estimation of OR-IVs, we provide theoretical guarantees for our estimation procedure, under the specified sparsity condition.

From a broad perspective, we see our method as a way to justify the use of certain instruments while dealing with CMRs. It is well-known that CMRs typically imply a large number of unconditional moments (Bierens, 1990; Carrasco and Florens, 2000). Although the asymptotic efficiency of estimators can be improved by considering a larger number of moments, an excessive number of them might not be recommendable in practice (Andersen and Sørensen, 1996; Newey and Smith, 2004). In applications then, a finite number of instruments are selected to obtain a finite number of unconditional moments, and estimation is performed using GMM. More often than not, this choice is "ad-hoc". The only choice that is theoretically justified, over other choices, is the so-called optimal IVs as moment based on them yields estimators that are semi-parametric efficient (e.g., Ackerberg et al., 2014; Ai and Chen, 2003, 2012; Chamberlain, 1992b; Chen and Pouzo, 2009). This paper argues that if one is interested in inference in a high-dimensional setting, only OR-IVs should be used, as moments based on them are debiased. Hence, our algorithm justifies the use of OR-IVs over other choices of instruments that do not necessarily yield a debiased moment and that are currently used in applied work. Interestingly, the optimal IV is a special OR-IV (van der Vaart, 1998), but it is not the only one. Indeed, our algorithm is able to convert any suitable function of the condition variables into an OR-IV that can be used in estimation as all these functions can be "residualized" using the approach that we described above. If multiple such functions are provided, then multiple orthogonal moments can be estimated, and GMM can be applied as usual. A natural approach is to start with the typical choices in applied work and transform them using our strategy to obtain debiased moments.

The rationale behind employing GMM stems from our theoretical findings, which assure that standard inference remains valid for the finite-dimensional parameter, despite machine learning tools being employed to estimate nuisance parameters in a first stage. Specifically, leveraging the asymptotic properties established for the OR-IV estimators, we demonstrate that D-CMRs follow an asymptotically normal distribution. For this, we only require that the nuisance parameters be estimated at a rate faster than $n^{1/4}$, which can be achieved by a variety of machine learners. When orthogonal moments are not considered, standard errors need to be correctly computed to account for such a first-stage estimation (Newey, 1994). Nonetheless, we show that standard errors based on our debiased moments properly account for such a first-stage estimation directly by using the usual "sandwich" formula, simplifying their calculation considerably.

We have assessed the finite sample performance of the introduced estimator through a number of Monte Carlo experiments. We acknowledge that one limitation our procedure presents is that it requires several choices by the user. For example, the choice of bases to construct the unknown regressors, the tuning parameter involved in the Lasso program, and the number of folds for cross-fitting, among others. Despite our theory suggesting that, under suitable conditions, these are innocuous for the performance of D-CMRs when n grows, we have been interested in studying the behavior of D-CMRs under distinct combinations of such choices, with finite sample sizes. To this end, we have conducted seven different Monte Carlo experiments in the context of production functions at the firm level, in a high-dimensional setting. Overall, we find that, in these numerical experiments, the finite sample performance of our estimator, in terms of bias and coverage, is satisfactory, regardless of the specific choices made within our methodology, in line with our theoretical results. Particularly, we observe that as the sample size increases, D-CMRs exhibit small bias across all parameters of interest, resulting in point estimates closely approximating the true values. More crucially, our computation of standard errors remains generally valid, despite employing machine learning techniques in the first stage, as evidenced by coverage levels closer to the nominal one. This shows the ability of our procedure to control size. The main takeaway is that our estimation of OR-IVs yields an estimator with good finite sample performance. These observations, then, allow us to be confident about the good properties of our estimation strategy and our theory.

The rest of the paper is organized as follows. Section 2 revises the most related works to ours, emphasizing what our contribution is. Section 3 works out two examples to motivate our results and describe the problem that we face. Section 4 defines LR moments in our context and discusses our approach informally. Section 5 is the core of the paper as it introduces our algorithm in a general setting. Section 6 obtains a convergence rate for the estimators of the OR-IVs. In Section 7 we present an estimator of the finite-dimensional parameter in a two-step setting, and Section 8 provides conditions under which its asymptotic distribution is a standard normal. Section 9 studies the performance of our estimator with finite sample sizes through different Monte Carlo experiments. Section 10 provides final remarks. An Appendix gathers the proofs of all the theoretical results, elaborates on implementation details of the Lasso-type program we introduce, and provides additional details on the Monte Carlo experiments.

2 Related Literature

Our work relates to various strands of the literature. First, this paper is connected to the recently developed literature on debiased moments. The general construction and asymptotic theory of such moments have been established by Chernozhukov, Escanciano, Ichimura, Newey, and Robins (2022a) for a high-dimensional context, i.e., when machine learners are used in a first stage. The debiasing properties of these moments generalize findings in several papers on modern orthogonal moment estimation and inference (see, e.g., Athey et al., 2018; Belloni et al., 2012, 2017; Bravo et al., 2020; Farrell, 2015; Nekipelov et al., 2022; Sasaki and Ura, 2023). Typically, a debiased moment function comprises two components: the initial identifying moment function and a suitable adjustment term. In many applications, this adjustment term is derived as the product of a residual function and the Riesz representer of the original moment function (Ichimura and Newey, 2022). Our paper builds on this literature by specifically examining general semiparametric models defined by CMRs, which may involve varying conditioning variables. While this setting is important in economics and statistics, previous literature has not extensively explored it with a comprehensive scope, as we undertake in this study. An exception is Chernozhukov et al. (2018), who present a general characterization of LR moments for models defined by CMRs. Our paper complements this work in two crucial ways.

Firstly, we consider a setting where we explicitly allow for CMRs depending on different conditioning variables. Secondly, while the approach outlined by Chernozhukov et al. (2018) involves the inversion of an unknown conditional variance-covariance matrix, our proposed method offers a practical solution that avoids cumbersome computations.

Second, our paper contributes to the body of literature focusing on models defined by CMRs. Indeed, our general model encompasses several settings extensively studied in this literature, where the emphasis has primarily been on efficiency. A seminal contribution in this regard is made by Ai and Chen (2012), who establish the semiparametric bound for structural parameters in semiparametric models defined by nested CMRs, and thus possibly different conditioning variables. Their work extends previous results by Chamberlain (1987) for the purely parametric case with one conditioning variable, by Chamberlain (1992b) and Ai and Chen (2003), who study semiparametric models with just one conditioning variable, and findings in Chamberlain (1992a) and Brown and Newey (1998) that consider the parametric case with nested CMRs. Moreover, Ackerberg et al. (2014) analyze CMRs with possibly non-nested or overlapping conditioning sets, where the nuisance functions are just identified parameters by conditional moments, while the structural one is identified through an unconditional moment restriction. Both Ai and Chen (2012) and Ackerberg et al. (2014) propose sieve estimators that are efficient from a semiparametric standpoint. The sieve approach enables them to estimate high-dimensional parameters. We build upon this literature by allowing for the utilization of a wide array of machine learning tools to estimate nuisance parameters in a first stage and automatically constructing LR moments for these models. Furthermore, we provide theoretical bounds for the estimation approach we propose. To the best of our knowledge, ours is the first paper to derive theoretical guarantees for LR moments in general settings defined by CMRs. Efficiency concerns are beyond the scope of this work, but it is an interesting avenue for future research.

Third, our paper also contributes to the literature on the automatic construction of debiased moments. In the context of Riesz representers, important progress has been made. Chernozhukov et al. (2022b) and Chernozhukov et al. (2022c) develop a Dantzig selector for the Riesz representer based on a sparse approximation assumption. Both papers focus on scenarios where the parameter of interest is an average of a linear functional of a regression function, which is estimated by a Dantzig selector as well. Chernozhukov et al. (2022d) propose a Lasso-type minimum distance estimator for the Riesz representer while Chernozhukov et al. (2021) explore neural networks. In both cases, the nuisance parameter can be estimated using general machine learning tools; they also study the case where the parameter of interest is an expectation of a nonlinear functional of the regression function. A unified framework for estimating Riesz representers is developed by Chernozhukov et al. (2020), where the estimation problem is formulated as an adversarial min-max program (Dikkala et al., 2020). While all the previous papers focus exclusively on the exogenous case, i.e., when the nuisance parameter depends on conditioning variables, Bakhitov (2022) proposes a penalized GMM approach for estimating the Riesz representer in endogenous settings. Moreover, Farrell et al. (2021b) proposes an automatic construction of adjustment terms (that yields a debiased moment) in circumstances where the first stage parameter is not necessarily a prediction, but any well-defined parameter that has some meaning in an economic model, and the parameter of interest is an average of some smooth function of the first stage. We extend these works by not focusing exclusively on situations where the parameter of interest represents an expectation. Instead, we deal with the general situation where the model implies conditional moments of arbitrary (but smooth) functions that depend on finite and infinite dimensional parameters. Then, our paper paves the way for extending the automatic construction of debiasing moments developed by those previous works to non-linear GMM settings, prevalent in important contexts such as general missing data problems (Graham, 2011; Hristache and Patilea, 2016, 2017), production functions (Ackerberg et al., 2014; Levinsohn and Petrin, 2003; Olley and Pakes, 1996), dynamic discrete choice models (Hotz and Miller, 1993), non-linear simultaneous equations models (see, e.g., Wooldridge, 2010, Section 14.3), to name a few.[1] We remark that while there is nothing in the estimation of OR-IVs that we propose that exploits the specific meaning of the first stage, we focus on conditional expectations (in the exogenous case) or functions that can be identified inverting conditional expectations (in the endogenous case) since there are well-known results that show their rates of convergence (upon which our asymptotic theory relies) for different machine learners allowing the use of all of them in our setting, while Farrell et al. (2021a) only provide theoretical bounds for Deep neural networks.

Last but not least, this paper builds on previous results by Argañaraz and Escanciano (2023). They present a characterization of debiased moments for settings with a finite number of CMRs, with varying conditioning variables, i.e., the same type of models that we study in this paper. Such a characterization serves for debiasing a general class of parameters, which include smooth functionals of high-dimensional parameters. Argañaraz and Escanciano (2023) do not discuss how such a characterization can be accomplished in practice, i.e., how their theoretical construction can be implemented. This construction is the starting point of the current work. To be precise, we leverage the theoretical results of Argañaraz and Escanciano (2023) to design a simple-to-implement algorithm to construct orthogonal moments in practice, which is data-driven and can be applied systematically in a wide variety of contexts, providing theoretical guarantees for it. Albeit we only focus on debiasing a particular class of structural parameters, excluding functionals of high-dimensional components, we believe that our current work is a crucial starting point for developing subsequent algorithms that can be employed to build debiased moments for more general parameters.

Notation: The norm $||\cdot||$ is a generic norm. For an arbitrary vector $x \in \mathbb{R}^r$, let $||x||_1$ and $||x||_2$ be the ℓ_1 and ℓ_2 norm, respectively. $||x||_0$ be the number of non-zero entries of x. For a random variable $a(W)$, let $||a(W)||_2 = \sqrt{\mathbb{E}\left[a(W)^2\right]}$. For a $n \times r$ matrix $A = [a_{ik}]$, let $||A||_\infty = \max_{i,k} |a_{ik}|$, and $||A||_{\ell_\infty} = \max_i \sum_{k=1}^r |a_{ik}|$. For a set of indexes $S \subseteq \{1, \cdots, r\}$, let x_S be the modification of x that places zeros in all entries of x whose indexes do not belong to S. Moreover, let S_x be the subset of S such that $x_k \neq 0$ for all $k \in S_x$, and S_x^c be the complement of S_x in S. For a bounded set Q, $|Q|$ denotes the cardinality of Q. For $a, b \in \mathbb{R}$, $a \vee b = \max\{a, b\}$. Let P be a probability distribution. Let L^2 be the space of functions of W that are square-integrable, when $W \sim P$, where the precise meaning of W will be established below.[2] In addition, let L_0^2 be a subset of L^2 with the additional mean-zero restriction. Similar definitions apply for objects such as $L^2(V)$ and $L_0^2(V)$ for functions of V,

[1] Additional examples can be found in Section 5 of Chen and Qiu (2016).

[2] Technically, we should index L^2 by some $\sigma-$finite measure. We avoid this to simplify our notation.

an arbitrary random variable. For any arbitrary subset $\mathcal{K}$, let $\overline{\mathcal{K}}$ denote the closure of $\mathcal{K}$ and $\overline{\mathcal{K}}^{\perp}$ be its orthocomplement, when a topology and inner product $\langle \cdot, \cdot \rangle$ is defined. Moreover, $\Pi_{\overline{\mathcal{K}}}$ is the orthogonal projection operator onto $\overline{\mathcal{K}}$. Let $J < \infty$ and $V = (V_1, V_2, \cdots, V_J)'$, we say that a vector-valued function $f(V) = (f_1(V_1), \cdots, f_J(V_J))'$ is in $L^2(V) \equiv \bigotimes_{j=1}^{J} L^2(V_j)$ when each of the elements in a such vector belongs to the corresponding $L^2(V_j)$, $j = 1, \cdots, J$. We let $||f(Z)||_{L^2(Z)} = \sqrt{\sum_{j=1}^{J} ||f_j(V_j)||_2^2}$. Finally, all the CMRs in the sequel are satisfied almost surely (a.s.), but we will not make it explicit to simplify the exposition.

3 Two Motivating Examples

To illustrate the key ideas of this paper we will consider two simple settings throughout. We emphasize that this choice is purely for the sake of clarity in exposition. Our theoretical framework is broad and applicable to models defined by various CMRs, where conditioning variables may vary, and the nuisance parameters may not necessarily be functions of these variables.

Example 1: Missing Data. Suppose that we start with the following identifying moment restriction for the parameter of interest θ_0:

$$\mathbb{E}\left[\rho\left(Y_1, Z_1, \theta_0\right) \middle| Z_1\right] = 0, \tag{3.1}$$

where ρ is some known (up to θ_0) residual function, Z_1 is a vector of exogenous variables and Y_1, e.g., income, is not always observed. In addition, we have a non-missing indicator variable $\delta \in \{0, 1\}$ such that Y_1 is observed iff $\delta = 1$. Also, there is an auxiliary variable T that is always observed. Hence, if we let $X = (Z_1, T)$, we observe $W = (\delta, \delta Y, X)$. Let us also denote $Z = (X, Z_1)$. We assume that $\mathbb{P}\left(\delta = 1 | Y_1, X\right) = \mathbb{P}\left(\delta = 1 | X\right) = \eta_0(X)$, where $\eta_0(X) > 0$ a.s. This model has been considered, e.g., by Hristache and Patilea (2017) and Hristache and Patilea (2016); see also Graham (2011). Under the assumed missingness mechanism, model (3.1) can be equivalently written at the observational level, using the CMRs

$$\mathbb{E}\left[\frac{\delta}{\eta_0(X)} - 1 \middle| X\right] = 0, \tag{3.2}$$

$$\mathbb{E}\left[\frac{\delta}{\eta_0(X)}\rho\left(Y_1, Z_1, \theta_0\right) \middle| Z_1\right] = 0. \tag{3.3}$$

We are interested in learning θ_0, in a context where η_0 is not known. We might proceed by exploiting Equation (3.2) to extract information of η_0, and then use it in the restriction (3.3). □

Example 2: Production Functions at the Firm Level. We observe a panel of n firms across T periods, where i and t index firms and periods, respectively. Let Y_{it} be the output of firm i at time t, and X_{it} be a vector of inputs, e.g., capital and labor. Output is determined by the following equation:[3]

$$Y_{it} = F\left(X_{it}, \theta_{0p}\right) + \omega_{it} + \epsilon_{it}, \tag{3.4}$$

[3] Unless otherwise stated, all the variables are expressed in logarithms.

where ω_{it} is firm i's productivity shock (anticipated productivity) in period t, which is allowed to be correlated with inputs, and ϵ_{it} is noise in output, which is independent and identically distributed (iid), and is assumed to be independent of the current and previous optimal decisions of the firm and anticipated productivities. The function F is assumed to be known up to θ_{0p}. Since ω_{it} is not observed and is correlated with inputs, OLS will provide inconsistent estimates.[4] To address this, we follow the so-called proxy variable approach, started by Olley and Pakes (1996); see also Levinsohn and Petrin (2003) and Wooldridge (2009). We assume that there exists some firm's choice, I_{it}, at t that is linked to ω_{it}:

$$I_{it} = I_t\left(\omega_{it}, X_{it}\right).$$

The precise meaning of variable I_{it}, a "proxy", differs across different formulations of the model.[5] In addition, let us assume that I_t is strictly monotonic in ω_{it}. No parametric assumptions are imposed on I_t. Then, we shall write

$$\omega_{it} = \omega_t\left(I_{it}, X_{it}\right),$$

where ω_t is also non-parametric. Hence, we were able to express the unobservable productivity in terms of observable inputs. It is immediate that Equation (3.4) becomes

$$Y_{it} = F\left(X_{it}, \theta_{0p}\right) + \omega_t\left(I_{it}, X_{it}\right) + \epsilon_{it}. \tag{3.5}$$

Let

$$\eta_{0t}\left(I_{it}, X_{it}\right) = F\left(X_{it}, \theta_{0p}\right) + \omega_t\left(I_{it}, X_{it}\right).$$

Then, by the independence assumption,

$$\mathbb{E}\left[Y_{it} - \eta_{0t}\left(I_{it}, X_{it}\right)\middle| I_{it}, X_{it}\right] = 0. \tag{3.6}$$

While Equation (3.6) identifies η_{0t}, it is not enough to identify all the parameters of the production function. This is true since X_{it} enters parametrically and non-parametrically in (3.5). The last element of the model is the evolution of ω_{it}. Typically, this is also treated non-parametrically, as in Olley and Pakes (1996) and Levinsohn and Petrin (2003). We, nonetheless, follow Ackerberg et al. (2014) and work with a more "natural" semiparametric model. Let us assume that ω_{it} follows a First-Order Markov's process in the sense that

$$\mathbb{E}\left[\omega_{it}\middle| X_{i,t-1}, I_{i,t-1}\cdots, \omega_{i,t-1}, \omega_{i,t-2}, \ldots, \omega_{i,0}\right] = \mathbb{E}\left[\omega_{it}\middle| \omega_{i,t-1}\right]. \tag{3.7}$$

Equation (3.7) is indeed assumed by Olley and Pakes (1996) and Levinsohn and Petrin (2003). What Ackerberg et al. (2014) suggests is to parameterize (3.7). To keep things simple, let us consider

$$\mathbb{E}\left[\omega_{it}\middle| \omega_{i,t-1}\right] = \theta_{0\omega}\omega_{i,t-1}. \tag{3.8}$$

[4] In the case where F is assumed to be linear in inputs.

[5] For instance, Olley and Pakes (1996) considers in I_{it} the firm's current investment towards future physical capital. In Levinsohn and Petrin (2003), I_{it} is the firm's choice of an intermediate input, e.g., electricity or material input.

Then, letting Ω_{it} be the firm information set at t, with $(X_{it}, I_{it}) \subseteq \Omega_{it}$, and using the independence assumption, Equations (3.5), (3.7), and (3.8), it is not difficult to show that

$$\mathbb{E}\left[Y_{it} - F\left(X_{it}, \theta_{0p}\right) - \theta_{0\omega}\left(\eta_{0,t-1}\left(Z_{i,t-1}\right) - F\left(X_{i,t-1}, \theta_{0p}\right)\right) \middle| \, \Omega_{i,t-1}\right] = 0. \tag{3.9}$$

Suppose that $T = 3$, i.e., firms are observed during three periods. Ignoring subscript i, and denoting $\eta_0 \equiv (\eta_{01}, \eta_{02})$, the model can be defined by the following CMRs:

$$\mathbb{E}\left[Y_1 - \eta_{01}\left(I_1, X_1\right) \middle| \, I_1, X_1\right] = 0, \tag{3.10}$$

$$\mathbb{E}\left[Y_2 - F\left(X_2, \theta_{0p}\right) - \theta_{0\omega}\left(\eta_{01}\left(I_1, X_1\right) - F\left(X_1, \theta_{0p}\right)\right) \middle| \, \Omega_1\right] = 0, \tag{3.11}$$

$$\mathbb{E}\left[Y_2 - \eta_{02}\left(I_2, X_2\right) \middle| \, I_2, X_2\right] = 0, \tag{3.12}$$

$$\mathbb{E}\left[Y_3 - F\left(X_3, \theta_{0p}\right) - \theta_{0\omega}\left(\eta_{02}\left(I_2, X_2\right) - F\left(X_2, \theta_{0p}\right)\right) \middle| \, \Omega_2\right] = 0. \tag{3.13}$$

Let $Y = (Y_1, Y_2, Y_3, X_2, X_3)$, $X = (I_1, X_1, I_2, X_2)$, $Z = (X, \Omega_1, \Omega_2)$, and $W = (Y, X, Z)$. Hence, based on W, our goal is to learn $\theta_0 = \left(\theta_{0p}', \theta_{0\omega}\right)'$, the parameter of interest, in the presence of an unknown η_0. To this end, we might exploit (3.10)/(3.12) to estimate η_0, and then plug the resulting estimator into moments based on (3.11)/(3.13) for estimation of θ_0. $\square$

Both Example 1 and Example 2 share a common characteristic: η_0 is unknown and ultimately needs to be estimated to estimate θ_0. One approach to estimating η_0 involves employing a low-dimensional framework, e.g., with a polynomial series using a small number of terms. For instance, a popular Stata command for estimation of production functions, introduced in Petrin et al. (2004), as a default option, treats the part of η_0 that depends on $\omega_t\left(I_t, X_t\right)$ as a third-degree polynomial. We, instead, model η_0 using machine learning tools, which includes the low-dimensional sieve approach of Petrin et al. (2004) as a special case.[6] These tools (e.g., Lasso, random forest, neural networks, and boosting) have proven useful in dealing with situations where functions of covariates (where the leading case involves conditional expectations) need to be estimated without imposing stringent parametric assumptions.

With a highly complex η_0, it would be difficult to estimate θ_0 without bias. This is a by-product of the regularization techniques that all those algorithms impose in estimation. Moreover, that bias would typically decay at slow rates, slower than $\sqrt{n}$ (Chernozhukov et al., 2022a). Consequently, a valid concern is that such a first-stage bias would be translated into bias in the estimation of θ_0, and then it would not hold that $\sqrt{n}\left(\hat{\theta}_{Non-orth} - \theta_0\right)$ is normally distributed. Indeed, $\sqrt{n}\left(\hat{\theta}_{Non-orth} - \theta_0\right)$ would not be $O_p(1)$, invalidating standard inference on $\hat{\theta}_{Non-orth}$. To illustrate the point, Figure 1 displays (in orange) the empirical distribution of the standardized $\left(\hat{\theta}_{Non-orth} - \theta_0\right)$ for one of the parameters of the production function of Example 2 obtained from simulated data on 1,000 firms and 1,000 Monte Carlo repetitions. This is computed by estimating η_0 using Random Forest in the first stage and then estimating θ_0 using GMM.[7] The figure illustrates that this estimator is substantially

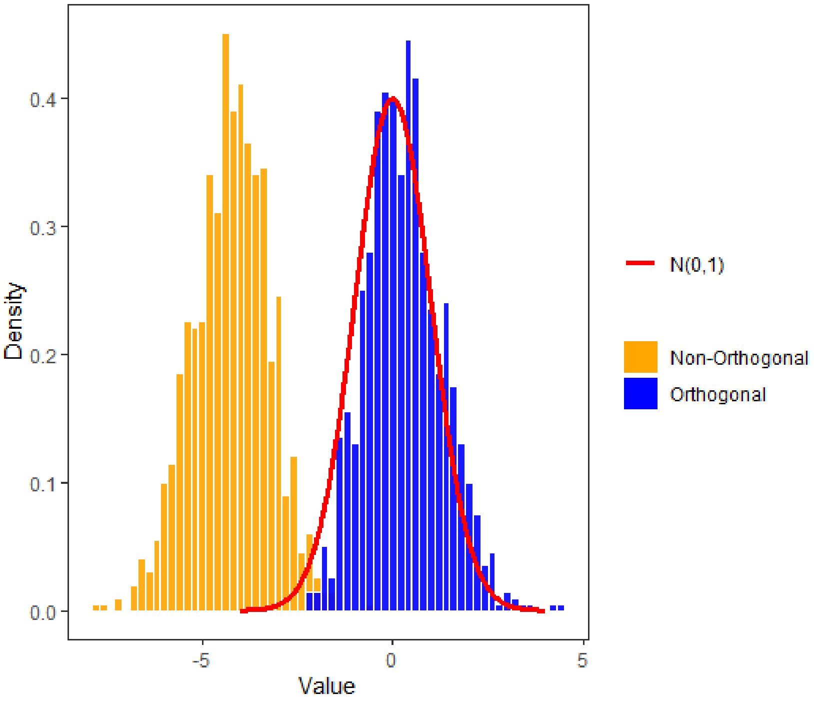
[6] The only generalization of Petrin et al. (2004)'s proposal that we are aware of is Cha et al. (2023), who model the nonparametric component using a high-dimensional basis. Notice that we allow for the use of many other suitable nonparametric estimation strategies.

[7] For this illustration, we have used the R function `ranger` and its default options.

biased, as its mean is not centered around zero. Furthermore, the shape of the distribution is certainly different from the standard normal distribution (depicted by the red curve), which would be similar to the expected asymptotic distribution if the bias were absent.

Therefore, our ultimate goal is to construct moments that, under regularity conditions, can be used for estimation of θ_0, and that will lead to an estimator that would remain $\sqrt{n}-$consistent, even when η_0 has been estimated flexibly. Such moments, known as LR/Orthogonal/Debiased moments, play a crucial role in achieving this goal. A key object will be a function of the conditioning variables Z that satisfies an important property. We will argue that to compute this function one needs to solve functional equations. The particular terms involved in it as well as the plausibility of it would be extremely linked to the particular setting that one is working on. Our aim is to construct LR moments in a data-driven manner that can be uniformly applied across various settings, e.g., missing data and production functions. This is the main contribution of this work. Moreover, the researcher does not have to derive explicit expressions for such functions, in the spirit of the fairly recent literature of automatic estimation of Riesz representers (cf. Section 2). The algorithm will provide the user with a suitable adjustment term that can subsequently be employed in the estimation of θ_0. This approach results in an estimator that indeed yields a $\sqrt{n}\left(\hat{\theta}_{Orth} - \theta_0\right)$ normally distributed. Figure 1 displays the standardized empirical distribution of such an estimator (in blue). Notably, it closely resembles the expected asymptotic distribution and, crucially, is centered around zero, implying that our construction is less affected by the first-stage bias. To accomplish this, we will show that the problem reduces to a Lasso-type problem, under an approximate sparsity condition. Thus, implementing our approach is relatively straightforward.

Figure 1: Comparison of Non-Orthogonal and Orthogonal Estimators



NOTE: The figure shows the histogram of the standardized and centered estimator $\left(\hat{\theta}-\theta_0\right)$ for one of the parameters of the production function of Example 2, using Non-orthogonal and Orthogonal moments, based on the construction we propose. The density of the standard normal distribution is also displayed in red. The sample size is $n = 1,000$. Results are based on $1,000$ Monte Carlo repetitions.

4 Debiased Moments and Informal Discussion

The first step in our automatic construction involves theoretically characterizing debiased moments (see definition below) within the specific setting that the researcher is concerned with. Hence, following the recent literature on debiasing moments in high-dimensional contexts (cf. Section 2), we aim to construct debiased moments for Example 1 and Example 2.

We provide a precise notion of a LR moment for θ_0 in our setting. Let $\kappa \in L^2(Z)$ be an arbitrary vector-valued function. Let κ_0 be an element of a *special* subset of $L^2(Z)$, which will be described below. Moreover, let $\eta_0 \in \boldsymbol{B}$, where $\boldsymbol{B}$ is a possibly infinite dimensional vector space. A debiased moment in our setting is a moment based on a function $\psi : \mathcal{W} \times \Theta \times \boldsymbol{B} \times L^2(Z) \mapsto \mathbb{R}$ satisfying the following two restrictions:

$$\frac{d}{d\tau}\mathbb{E}\left[\psi\left(W,\theta_0,\eta_0+\tau b,\kappa_0\right)\right]=0, \quad \text{for all } b \in \boldsymbol{B}, \tag{4.1}$$

$$\mathbb{E}\left[\psi\left(W,\theta_0,\eta_0,\kappa\right)\right]=0, \quad \text{for all } \kappa \in L^2(Z), \tag{4.2}$$

where $\frac{d}{d\tau}$ denotes derivatives from the right (i.e., from non-negatives values of τ). Equation (4.1) implies that the Gateaux derivative at η_0 of a moment based on ψ is zero. Here b represents a possible direction of deviation from η_0, and belongs to $\boldsymbol{B}$. Intuitively, local perturbations to the nuisance parameter do not affect the moment. This is an appealing property as it would be hard to learn η_0 exactly, which is particularly true in high-dimensional contexts. This is the reason why the estimation of θ_0 based on a moment that is orthogonal is less affected by first-stage bias, typically present in the estimation of η_0, and will be a key attribute to establish standard inference results on an estimator of θ_0. As we will show below, we need to introduce the nuisance parameter κ_0 to achieve (4.1). Nonetheless, this would not affect the moment itself as Equation (4.2) implies that the expectation of ψ is *globally* insensitive to deviations from κ_0.

Now, how can we construct moments that satisfy the key properties (4.1)-(4.2) in the contexts of Example 1 and Example 2? The basic idea will be to obtain a function κ_0 that is orthogonal to a set that contains all the possible derivatives of the CMRs with respect to η. This corresponds to a more general result of orthogonality for general functionals in models defined by several CMRs, studied by Argañaraz and Escanciano (2023).

Example 1: Under regularity conditions, it is not difficult to see that an orthogonal moment in this case will be based on

$$\psi\left(W, \theta_0, \eta_0, \kappa_0\right) = \left(\frac{\delta}{\eta_0(X)} - 1\right)\kappa_{01}(X) + \frac{\delta}{\eta_0(X)}\rho\left(Y_1, Z_1, \theta_0\right)\kappa_{02}(Z_1), \tag{4.3}$$

where $\kappa_0 = (\kappa_{01}, \kappa_{02}) \in L^2(X) \times L^2(Z_1)$ is such that

$$\frac{d}{d\tau}\mathbb{E}\left[\psi\left(W, \theta_0, \eta_0 + \tau b, \kappa_0\right)\right] = \mathbb{E}\left[\frac{\delta b(X)}{\eta_0^2(X)}\left(-\kappa_{01}(X) - \rho\left(Y_1, Z_1, \theta_0\right)\kappa_{02}(Z_1)\right)\right] = 0, \quad \text{for all } \ b \in \boldsymbol{B}. \tag{4.4}$$

Then, κ_0 satisfies (4.1). In addition, as $\kappa_0 \in L^2(Z)$, condition (4.2) also holds. □

Example 2: For this case, given some smoothness conditions, we can obtain a debiased moment by means of

$$\begin{aligned}\psi\left(W, \theta_0, \eta_0, \kappa_0\right) = {} & \left(Y_1 - \eta_{01}\left(I_1, X_1\right)\right)\kappa_{01}\left(Z_1\right) + \left(Y_2 - F\left(X_2, \theta_{0p}\right) - \theta_{0\omega}\left(\eta_{01}\left(Z_1\right) - F\left(X_1, \theta_{0p}\right)\right)\right)\kappa_{02}\left(Z_1\right) \\ & + \left(Y_2 - \eta_{02}\left(Z_2\right)\right)\kappa_{03}\left(Z_2\right) + \left(Y_3 - F\left(X_3, \theta_{0p}\right) - \theta_{0\omega}\left(\eta_{02}\left(Z_2\right) - F\left(X_2, \theta_{0p}\right)\right)\right)\kappa_{04}\left(Z_2\right),\end{aligned} \tag{4.5}$$

where $Z_1 = (I_1, X_1)$, $Z_2 = (I_2, X_2)$, $\kappa_0 = (\kappa_{01}, \kappa_{02}, \kappa_{03}, \kappa_{04}) \in L^2(Z_1) \times L^2(Z_1) \times L^2(Z_2) \times L^2(Z_2)$ is such that

$$\begin{aligned}\frac{d}{d\tau}\mathbb{E}\left[\psi\left(W, \theta_0, \eta_0 + \tau b, \kappa_0\right)\right] &= \mathbb{E}\left[b_1\left(Z_1\right)\left(-\kappa_{01}\left(Z_1\right) - \theta_{0\omega}\kappa_{02}\left(Z_1\right)\right) + b_2\left(Z_2\right)\left(-\kappa_{02}\left(Z_2\right) - \theta_{0\omega}\kappa_{02}\left(Z_2\right)\right)\right] \\ &= 0, \quad \text{for all } \ b \in \boldsymbol{B}.\end{aligned} \tag{4.6}$$

Then, we can verify that (4.5) indeed satisfies (4.1)-(4.2). □

Expressions (4.3) and (4.5) indicate that we can obtain debiased moments by linearly combining the initial CMRs: sum the products of the initial residual functions and the elements in κ_0. This linear combination is special in the sense that we are required to find $\kappa_0 \in L^2(Z)$ such that the moment is invariant to local perturbations to η. Therefore, by construction, the resulting moments (4.5)/(4.3) are debiased. Since κ_0 yields an orthogonal moment by properly combining the initial residual functions, we denote them Orthogonal Instrumental Variables (OR-IVs).

In both situations, obtaining a valid κ_0 entails solving functional equations (4.4)/(4.6). Various approaches can be pursued to achieve this goal. One option is to directly characterize the solutions to (4.4)/(4.6). However, this strategy may not be ideal, as its plausibility hinges on the specific expressions involved in the functional equations defining κ_0. Consequently, there is no guarantee that a particular procedure for solving κ_0 in one model can be readily applied in other contexts. For instance, solving Equation (4.4) differs from solving Equation (4.6), given that the former involves η_0, while κ_{01} and κ_{02} are functions of different random variables. Moreover, Equation (4.6) requires finding four terms, instead of two, although each pair involves functions of the same random variables. In more complex settings, additional terms may emerge. Instead of relying on the tractability of the equations defining the OR-IVs, we propose an approach that can be applied generally. In addition, as it is evident from (4.4)/(4.6), these equations might involve unknown quantities, and thus the direct computations might not be feasible. Our goal is to design a *feasible* algorithm where unknown quantities are treated intelligently.

Another possibility, inspired by the literature on the automatic construction of Riesz representers for orthogonal moments (see Bakhitov, 2022; Chernozhukov et al., 2022d; Ichimura and Newey, 2022), is to view Equations (4.4)/(4.6) as moment conditions for κ_0. However, we cannot directly apply the same techniques developed for Riesz representers. There are two main reasons for this. First, unlike Riesz representers, which are unique, for reasons that will become clear below, κ_0 is not uniquely identified—there might be more than one OR-IV. Second, in our setting, the trivial solution is always a solution, making $\kappa_0 = 0$ an OR-IV, which leads to the orthogonal moment $\psi = 0$, providing no information about θ_0. To avoid trivial solutions, one might consider imposing an additional constraint to disregard them. However, determining the most suitable approach to impose such restrictions can be challenging and context-dependent. For example, one might attempt to estimate the null space associated with some unknown linear operator defined by (4.4)/(4.6). Although this could theoretically work, estimating unknown null spaces is likely unfamiliar to the average practitioner, which is why we opted not to pursue this route further. Instead, we propose a more "natural" idea based on a Lasso-type program, as detailed below.

Finally, in cases where the conditioning variables are fixed across the CMRs, Chernozhukov et al. (2018) (Section 2.2.4) present a general expression for κ_0's. This approach has some caveats. The formula involves unknown quantities and thus one might not apply it directly. More importantly, it requires the inversion of an unknown conditional variance-covariance matrix, which is challenging to handle in practice (see Equations (2.23)-(2.25) in Chernozhukov et al. (2018)).

Differently from all the previous ideas, we aim to design a procedure that allows the researcher to find functions κ_0's without solving complicated equations. Additionally, we want to obtain these functions without dealing with sensitive terms such as conditional variances that need to be inverted. More importantly, we aim to provide a method that can be applied in general settings, independent of the specific structure of the model. Our algorithm will convert any vector of instruments $f \in L^2(Z)$ into a unique and valid κ_0. In other words, our approach will transform any suitable function of the conditioning variables such that it satisfies the equations implied by (4.4) or (4.6). If multiple functions are considered, several ψ's can be constructed and orthogonal moments derived, and hence GMM can be applied as usual to estimate θ_0. An estimator $\hat{\theta}$ based on this will be shown to be asymptotically normally distributed. Hence, the associated "sandwich" formula will yield proper standard errors, which account for the first stage estimation of η_0 and κ_0.

Let us discuss informally the core idea of our procedure in the context of Example 1. We start with an *arbitrary* and *known* function $f = (f_1, f_2) \in L^2(Z)$, chosen by the researcher. For instance, $f_1(X) = T$ and $f_2(Z_1) = Z_1$, provided that these variables have finite second moments. Under suitable conditions and for suitable vectors $\hat{M}_1$ and $\hat{M}_2$, our algorithm will find some finite-dimensional vector $\hat{\beta}$ such that an estimator $\hat{\kappa}$ is constructed as follows:

$$\hat{\kappa}_1(X) = f_1(X) - \hat{M}_1'\hat{\beta}, \tag{4.7}$$

$$\hat{\kappa}_2(Z_1) = f_2(Z_1) - \hat{M}_2'\hat{\beta}, \tag{4.8}$$

and satisfies (4.4), with probability approaching one.

In particular, we interpret $\hat{\beta}$ as being the solution to a Lasso-type problem with regressors $\hat{M}_1$ and $\hat{M}_2$. Essentially, the vector $\hat{\beta}$ makes $\hat{\kappa}$, constructing as in (4.7)-(4.8), orthogonal to a linear operator whose range is equal to all the possible deviations of the moment based on ψ. This orthogonality condition can be seen as the minimization of a mean squared error. As we allow for the dimension of each of the regressors $\hat{M}$'s above to be greater than the sample size, we add a ℓ_1-norm penalization term. Hence, our problem of finding $\hat{\beta}$ can be regarded as the solution to a Lasso problem. Under an approximate sparsity condition, we show that if this $\hat{\beta}$ is plugged into (4.7)-(4.8), a valid $\hat{\kappa}$ and thus a debiased moment can be obtained. Therefore, starting from an arbitrary f, we were able to construct an orthogonal moment.

Based on this construction, we derive below a convergence rate for our OR-IV estimator, $\hat{\kappa}$. Letting $\hat{\eta}$ be a suitable machine learning estimator of η_0, under the key condition

$$\sqrt{n}\,||\hat{\eta} - \eta_0||\,||\hat{\kappa} - \kappa_0|| \to 0,$$

which allows slower than $\sqrt{n}$ rates for $\hat{\eta}$ and $\hat{\kappa}$, we prove the asymptotic normality of a two-step GMM estimator $\hat{\theta}$, implying that standard inference can be conducted on it straightforwardly. Debiasing plays a pivotal role in this derivation. The following sections develop these ideas in a general framework, derive the exact program that β has to solve, and provides the technical conditions required. Readers less interested in these details can skip these sections, except for Section 7 where we present our estimator.

5 Computation of the OR-IVs

5.1 General Setting

Let us introduce the general setting. The data $W_i = (Y_i, X_i, Z_i)$, $i = 1, \cdots, n$, is iid with support $\mathcal{W}$, where Y is a random vector of endogenous variables taking values in $\mathcal{Y} \subseteq \mathbb{R}^{d_Y}$, X is another random vector of potentially endogenous variables, and Z is random vector of exogenous variables. Let $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ denote a finite-dimensional parameter vector. In addition, let $\eta \in \boldsymbol{B}$ be a vector of real-valued measurable functions of X that may depend on additional unknown parameters that we do not specify, and $\boldsymbol{B}$ is some linear vector space. To be specific, $\eta = \left(\eta_1, \cdots, \eta_{d_\eta}\right)$ with $\eta_s \equiv \eta_s(X)$. We assume that there is a vector of residual functions $m_j : \mathcal{Y} \times \Theta \times \boldsymbol{B} \mapsto \mathbb{R}$ such that:

$$\mathbb{E}\left[\left. m_j\left(Y, \theta_0, \eta_0\right)\right| Z_j\right] = 0, \quad \mu_j - a.s., \quad j = 1, 2, \cdots, J, \tag{5.1}$$

where $\mathbb{E}[\cdot]$ is expectation under the distribution of Y given Z_j, μ_j is probability measure of Z_j, $Z = (Z_1, \cdots, Z_J)$, and each m_j is known up to the parameters (θ_0, η_0). To be precise, m_j might depend on θ_0 arbitrarily. Observe that we are not imposing anything regarding how the conditioning variables relate. These might have all or some elements in common. Note that in the case where Z_j is a constant, for some j, we have an unconditional moment. Hereafter, we assume that there exists a unique $(\theta_0, \eta_0) \in \Theta \times \boldsymbol{B}$ such that (5.1) holds.

As before, let $\kappa = (\kappa_1, \cdots, \kappa_J)$, where $\kappa_j \equiv \kappa_j(Z_j)$, and $\kappa_j \in L^2(Z_j)$, $1 \leq j \leq J$. Hence, we say $\kappa \in L^2(Z)$, where $L^2(Z) = \bigotimes_{j=1}^J L^2(Z_j)$. Let $\boldsymbol{B} \subseteq \bigotimes^{d_\eta} L^2(X)$ be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\boldsymbol{B}}$ and define

$$h_j\left(Z_j, \theta, \eta\right) = \mathbb{E}\left[\left. m_j\left(Y, \theta, \eta\right)\right| Z_j\right].$$

Our results will rely heavily on smoothness conditions of the functions g_j's. In particular, throughout we maintain the key assumption:

Assumption 1. *Given some $||\cdot||$, $h_j\left(Z_j, \theta_0, \cdot\right) : \boldsymbol{B} \mapsto L^2(Z_j)$ is Fréchet differentiable in a neighborhood of η_0, where the derivative is given by*

$$\begin{aligned}\left[\nabla h_j\left(Z_j, \theta_0, \eta_0\right)\right](b) &\equiv \frac{d}{d\tau} h_j\left(Z_j, \theta_0, \eta_0 + \tau b\right) \\ &= \left[S^{(j)}_{\theta_0, \eta_0} b\right]\left(Z_j\right),\end{aligned} \tag{5.2}$$

for some $b \in \boldsymbol{B}$.

We make the observation that (5.2) defines a linear operator $S^{(j)}_{\theta_0,\eta_0} : \boldsymbol{B} \mapsto L^2(Z_j)$ (Carrasco et al., 2007; Luenberger, 1997). In addition, let us define

$$S_{\theta_0,\eta_0} b = \left(S^{(1)}_{\theta_0,\eta_0} b, \cdots, S^{(J)}_{\theta_0,\eta_0} b\right).$$

Then, $S_{\theta_0,\eta_0} : \boldsymbol{B} \mapsto L^2(Z)$ is also a linear operator. We equipped $L^2(Z)$ with the inner product $\langle f_1(Z), f_2(Z) \rangle_{L^2(Z)} = \sum_{j=1}^J \mathbb{E}\left[f_{1j}(Z_j) f_{2j}(Z_j)\right]$, where $f_1 = (f_{11}, \cdots, f_{1J})$ and $f_2 = (f_{21}, \cdots, f_{2J})$. Therefore, $L^2(Z)$ is a Hilbert space. The range of that operator can be defined as follows

$$\mathcal{R}\left(S_{\theta_0,\eta_0}\right) = \left\{f \in L^2(Z) : f = S_{\theta_0,\eta_0} b \text{ for some } b \in \boldsymbol{B}\right\}.$$

A key object for us is $\overline{\mathcal{R}\left(S_{\theta_0,\eta_0}\right)}^{\perp}$, i.e., the orthogonal complement of the closure of the range of S_{θ_0,η_0} in $L^2(Z)$, which can be defined as

$$\overline{\mathcal{R}\left(S_{\theta_0,\eta_0}\right)}^{\perp}=\left\{f\in L^2\left(Z\right):\sum_{j=1}^{J}\mathbb{E}\left[f_j\left(Z_j\right)h_j\left(Z_j\right)\right]=0,\quad\text{for all}\quad h=\left(h_1,\cdots,h_J\right)\in\overline{\mathcal{R}\left(S_{\theta_0,\eta_0}\right)}\right\}.$$

Let $\kappa_0\in\overline{\mathcal{R}\left(S_{\theta_0,\eta_0}\right)}^{\perp}$. Then, it can be easily verified that a debiased moment in model (5.1) can be constructed as follows:

$$\psi\left(W,\theta_0,\eta_0\right)=\sum_{j=1}^{J}m_j\left(Y,\theta_0,\eta_0\right)\kappa_{0j}\left(Z_j\right). \tag{5.3}$$

We point out two important observations. First, κ_0 might not be unique. In fact, there can potentially exist an infinite number of such κ_0's. Despite this, not every IV, i.e., functions in $L^2\left(Z\right)$, belongs to $\overline{\mathcal{R}\left(S_{\theta_0,\eta_0}\right)}^{\perp}$, and thus not every IV may serve as a valid κ_0. Consequently, choices of instruments commonly made in applied work might not lead to orthogonal moments. Nevertheless, we will demonstrate how to convert any possible function in $L^2\left(Z\right)$ into a valid κ_0. Therefore, one may begin with the common choices of instrument functions and then apply our transformation to directly obtain an orthogonal moment.

Second, κ_0 does not necessarily exist. This situation occurs when $\overline{\mathcal{R}\left(S_{\theta_0,\eta_0}\right)}^{\perp}=\{0\}$. This implies that the only valid OR-IV is $\kappa_0=0$, which will lead to a trivial LR moment that cannot be used to learn θ_0. Following the terminology in Argañaraz and Escanciano (2023), when $\overline{\mathcal{R}\left(S_{\theta_0,\eta_0}\right)}^{\perp}=\{0\}$ we say that the model satisfies a *local surjectivity* property.[8] A practical implication of this situation is that this would result in a trivial orthogonal moment, i.e., $\psi=0$. Unfortunately, the plausibility of encountering this situation depends on the particular model at hand.[9]

5.2 Estimation of OR-IVs

We next explain how to obtain OR-IVs automatically in the general model (5.1). We need to impose some conditions to simplify our automatic construction below while still maintaining a general framework. We assume

Assumption 2. *(i) There exists a known (up to θ_0 and η_0) function ν_j such that*

$$\left[S^{(j)}_{\theta_0,\eta_0}b\right]\left(Z_j\right)=\mathbb{E}\left[\nu_j\left(Y,\theta_0,\eta_0,b\right)\middle|\,Z_j\right];$$

(ii) $\nu_j\left(Y,\theta_0,\eta_0,b\right)=b\left(X\right)'\tilde{\nu}_j\left(Y,\theta_0,\eta_0\right)$, for some $d_\eta-$vector of known (up to θ_0 and η_0) functions $\tilde{\nu}_j$.

[8] Similar notions have appeared elsewhere, e.g., in Bonhomme (2012).

[9] Using our approach for estimating OR-IVs, we may check for local surjectivity in practice. For example, local surjectivity implies $\hat{\kappa}\approx 0$ for each individual in the sample for a large class of initial functions f's. An alternative approach consists of using duality theory and exploiting the fact that $\overline{\mathcal{R}\left(S_{\theta_0,\eta_0}\right)}^{\perp}=\mathcal{N}\left(S^{*}_{\theta_0,\eta_0}\right)$, where $S^{*}_{\theta_0,\eta_0}$ is the adjoint operator of S_{θ_0,η_0} and $\mathcal{N}\left(\cdot\right)$ denotes the null space of an operator (Luenberger, 1997, Theorem 6.6.3). Theoretically checking that $\mathcal{N}\left(S^{*}_{\theta_0,\eta_0}\right)=\{0\}$ might be easier than studying if $\overline{\mathcal{R}\left(S_{\theta_0,\eta_0}\right)}^{\perp}=\{0\}$.

(iii) For all $b \in \boldsymbol{B}$, $\mathbb{E}\left[b(X)' \tilde{\nu}_j (Y, \theta_0, \eta_0) \middle| Z_j \right] \in L^2(Z_j)$.

Assumption 2 (i) says that there exists a function ν_j that links the direction b to the operator $S^{(j)}_{\theta_0,\eta_0}$ through the conditional expectation given Z_j. Assumption 2 (ii) requires that the function ν_j has to be linear in the direction b. Assumption 2 (iii) assures the well-definiteness of the operator for all $b \in \boldsymbol{B}$. Notice that assuming the residual functions m_j's are sufficiently smooth such that their Fréchet differentiable exists and the interchange between derivatives and integrals holds, Assumption 2 (i) and (ii) are satisfied. Assumption 2 (iii) is a high-level condition that restricts the conditional distribution of the data given Z_j and simplifies our calculations below.[10] We can easily verify that Assumption 2 holds in Example 1 and Example 2.

As we explained above, our goal is to find a function $\kappa_0 \in L^2(Z)$ such that it is orthogonal to $\overline{\mathcal{R}(S_{\theta_0,\eta_0})}$. This implies that

$$\sum_{j=1}^{J} \mathbb{E}\left[\mathbb{E}\left[\nu_j (Y, \theta_0, \eta_0, b) \middle| Z_j \right] \kappa_{0j}(Z_j) \right] = 0,$$

for all $b \in \boldsymbol{B}$. We can obtain such κ_0 by picking some function $f \in L^2(Z)$, and then computing $\kappa_0 = f - \Pi_{\overline{\mathcal{R}(S_{\theta_0,\eta_0})}} f$, where recall that $\Pi_{\overline{\mathcal{R}(S_{\theta_0,\eta_0})}}$ denotes the orthogonal projection operator onto $\overline{\mathcal{R}(S_{\theta_0,\eta_0})}$, and is defined as follows

$$\Pi_{\overline{\mathcal{R}(S_{\theta_0,\eta_0})}} f := \underset{\tilde{f} \in \overline{\mathcal{R}(S_{\theta_0,\eta_0})}}{\arg\min} \sum_{j=1}^{J} \mathbb{E}\left[\left(f_j(Z_j) - \tilde{f}_j(Z_j) \right)^2 \right]. \tag{5.4}$$

By the Projection Theorem, (5.4) exists and is unique (see Luenberger, 1997, Theorem 3.3.2). Next, we exploit the following facts. Notice that $\Pi_{\overline{\mathcal{R}(S_{\theta,\eta_0})}} f \in \overline{\mathcal{R}(S_{\theta_0,\eta_0})}$, and that the range of $S_{\theta_0,\eta_0} S^*_{\theta_0,\eta_0}$ is dense in $\overline{\mathcal{R}(S_{\theta_0,\eta_0})}$, where $S^*_{\theta_0,\eta_0}$ is the adjoint operator of S_{θ_0,η_0}. The following proposition provides a general expression for this operator, which will be useful in the sequel.

Proposition 1. *Suppose Assumptions 1 and 2 hold. In addition, suppose that* S_{θ_0,η_0} *is bounded and* $\langle b_1, b_2 \rangle_{\boldsymbol{B}} = \mathbb{E}\left[b_1(X)' b_2(X) \right]$. *Then, the adjoint* $S^*_{\theta_0,\eta_0} : L^2(Z) \mapsto \boldsymbol{B}$ *exists, is lineal, continuous, and is given by*

$$\left[S^*_{\theta_0,\eta_0} g \right](X) = \sum_{j=1}^{J} \mathbb{E}\left[\tilde{\nu}_j (Y, \theta, \eta_0) \, g_j(Z_j) \middle| X \right].$$

Let $\Pi_{\overline{\mathcal{R}(S_{\theta_0,\eta_0})}} f = f^*$, then for any $\varepsilon > 0$, there exists at least one function g^* such that

$$\sum_{j=1}^{J} \mathbb{E}\left[\left(f^*_j(Z_j) - S^{(j)}_{\theta_0,\eta_0} S^*_{\theta_0,\eta_0} g^* \right)^2 \right] < \varepsilon. \tag{5.5}$$

Notice that it is not necessarily the case that f^* is an interior point of $\overline{\mathcal{R}\left(S_{\theta_0,\eta_0} S^*_{\theta_0,\eta_0} \right)}$. However, as the previous display indicates, we can approximate the orthogonal projection by exploiting the operator

[10] Similar conditions to Assumption 5.2.1 in Bonhomme (2012) are sufficient for Assumption 2 (iii).

$S_{\theta_0,\eta_0}S^*_{\theta_0,\eta_0}$, objects that, as we have shown, are given by the model. These types of observations have been used in previous works, e.g., Bonhomme (2012) (see Section D of the Appendix of this paper). In other words, we may say that f^* is approximately smooth relative to $S_{\theta_0,\eta_0}S^*_{\theta_0,\eta_0}$.[11] Moreover, notice that for any function $g \in L^2(Z)$, by orthogonality we have that

$$\sum_{j=1}^{J} \mathbb{E}\left[\left(f_j\left(Z_j\right) - S^{(j)}_{\theta_0,\eta_0}S^*_{\theta_0,\eta_0}g\right)^2\right] = \sum_{j=1}^{J} \mathbb{E}\left[\left(f^*_j\left(Z_j\right) - S^{(j)}_{\theta_0,\eta_0}S^*_{\theta_0,\eta_0}g\right)^2\right] + \sum_{j=1}^{J} \mathbb{E}\left[\left(f_j\left(Z_j\right) - f^*_j\left(Z_j\right)\right)^2\right]. \tag{5.6}$$

The above implies that a necessary and sufficient condition for $S_{\theta_0,\eta_0}S^*_{\theta_0,\eta_0}g$ to be close to f^* is that $S_{\theta_0,\eta_0}S^*_{\theta_0,\eta_0}g$ has to be close to f. This point is important since the left-hand side of (5.6) can be used in estimation, as we will explain below.

In this paper, we focus on a particular function space for functions that satisfy (5.5). To be precise, let $\mathcal{G}$ be some space of functions equipped with norm $||\cdot||_{\mathcal{G}}$ such that $\mathcal{G} \subseteq L^2(Z)$. Notice that this condition is not necessarily restrictive. In particular, if we consider $\mathcal{G}$ to be the space of functions in $L^2(Z)$ that has finite L_1-norm, i.e., $\sum_{j=1}^{J} \mathbb{E}\left[\left|f_j\left(Z_j\right)\right|\right] < \infty$, then, all functions in $L^2(Z)$ belongs to $\mathcal{G}$. In general, we are interested in solving

$$\min_{g\in\mathcal{G}} \sum_{j=1}^{J} \mathbb{E}\left[\left(f_j(Z_j) - S^{(j)}_{\theta_0,\eta_0}S^*_{\theta_0,\eta_0}g\right)^2\right]. \tag{5.7}$$

Using Proposition 1, we can write (5.7) more explicitly:

$$\min_{g\in\mathcal{G}} \sum_{j=1}^{J} \mathbb{E}\left[\left(f_j\left(Z_j\right) - \mathbb{E}\left[\left(\sum_{j'=1}^{J} \mathbb{E}\left[\tilde{\nu}_{j'}\left(Y,\theta_0,\eta_0\right) g_{j'}\left(Z_j\right)\middle| X\right]\right)' \tilde{\nu}_j\left(Y,\theta_0,\eta_0\right)\middle| Z_j\right]\right)^2\right]. \tag{5.8}$$

Typically, there will be more than one element g that solves (5.8). This can be seen from the fact that if g^* is a solution, then any $g^* + h$ is also a solution if $h \in \mathcal{N}\left(S_{\theta_0,\eta_0}S^*_{\theta_0,\eta_0}\right)$, where $\mathcal{N}(\cdot)$ denotes the null space of an operator. Nevertheless, this point is not problematic for us since any of them will lead to the same approximation to f^*, i.e., we can take any function that solves (5.8). A natural thing to do is to focus on the g^* of minimum norm, denote it by g_0. Then $||g_0||_{\mathcal{G}} \leq ||g||_{\mathcal{G}}$ for all g that solves (5.8). If the minimum norm solution exists, it is unique since $\mathcal{N}\left(S_{\theta_0,\eta_0}S^*_{\theta_0,\eta_0}\right)$ is a closed linear subspace (Luenberger, 1997, Them 3.10.1). We will let $||\cdot||_{\mathcal{G}}$ be the ℓ_1-norm.

In what follows we will construct an estimator of g_0 and use it to obtain an estimator of f^*. We propose to estimate g_0 by means of

$$\hat{g}_n = \arg\min_{g\in\mathcal{G}_n} \sum_{j=1}^{J} \mathbb{E}\left[\left(f_j(Z_j) - \hat{S}^{(j)}_{\hat{\theta},\hat{\eta}}\hat{S}^*{}_{\hat{\theta},\hat{\eta}}g\right)^2\right] + 2\lambda_n\, ||g||^2_{\mathcal{G}}\,, \tag{5.9}$$

[11] Interestingly, in other settings, conditions such as this one, have been imposed. Recently, in the context of Riesz representers and for non-parametric IV ill-posed problems, Bennett et al. (2022) impose a similar assumption for the Riesz representer associated with a parameter of interest that can be written as an expectation of a function of the nonparametric component of the model. This assumption assures strong identification of such a parameter, even if the nonparametric part is non-identified.

where $\hat{S}^{(j)}_{\hat{\theta},\hat{\eta}}\hat{S}^{*}{}_{\hat{\theta},\hat{\eta}}$ is a suitable estimator of $S^{(j)}_{\theta_0,\eta_0}S^{*}_{\theta_0,\eta_0}$, $\lambda_n \geq 0$ is a regularization term, and $\mathcal{G}_n$ is a function class. We impose the key assumption that $\mathcal{G}_n \subseteq \bar{\mathcal{G}}$ for some fixed normed set of functions that do not depend on n with norm $||\cdot||_{\mathcal{G}}$ and, importantly, $\bar{\mathcal{G}} \subseteq \mathcal{G}$. We note that $\mathcal{G}_n$ can be any suitable class, e.g., Reproducing Kernel Hilbert Spaces or Neural Networks.

Particularly, we work with the space of sparse functions.[12] Let $\gamma(Z) = \left(\gamma_1(Z_1)', \cdots, \gamma_J(Z_J)'\right)'$ be a vector of basis functions, where $\gamma_j(Z_j)$ is a r_j-dimensional vector of known real-valued functions, with $\mathbb{E}[\gamma_{jk}(Z_j)] = 0$ and $\mathbb{E}\left[\gamma^2_{jk}(Z_j)\right] = 1$, $k = 2, \cdots, r_j$.[13] Let $\beta = (\beta_{11}, \cdots, \beta_{1r_1}, \cdots, \beta_{Jr_J})'$ be a $r-$dimensional vector, where $r = \sum_{j=1}^{J} r_j$. Then, we consider the s-sparse function class in $r-$dimension with bounded coefficients:

$$\mathcal{G}_n = \left\{ g \in \bar{\mathcal{G}} : g_j(Z_j) = \gamma_j(Z_j)'\beta_j, \ \ ||\beta||_0 = s, \ \ ||\beta||_\infty < c \right\}.$$

For estimation, we use cross-fitting.[14] Randomly partition the sample into L subgroups, $I_1, \cdots, I_L$, of the same size. Let I^c_ℓ be the complement of I_ℓ. Next, split I^c_ℓ into three pieces such that $I^c_\ell = A_\ell + B_\ell + C_\ell$. Possibly, this partition will depend on j, but we omit this dependence for simplicity. We compute $\hat{\beta}_\ell$, using observations in I^c_ℓ only, and base estimation of each of the components in (5.9) using each of the pieces of I^c_ℓ. To make our computation clear, we need to add some notation. For any estimator, a subscript will indicate the part of I^c_ℓ that has been used to compute it. For example, $\hat{\theta}_{A_\ell}$ means that θ_0 has been estimated using observations in A_ℓ and so on. In addition, let n_ℓ denote the number of observations in I_ℓ. The estimator $\hat{\beta}_\ell$ can be written as follows

$$\begin{aligned}\hat{\beta}_\ell = \underset{\beta \in \mathbb{R}^r}{\arg\min} \sum_{j=1}^{J} \frac{1}{n - n_\ell} \sum_{i \notin I_\ell} \Bigg(& f_j(Z_{ji}) \\ & - \sum_{j'=1}^{J} \sum_{k=1}^{r_{j'}} \beta_{j'k} \hat{\mathbb{E}}_{C_\ell}\left[\left(\hat{\mathbb{E}}_{B_\ell}\left[\tilde{\nu}_{j'}\left(Y_i, \hat{\theta}_{A_\ell}, \hat{\eta}_{A_\ell}\right) \gamma_{j'k}(Z_{ji}) \middle| X_i \right] \right)' \tilde{\nu}_j\left(Y_i, \hat{\theta}_{B_\ell}, \hat{\eta}_{B_\ell}\right) \middle| Z_{ji} \right] \Bigg)^2 + 2\lambda_n ||\beta||_1 ,\end{aligned} \tag{5.10}$$

where $\hat{\eta}_{A_\ell}$, $\hat{\eta}_{B_\ell}$, $\hat{\mathbb{E}}_{B_\ell}[\cdot|\, X]$, and $\hat{\mathbb{E}}_{C_\ell}[\cdot|\, Z_j]$ are non-parametric estimators, possibly based on some Machine Learning tool, and $\hat{\theta}_{A_\ell}$ and $\hat{\theta}_{B_\ell}$ are possibly non-LR estimators of θ_0.[15] Recall that $\tilde{\nu}_j$ are known functions, given estimators of θ_0 and η_0, thus, these conditional expectations can be evaluated. Notice that the ℓ_1-penalization term allows for $r > n$. We can also write (5.10) using matrix notation. Let $\boldsymbol{f_{j\ell}}$ be a $n_\ell-$dimensional vector containing each $f_j(Z_{ji})$, $i \notin I_\ell$. Let $\hat{\boldsymbol{M}}_{\boldsymbol{j\ell}}$ be a $n_\ell \times r$ design matrix such that its (i,l)-entry is given by

$$\left[\hat{\boldsymbol{M}}_{\boldsymbol{j\ell}}\right]_{il} = \hat{\mathbb{E}}_{C_\ell}\left[\left(\hat{\mathbb{E}}_{B_\ell}\left[\tilde{\nu}_{j'}\left(Y_i, \hat{\theta}_{A_\ell}, \hat{\eta}_{A_\ell}\right) \gamma_{j'k}(Z_{ji}) \middle| X_i \right] \right)' \tilde{\nu}_j\left(Y_i, \hat{\theta}_{B_\ell}, \hat{\eta}_{B_\ell}\right) \middle| Z_{ji} \right]. \tag{5.11}$$

[12] Extending our results to other functional classes such as Reproducing Kernel Hilbert Spaces or Neural Networks is an interesting avenue of research.

[13] We let the first element of γ_j be 1.

[14] Cross-fitting has a long tradition in the semiparametric literature; see, e.g., Bickel (1982), Klaassen (1987), van der Vaart (1998), Robins et al. (2008), Zheng and van der Laan (2010), and more recently, Chernozhukov et al. (2018).

[15] I.e., estimators based on non-LR moments.

Then, (5.10) can be equivalently written as

$$\hat{\beta}_\ell = \arg\min_{\beta \in \mathbb{R}^r} \sum_{j=1}^{J} \frac{1}{n-n_\ell} \left(\boldsymbol{f_{j\ell}} - \boldsymbol{\hat{M}_{j\ell}} \beta \right)' \left(\boldsymbol{f_{j\ell}} - \boldsymbol{\hat{M}_{j\ell}} \beta \right) + 2\lambda_n \left|\left|\beta\right|\right|_1 .$$

Remarkably, we have transformed our initial problem into a Lasso-type one for which there exist well-known and fast algorithms to find the solution.

Ultimately, we propose to estimate each component in κ_0, for each individual $i \in I_\ell$, by means of

$$\begin{aligned} \hat{\kappa}_{j\ell}\left(Z_{ji}\right) &= f_j\left(Z_j\right) - \hat{f}_j^*\left(Z_j\right) \\ &= f_j\left(Z_{ji}\right) - \sum_{j'=1}^{J} \sum_{k=1}^{r_{j'}} \hat{\beta}_{j'k\ell} \hat{\mathbb{E}}_{C_\ell} \left[\left(\hat{\mathbb{E}}_{B_\ell} \left[\tilde{\nu}_{j'} \left(Y_i, \hat{\theta}_{A_\ell}, \hat{\eta}_{A_\ell} \right) \gamma_{j'k}\left(Z_{ji}\right) \middle| X \right] \right)' \tilde{\nu}_j \left(Y_i, \hat{\theta}_{B_\ell}, \hat{\eta}_{B_\ell} \right) \middle| Z_{ji} \right] . \end{aligned} \tag{5.12}$$

We can outline our algorithm as follows:

Algorithm to estimate OR-IVs:
Step 0: Choose a real-valued function $f \in L^2(Z)$. For instance, $f(Z) = (f_1(X), f_2(Z_1)) = (T, Z_1)$ in Example 1. Choose a basis for each $\gamma_j(Z_j)$, e.g., exponential, Fourier, splines, or power. In addition, specify a low-dimensional dictionary, say $\gamma^{low}(Z)$, which is a sub-vector of $\gamma(Z)$.[16]

Step 1: For each $\ell = 1, \cdots L$, compute (possible) non-LR estimators $\hat{\theta}_{A_\ell}$ and $\hat{\theta}_{B_\ell}$. Moreover, using some Machine Learning algorithm, compute $\hat{\eta}_{A_\ell}$, $\hat{\eta}_{B_\ell}$, $\hat{\mathbb{E}}_{B_\ell}\left[\cdot \middle| X\right]$, and $\hat{\mathbb{E}}_{C_\ell}\left[\cdot \middle| Z_j\right]$. These conditional expectations depend on known $\tilde{\nu}_j$, and thus can be evaluated.

Step 2: Compute design matrix $\boldsymbol{\hat{M}_{j\ell}}$ such that its $(i,l)-$entry is (5.11).

Step 3: Initialize $\hat{\beta}_\ell$ using $\gamma^{low}(Z)$ such that

$$\begin{aligned} \left[\boldsymbol{\hat{M}}_{j\ell}\right]_{il} &= \hat{\mathbb{E}}_{C_\ell} \left[\left(\hat{\mathbb{E}}_{B_\ell} \left[\tilde{\nu}_{j'} \left(Y_i, \hat{\theta}_{A_\ell}, \hat{\eta}_{jA_\ell} \right) \gamma_{j'k}^{low} \left(Z_{j'i} \right) \middle| X_i \right] \right)' \tilde{\nu}_j \left(Y_i, \hat{\theta}_{B_\ell}, \hat{\eta}_{jB_\ell} \right) \middle| Z_{ji} \right] , \\ \hat{\beta}_\ell &= \begin{pmatrix} \left(\sum_{j=1}^J \boldsymbol{\hat{M}'_{j\ell} \hat{M}_{j\ell}} \right)^{-1} \left(\sum_{j=1}^J \boldsymbol{\hat{M}'_{j\ell} f_{j\ell}} \right) \\ 0 \end{pmatrix} \end{aligned}$$

Step 4: (While $\hat{\beta}_\ell$ has not converged)

(a) Update normalization

$$\begin{aligned} \hat{D}_{j'k\ell} &= \left[\frac{1}{n-n_\ell} \sum_{i \notin I_\ell} \left\{ \sum_{j=1}^{J} \hat{\mathbb{E}}_{C_\ell} \left[\left(\hat{\mathbb{E}}_{B_\ell} \left[\tilde{\nu}_{j'} \left(Y_i, \hat{\theta}_{A_\ell}, \hat{\eta}_{jA_\ell} \right) \gamma_{j'k} \left(Z_{j'i} \right) \middle| X_i \right] \right)' \tilde{\nu}_j \left(Y_i, \hat{\theta}_{B_\ell}, \hat{\eta}_{jB_\ell} \right) \middle| Z_{ji} \right] \hat{\epsilon}_{ji\ell} \right\}^2 \right]^{1/2} \\ \hat{\epsilon}_{ji\ell} &= f_j\left(Z_{ji}\right) - \sum_{j'=1}^{J} \sum_{k=1}^{r_{j'}} \hat{\beta}_{j'k\ell} \hat{\mathbb{E}}_{C_\ell} \left[\left(\hat{\mathbb{E}}_{B_\ell} \left[\tilde{\nu}_{j'} \left(Y_i, \hat{\theta}_{A_\ell}, \hat{\eta}_{jA_\ell} \right) \gamma_{j'k} \left(Z_{j'i} \right) \middle| X \right] \right)' \tilde{\nu}_j \left(Y_i, \hat{\theta}_{B_\ell}, \hat{\eta}_{jB_\ell} \right) \middle| Z_{ji} \right] . \end{aligned}$$

[16] E.g., take the first $\tilde{r}_j$ components of each γ_j.

(b) Update $\hat{\beta}_\ell$, where

$$\hat{\beta}_\ell = \underset{\beta \in \mathbb{R}^r}{\arg\min} \sum_{j=1}^{J} \frac{1}{n - n_\ell} \left(\boldsymbol{f_{j\ell}} - \boldsymbol{\hat{M}_{j\ell}} \beta \right)' \left(\boldsymbol{f_{j\ell}} - \boldsymbol{\hat{M}_{j\ell}} \beta \right) + 2\lambda_n \sum_{j=1}^{J} \sum_{k=1}^{r_j} \left| \hat{D}_{jk\ell} \beta_{jk} \right|,$$

and

$$\lambda_n = \frac{c_1}{\sqrt{n - n_\ell}} \Phi^{-1} \left(1 - \frac{c_2}{2r} \right),$$

where $\Phi(.)$ is the standard normal cdf.

Step 5: Given the optimal $\hat{\beta}_\ell$, compute $\hat{\kappa}_{j\ell}$ as in (5.12).

Following Belloni et al. (2012), we need to include a normalization term, $\hat{D}_{jk\ell}$, in the ℓ_1 norm, which is necessary for the good properties of the Lasso estimator we are considering. For this same reason, we suggest computing λ_n as given in **Step 4**, as recommended by Belloni et al. (2012) (p. 2380). We initially set $c_1 = 1.1$ and $c_2 = 0.1/\log(n \vee p)$, as recommended by Belloni et al. (2012) (footnote 7). In our Monte Carlo experiments, we have also considered other choices for these constants. To improve numerical stability we follow Chernozhukov et al. (2022d) and cap the maximum number of iterations at 10. In addition, we use warm start. This means that in a given iteration, the initial parameter value is equal to the $\hat{\beta}_\ell$ obtained in the previous iteration. In Section B of the Appendix we provide a justification for the previous optimization of $\hat{\beta}_\ell$. Notice that **Step 4** (b) requires solving for $\hat{\beta}_\ell$. For this, we use an extension of the coordinate descent approach for Lasso (Friedman et al., 2007, 2010; Fu, 1998); see Section C of the Appendix for details and justification.

We make the observation that our previous algorithm applies to the most general case. In some applications, it will be implemented with some simplifications, depending on the particular expression of $\tilde{\nu}_j$ and how the random variables relate. For example, it might not be necessary to partition I_ℓ into three pieces, as it might not be necessary to compute all the estimators in **Step 1**. This occurs, for example, when $\tilde{\nu}_j$ does not depend on θ_0 or η_0 or both. Another case is when $\mathbb{E}\left[\tilde{\nu}_j\left(Y, \theta_0, \eta_0\right) \gamma_{jk}(Z_{ji}) \middle| X\right] = \tilde{\nu}_j\left(Y, \theta_0, \eta_0\right) \gamma_{jk}(Z_{ji})$ as (Y, Z_j) and the variables that η_0 depends on are contained in X. More simplifications can emerge if $\tilde{\nu}_j$ depends only on Z_j. As we will illustrate below, in Example 1 and Example 2, $\left[\boldsymbol{\hat{M}}_{j\ell}\right]_{i\ell}$ has a simpler expression than in (5.11).

We can simplify our implementation further. Notice that by nature of our general model, we need to deal with a dimension given by j, which indexes the CMRs. This might be inconvenient in practice as the researcher has to make choices for each j. For example, the type of basis γ_j and the dimension of each of these vectors r_j. Moreover, this might lead to work with large matrices $\boldsymbol{\hat{M}_{j\ell}}$. In particular, in the case where $r_j = r^*$ for all j, the number of columns of the design matrix is $r = J \times r^*$. Nevertheless, the way in which we have formulated our problem allows us to deal with this point. We are trying to estimate simultaneously a number of functions in $L^2(Z)$ by using a linear combination of elements. Then we can straightforwardly impose restrictions among the approximating terms, which places restrictions on $\mathcal{G}_n$. One natural choice is to restrict γ_j and β_j to be constant across j. Then, each basis $\gamma_j = \tilde{\gamma}$ has the same dimension, and $g_j(Z_j) = \tilde{\gamma}(Z_j)' \beta$. In this case, $\beta = (\beta_1, \cdots, \beta_r)$, and

$$\left[\hat{\boldsymbol{M}}_{\boldsymbol{j\ell}}\right]_{ik} = \hat{\mathbb{E}}_{C_\ell}\left[\left(\sum_{j'=1}^{J} \hat{\mathbb{E}}_{B_\ell}\left[\tilde{\nu}_{j'}\left(Y_i, \hat{\theta}_{A_\ell}, \hat{\eta}_{A_\ell}\right) \tilde{\gamma}_k\left(Z_{j'i}\right) \middle| X_i\right]\right)' \tilde{\nu}_j\left(Y_i, \hat{\theta}_{B_\ell}, \hat{\eta}_{B_\ell}\right) \middle| Z_{ji}\right]. \tag{5.13}$$

Needless to say, our approximating condition will be more likely to hold with more flexible models.[17] As we will evaluate this simpler estimation of OR-IVs in Section 9 through different Monte Carlo exercises, let us show (5.13) in the contexts of Example 1 and Example 2.

Example 1: In this case, the researcher has to provide a function $f \in L^2(Z)$ and basis $\gamma(X)$ and $\gamma(Z_1)$. The regressors can be written as

$$\left[\hat{\boldsymbol{M}}_{\boldsymbol{1\ell}}\right]_{ik} = \frac{\gamma_k(X_i)}{\hat{\eta}_\ell^2(X_i)} + \hat{\mathbb{E}}_{B_\ell}\left[\frac{\gamma_k(Z_{1i})\,\delta_i}{\hat{\eta}_{A_\ell}^3(X_i)} \rho\left(Y_{1i}, Z_{1i}, \hat{\theta}_{A_\ell}\right) \middle| X_i, \right],$$

$$\left[\hat{\boldsymbol{M}}_{\boldsymbol{2\ell}}\right]_{ik} = \hat{\mathbb{E}}_{C_\ell}\left[\frac{\delta_i}{\hat{\eta}_{B_\ell}^2(X_i)} \rho\left(Y_{1i}, Z_{1i}, \hat{\theta}_{B_\ell}\right) \left(\frac{\gamma_k(X_i)}{\hat{\eta}_{B_\ell}(X_i)} + \hat{\mathbb{E}}_{B_\ell}\left[\frac{\gamma_k(Z_{1i})\,\delta_i}{\hat{\eta}_{A_\ell}^2(X_i)} \rho\left(Y_{1i}, Z_{1i}, \hat{\theta}_{A_\ell}\right) \middle| X_i\right]\right) \middle| Z_{1i}\right].$$

In this example, while $\hat{\boldsymbol{M}}_{\boldsymbol{2\ell}}$ involves all the estimated objects in **Step 1**, in $\hat{\boldsymbol{M}}_{\boldsymbol{1\ell}}$, only $\hat{\theta}_{A_\ell}$, $\hat{\eta}_{A_\ell}$, and one conditional expectation appear. Thus, to estimate the entries of $\hat{\boldsymbol{M}}_{\boldsymbol{1\ell}}$, we only need to split I_ℓ^c into two pieces such that $I_\ell^c = A_\ell + B_\ell$, while for estimation of $\hat{\boldsymbol{M}}_{\boldsymbol{2\ell}}$, we make the partition $I_\ell^c = A_\ell + B_\ell + C_\ell$. □

Example 2: The user has to provide $f(Z) = (f_1(Z_1), f_2(Z_1), f_3(Z_2), f_4(Z_2)) \in L^2(Z)$ and basis $\gamma(Z_1)$ and $\gamma(Z_2)$. In this example, the regressors have the following expression

$$\left[\hat{\boldsymbol{M}}_{\boldsymbol{1\ell}}\right]_{ik} = \gamma_k(Z_{1i}) + \hat{\theta}_{\omega\ell}\gamma_k(Z_{1i}), \tag{5.14}$$

$$\left[\hat{\boldsymbol{M}}_{\boldsymbol{2\ell}}\right]_{ik} = \hat{\theta}_{\omega\ell}\left(\gamma_k(Z_{1i}) + \hat{\theta}_{\omega\ell}\gamma_k(Z_{1i})\right), \tag{5.15}$$

$$\left[\hat{\boldsymbol{M}}_{\boldsymbol{3\ell}}\right]_{ik} = \gamma_k(Z_{2i}) + \hat{\theta}_{\omega\ell}\gamma_k(Z_{2i}), \tag{5.16}$$

$$\left[\hat{\boldsymbol{M}}_{\boldsymbol{4\ell}}\right]_{ik} = \hat{\theta}_{\omega\ell}\left(\gamma_k(Z_{2i}) + \hat{\theta}_{\omega\ell}\gamma_k(Z_{2i})\right). \tag{5.17}$$

As we can see from above, no conditional expectations or estimators of η_0 appear. Hence, it is not necessary to partition I_ℓ^c. This is, we would use all observations in I_ℓ^c to obtain an estimator of $\theta_{0\omega}$. □

6 Asymptotic Properties of OR-IVs

This section provides the mean square convergence rate for $\hat{\kappa}$ based on the Lasso estimator that we introduced above, which is fundamental for deriving the asymptotic properties of $\hat{\theta}$.

[17] Notice that our algorithm remains as given above, except for the fact that now the expression $\left[\hat{\boldsymbol{M}}_{\boldsymbol{j\ell}}\right]_{ik}$ (and thus the normalization term $\hat{D}_{jk\ell}$) needs to account for the aggregation before the expectation $\hat{\mathbb{E}}_{B_\ell}[\cdot|\, X]$, as suggested by (5.13).

Let $\boldsymbol{M}_j$ be the population analog of matrix $\hat{\boldsymbol{M}}_{\boldsymbol{j\ell}}$. Let $\hat{M}_{j\ell}\left(Z_{ji}\right)$ be a $r-$dimensional vector containing the $i-$ row of $\hat{\boldsymbol{M}}_{\boldsymbol{j\ell}}$. A similar definition applies to $M_j\left(Z_{ji}\right)$. We define

$$\hat{F}_{j\ell}=\frac{1}{n-n_\ell}\sum_{i\notin I_\ell}f_j\left(Z_{ji}\right)\hat{M}_{j\ell}\left(Z_{ji}\right),\qquad F_j=\mathbb{E}\left[f_j\left(Z_j\right)M_j\left(Z_j\right)\right],$$
$$\hat{G}_{j\ell}=\frac{1}{n-n_\ell}\sum_{i\notin I_\ell}\hat{M}_{j\ell}\left(Z_{ji}\right)\hat{M}_{j\ell}\left(Z_{ji}\right)^{'},\quad G_j=\mathbb{E}\left[M_j\left(Z_j\right)M_j\left(Z_j\right)^{'}\right].$$

Then, $\hat{\beta}_\ell$ can equivalently be written as

$$\hat{\beta}_\ell=\underset{\beta\in\mathbb{R}^r}{\arg\min}\sum_{j=1}^{J}\left(-2\hat{F}_{j\ell}^{'}\beta-\beta^{'}\hat{G}_{j\ell}\beta\right)+2\lambda_n\left|\left|\beta\right|\right|_1. \tag{6.1}$$

Interestingly, the previous characterization of the program is similar to the one proposed by Chernozhukov et al. (2022d) for automatic estimation of Riesz representers (see Equation (3.7) in this paper), but with different matrices. An important difference is that $\hat{G}_j$, the estimated Gram matrix, depends on regressors $\hat{M}_j$'s that are estimated. This also contrasts with the formulation considered by Bakhitov (2022). These previous works formulate the estimation problem in terms of known objects analogous to M_j's. In what follows, we work with the characterization given in (6.1). We start by assuming

Assumption 3. *There are constants $c_1,\cdots,c_J$ such that with probability approaching one*

$$\max_{1\leq k\leq r}\left|M_{jk}\left(Z_j\right)\right|\leq c_j,\quad \mu_j-a.s.,\quad j=1,\cdots,J.$$

Since $\hat{G}_j$ depends on the estimated $\hat{M}_j(Z_j)\hat{M}_j(Z_j)^{'}$, it is natural that we require a convergence rate for them such that we can assure that $\hat{G}_j$ provides a good approximation to G_j. Let F_0 be the distribution of the data W, then we assume

Assumption 4.

$$\int\left|\left|\hat{M}_{j\ell}(z_{ji})\hat{M}_{j\ell}(z_{ji})^{'}-M_{j\ell}(z_{ji})M_{j\ell}(z_{ji})^{'}\right|\right|_{\infty}F_0\left(dw\right)=O_p\left(\varepsilon_n^2\right),\quad \varepsilon_n=\sqrt{\frac{\log(r)}{n}}.$$

Assumption 4 is a key assumption. For our Lasso results to be valid, it is sufficient to approximate the population counterparts of the entries of the Gram matrix well, at a rate given by ε_n^2. Adding more estimated regressors in the Lasso program is worth it as long as their estimation is good enough. This emphasizes the need to leverage machine learning tools to effectively approximate the unknown regressors, and thereby G_j, at the required rate. The flexibility of these algorithms can be important in ensuring that Assumption 4 holds. Assumptions 3 and 4 imply

$$\left|\left|\hat{G}_{j\ell}-G_j\right|\right|_{\infty}=O_p\left(\varepsilon_n\right).$$

Chernozhukov et al. (2022d) and Bakhitov (2022) present a similar result to the previous one. It should not be surprising that we obtain this, even though we do not know the M_j's, as Assumption 4 addresses the fact that we are working with estimated $\hat{M}_j$'s in the sense that we can proceed "as if" we knew them. Next, we impose a sparse approximate condition on the orthogonal projection f^*.

Assumption 5. *There exist $C > 1$ and $\bar{\beta}$ with s non-zero elements such that*

$$\sum_{j=1}^{J} \mathbb{E}\left[\left\{f_j^*\left(Z_j\right) - M_j(Z_j)'\bar{\beta}\right\}^2\right] \leq Cs\varepsilon_n^2,$$

with $\left|\left|\bar{\beta}\right|\right|_1 = O(1)$.

The previous assumption controls the squared approximation error from using the linear combination $M_j'\bar{\beta}$ to approximate the orthogonal projection. Remark that Assumption 5 does not impose that f^* can be written as a linear combination of s terms, i.e., that the orthogonal projection is strictly sparse. Instead, Assumption 5 only requires the existence of $\bar{\beta}$ with s terms such that the approximation error across the J elements is bounded by $Cs\varepsilon_n^2$. Moreover, the above assumption does allow the unknown identity of the elements in M_j that give a good approximation, i.e., the researcher does not have to specify which elements are important, a task that will be typically hard to accomplish as we are dealing with highly complex functional objects; see Bradic et al. (2022). We also notice that a very sparse approximation, with a small number of terms s, will typically lead to faster convergence rates. For a more detailed discussion of approximation bias conditions with sparse specifications, see Belloni et al. (2012). For the remainder of this section let us drop the dependence of random elements on ℓ to simplify our notation.

We next impose a sparse eigenvalue condition, following the Lasso literature (e.g., Bickel et al., 2009), on the empirical matrix $\sum_{j=1}^{J}\hat{G}_j$:[18]

Assumption 6. *The largest eigenvalue of $\sum_{j=1}^{J} G_j$ is uniformly bounded in n and there is a $c > 0$ such that with probability approaching one*

$$\phi^2(s) = \inf\left\{\frac{\delta'\sum_j^J \hat{G}_j\delta}{\left|\left|\delta_{S_\beta}\right|\right|_2^2}, \quad \delta \in \mathbb{R}^r \setminus \{0\}, \left|\left|\delta_{S_\beta^c}\right|\right|_1 \leq 3\left|\left|\delta_{S_\beta}\right|\right|_1, \ \ |S_\beta| \leq s\right\} > c.$$

Notice that the objective function in (6.1) depends on a sample counterpart of F_j, $\hat{F}_j$, and thus we hypothesize a convergence rate for it.

Assumption 7. $\left|\left|\hat{F}_{j\ell} - F_j\right|\right|_\infty = O_p\left(\varepsilon_n\right)$.

Assumption 4 involves the estimated components of the estimated design matrix $\hat{G}_j$, $\hat{M}_j\hat{M}_j'$. This assumption is crucial for the good properties of the Lasso estimator $\hat{\beta}$. Nonetheless, it is not enough to assure that a good approximation will be obtained for f^*. If $\hat{\beta}$ is sufficiently accurate but $\hat{M}_j$ is not, in a precise sense, we will not be able to approximate f^* well. Given this, we also need to impose some rate for the mean square distance between $\hat{M}_j$ and its population counterparts M_j. We do this with the following assumption:

[18] Imposing a sparse eigenvalue condition on the empirical Gram matrix has been done elsewhere, e.g., Belloni and Chernozhukov (2013).

Assumption 8. *Let*

$$B = \sum_{j=1}^{J} \int \left(M_j\left(z_j\right) - \hat{M}_j\left(z_j\right) \right) \left(M_j\left(z_j\right) - \hat{M}_j\left(z_j\right) \right)' F_0\left(dw\right).$$

Then, the maximum eigenvalue of B is $O_p\left(\varepsilon_n^2\right)$.

Finally, we allow the Lasso regularization parameter λ_n to shrink slightly slower than $\varepsilon_n = \sqrt{\log(r)/n}$ (as in Bakhitov (2022) and Chernozhukov et al. (2022d)), which restricts the rate at which r grows as a function of n. This completes the list of sufficient conditions that yields one of the main results of the paper:

Theorem 2. *Let Assumptions 3-8 hold. In addition, suppose that $\varepsilon_n = o\left(\lambda_n\right)$. Then,*

$$||\hat{\kappa}(Z) - \kappa_0(Z)||_{L^2(Z)} = O_p\left(\mu_n^{\kappa}\right), \quad \mu_n^{\kappa} = \sqrt{s}\lambda_n.$$

Notice that the rate depends on s, which controls the degree of approximate sparsity in f^*. The smaller s, the faster the rate. Additionally, the rate depends on ε_n, which controls the approximation of the estimated Gram matrix $\hat{G}_j$ to G_j. As we emphasized above, this depends heavily on the good properties of $\hat{M}_j$. Interestingly, despite allowing for an endogenous setting, as in Bakhitov (2022), which derives slower rates for its Riesz representer estimator relative to the one in Chernozhukov et al. (2022d), endogeneity does not affect the rate we obtain.[19] Intuitively, this is due to the fact that to construct our estimator $\hat{\kappa}$, we always work with functions or dictionaries of the exogenous variables of the model. Particularly, g^* is a function of the conditioning variables only. In contrast, Bakhitov (2022) deals with dictionaries of endogenous variables. We regard this as an advantage of the way we have formulated our problem.

7 Estimation of the Parameter of Interest in a Two-Step Setting

We now propose an estimator of θ_0. To this end, we will simplify some aspects of our general model (5.1). Up to now, we have allowed the functions m_j's to depend on the entire vector η_0 and θ_0. This is the most general case that we can think of in our context. As we have shown, our construction of debiased moments can handle it. We, however, will be interested in the common situation where the researcher works with a two-step setting, in which there are functions m_j's that depend on η_0 only. Then, CMRs based on those functions can be used to obtain an estimator of η_0, as it occurs with our examples. We focus on this case as many relevant scenarios in applied work present this feature (see, e.g., Chen and Qiu, 2016, Section 5 and references therein). Additionally, to simplify our theoretical derivations below and to be able to obtain more familiar conditions for rates of first-stage estimators, we will focus on the case where m_j depends on η_j only and η_{0j} is a conditional expectation.[20] Focusing

[19] See Theorem 1 and the author's comments in Bakhitov (2022).

[20] We remark that this is not necessary. Our theory can be extended to the more general case, at the cost of making derivations more involved. We focus on the situation where η_{0j} is a conditional expectation since the asymptotic result we obtain for $\hat{\theta}$ depends on the mean square convergence of $\hat{\eta}_j$, which has been derived for various machine learners. This is not the case for general first-stages, except for the theoretical guarantees obtained by Farrell et al. (2021b) for Deep Neural Networks.

on conditional expectations as first-stage nuisance parameters is appealing as we can leverage machine learning algorithms to model such objects very flexibly. Furthermore, there exist well-known theoretical results that guarantee their convergence.

Recall from our previous discussion that for a given instrument $f \in L^2(Z)$, we can obtain an OR-IV $\kappa_0(Z)$, with J elements. Then, for different choices of instruments, say q of them, we can construct J vectors $\boldsymbol{\kappa_{0j}(Z_j)}$, of dimension q. We use the bold notation to emphasize that $\boldsymbol{\kappa_{0j}(Z_j)}$ is a $q-$vector. For the remainder of this paper, let us re-define (5.3) such that

$$\psi\left(W, \theta, \eta, \boldsymbol{\kappa}\right) = \sum_{j=1}^{J} m_j\left(Y_i, \theta, \eta_j\right) \boldsymbol{\kappa_j(Z_j)},$$

where we should notice that ψ is a now a $q-$vector function, and m_j depends on η_j only. Let $\hat{\eta}_\ell$ be an estimator of η_0, using observations in I_ℓ^c. In addition, let

$$\hat{\psi}\left(\theta\right) = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \psi\left(W_i, \theta, \hat{\eta}_\ell, \hat{\boldsymbol{\kappa}}_\ell\right).$$

Our proposed estimator $\hat{\theta}$ is defined as the solution to the GMM program

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \hat{\psi}\left(\theta\right)' \hat{\Lambda} \hat{\psi}\left(\theta\right), \tag{7.1}$$

where $\hat{\Lambda}$ is a positive semi-definite symmetric weighting matrix. A choice that asymptotically minimizes the asymptotic variance is $\hat{\Lambda} = \hat{\Psi}^{-1}$, where

$$\hat{\Psi} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \hat{\psi}_{i\ell} \hat{\psi}_{i\ell}', \quad \hat{\psi}_{i\ell} \equiv \psi\left(W_i, \tilde{\theta}_\ell, \hat{\eta}_\ell, \hat{\boldsymbol{\kappa}}_\ell\right),$$

and $\tilde{\theta}_\ell$ is a possibly non-LR estimator of θ_0, based on observations in I_ℓ^c. Matrix $\hat{\Psi}$ directly accounts for estimating η_0 and κ_0 in a previous stage. We refer to (7.1) as the *Debiased-CMRs Estimator* (D-CMRs). Let us summarize our estimation procedure with the following steps:

Step 1: For each subsample $\ell = 1, \cdots, L$, compute estimates $\hat{\eta}_\ell$ and $\hat{\kappa}_\ell$, using observations not in I_ℓ.
Step 2: Obtain our estimator $\hat{\theta}$ by means of (7.1). The estimator of the asymptotic variance, which accounts for the estimation of the previous objects, takes the "sandwich" form

$$\hat{V} = \left(\hat{\Upsilon}' \hat{\Lambda} \hat{\Upsilon}\right)^{-1} \hat{\Upsilon}' \hat{\Lambda} \hat{\Psi} \hat{\Lambda} \hat{\Upsilon} \left(\hat{\Upsilon}' \hat{\Lambda} \hat{\Upsilon}\right)^{-1}, \quad \hat{\Upsilon} = \frac{\partial}{\partial \theta} \hat{\psi}(\hat{\theta}). \tag{7.2}$$

As η_0 is a vector of conditional expectations, any suitable machine learning procedure can be used to obtain $\hat{\eta}_\ell$. Standard assumptions in the machine learning literature need to be imposed on the mean-square convergence rate of this estimator. Conveniently, for the exogenous case where η_0 depends on the conditioning variables Z only, the convergence rates for different machine learners have been already obtained, for example, neural networks (Chen and White, 1999; Farrell et al., 2021a;

Schmidt-Hieber, 2020), random forests (Syrgkanis and Zampetakis, 2020), Lasso (Bickel et al., 2009), and boosting (Luo et al., 2022). As we stated previously, we want to consider the general case where η_0 can also be a function of variables different from the conditioning ones, i.e., we allow for endogeneity. In the endogenous setting, identification of η_0 is trickier and typically involves an ill-posed problem. In particular, the identification of this parameter requires the so-called "completeness condition" (Newey and Powell, 2003). The ill-posedness results in slower rates for the mean square norm of estimators of η_0 than in the exogenous situation. One strategy to alleviate this is to consider rates for the projected mean square norm. Let $T_j : L^2(X) \mapsto L^2(Z_j)$ denote the conditional expectation operator given by

$$T_j \eta_j = \mathbb{E}\left[\left.\eta_j(X)\right| Z_j\right].$$

The projected norm is $||T_j(\eta_j - \eta_{0j})||_2 = \sqrt{\mathbb{E}\left[\mathbb{E}\left[\left.\eta_j(X) - \eta_{0j}(X)\right| Z_j\right]^2\right]}$. That this norm projects onto the exogenous Z_j allows us to hypothesize convergence rates without having to control the degree of ill-posedness. Moreover, due to ill-posedness, rates for the square norm will be typically slower than for the projected norm.[21] Hence, conditions on rates for the projected norm will be weaker. This is what we will do to derive the asymptotic properties of $\hat{\theta}$. To be concrete, we will require

$$\begin{aligned} ||T(\eta - \eta_0)||_{L^2(Z)} &\equiv \sqrt{\sum_{j=1}^{J} ||T_j(\eta_j - \eta_{0j})||_2^2} \\ &= O_p(\mu_n^{\eta}), \end{aligned} \tag{7.3}$$

where we will allow μ_n^{η} to be slower than the $\sqrt{n}$−rate. Then, we can use a variety of machine learners that deal with the estimation of conditional expectations under the presence of endogeneity (e.g., Gold et al., 2020; Singh et al., 2019). Considering the previous discussion, and to avoid ambiguity, for the remainder of this paper, we treat η as belonging to $\tilde{\boldsymbol{B}} \subseteq \boldsymbol{B}$, equipped with the projected mean square norm, given on the right-hand side of (7.3).

8 Asymptotic Properties of D-CMRs

The theoretical guarantees of $\hat{\theta}$ will be obtained by applying many of the results derived by Chernozhukov et al. (2022a). To this end, let us impose some regularity conditions:

Assumption 9. $\mathbb{E}\left[||\psi(W, \theta_0, \eta_0, \boldsymbol{\kappa_0})||^2\right] < \infty$, *and*

i) $\int |m_j(y, \theta_0, \hat{\eta}_{j\ell}) - m_j(y, \theta_0, \eta_{0j})|^2 F_0(dw) \xrightarrow{p} 0$,

ii) $\int |m_j(y, \theta_0, \hat{\eta}_{j\ell}) - m_j(y, \theta_0, \eta_{0j})|^2 ||\boldsymbol{\kappa_{0j}(z_j)}||^2 F_0(dw) \xrightarrow{p} 0$,

iii) $\int |m_j(y, \theta_0, \eta_{0j})|^2 ||\boldsymbol{\hat{\kappa}_{j\ell}(z_j)} - \boldsymbol{\kappa_{0j}(z_j)}||^2 \xrightarrow{p} 0$.

[21] See Section 2.1 in Bakhitov (2022) to see why this is the case; see also Bennett et al. (2022).

Assumption 9 *i)* and *ii)* are mean square convergence conditions for $\hat{\eta}$, while *iii)* is a convergence condition for $\hat{\boldsymbol{\kappa}}$. Next, let us define

$$\hat{\Delta}_\ell(w) = \sum_{j=1}^{J} \left(m_j\left(y, \theta_0, \hat{\eta}_{j\ell}\right) - m_j\left(y, \theta_0, \eta_{0j}\right)\right)\left(\boldsymbol{\hat{\kappa}_{j\ell}(Z_j)} - \boldsymbol{\kappa_{0j}(Z_j)}\right).$$

We need to guarantee that the estimators of OR-IVs are well-defined, for this, we impose

Assumption 10. *There are constants $c_1, \cdots, c_j$ such that with probability approaching one*

$$\max_{1 \leq k \leq r} \left|\hat{M}_{jk}\left(Z_j\right)\right| \leq c_j, \quad j = 1, \cdots, J, \quad a.s.$$

Also, we hypothesize a rate of convergence for the estimator $\hat{\eta}$, which is required to be faster than $n^{-1/4}$ only, and the same requirement applies for our estimators of OR-IVs, as the following assumption states:

Assumption 11. *i)* $||T\left(\hat{\eta}_\ell - \eta_0\right)||_{L^2(Z)} = O_p\left(\mu_n^\eta\right), \quad \mu_n^\eta = o\left(n^{-1/4}\right)$*; ii)* $\sqrt{n}\mu_n^\eta\mu_n^\kappa \to 0$.

However, to make the rate for $\hat{\eta}$ useful in our context, it must be the case that m_j is continuous in η_j, which is imposed by the following condition:

Assumption 12. *For* $||T\left(\hat{\eta}_\ell - \eta_0\right)||_{L^2(Z)}^2$ *small enough,*

$$\sum_{j=1}^{J} ||T_j\left(m_j\left(y, \theta_0, \eta_j\right) - m_j\left(y, \theta_0, \eta_{0j}\right)\right)||_2^2 \leq C\, ||T\left(\hat{\eta}_\ell - \eta_0\right)||_{L^2(Z)}^2.$$

Assumptions 3, 4, 7, 10, 11, 12, and $\varepsilon_n = o(\lambda_n)$ imply

$$i) \int \left|\left|\hat{\Delta}_\ell(w)\right|\right|^2 F_0(dw) \xrightarrow{p} 0, \quad and \quad ii) \sqrt{n}\int \hat{\Delta}_\ell(w) F_0(dw) \xrightarrow{p} 0. \tag{8.1}$$

Expression (8.1) is a $\sqrt{n}-$convergence result for object $\hat{\Delta}_\ell$, which will be key to derive the asymptotic properties of $\hat{\theta}$. Additionally, let

$$\overline{\psi}\left(\theta, \eta, \boldsymbol{\kappa}\right) = \mathbb{E}\left[\psi\left(W, \theta, \eta, \boldsymbol{\kappa}\right)\right].$$

Assumption 13. $\overline{\psi}\left(\theta, \eta, \boldsymbol{\kappa}\right)$ *is twice continuously Fréchet differentiable in a neighborhood of* η_0.

Then it can be shown that since ψ leads to a debiased moment, there exists a $C > 0$ such that

$$\left|\left|\overline{\psi}\left(\theta_0, \eta, \boldsymbol{\kappa_0}\right)\right|\right| \leq C\, ||T\left(\hat{\eta}_\ell - \eta_0\right)||_{L^2(Z)}^2.$$

All the previous conditions yield the most important result of this section:

Lemma 3. *Let Assumptions 3, 4, 7, 9, 10, 11, 12, and 13 hold. Then,*

$$\sqrt{n}\hat{\psi}(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi\left(W_i, \theta_0, \eta_0, \boldsymbol{\kappa_0}\right) + o_p(1). \tag{8.2}$$

The result in (8.2) is essential for obtaining asymptotic normality of $\hat{\theta}$. Interestingly, cross-fitting enables to show (8.2) in a simple manner, without the need to impose the so-called Donsker conditions for η_0, as discussed in Chernozhukov et al. (2018) and Chernozhukov et al. (2022a). Avoiding Donsker conditions is important as it is unknown if machine learners satisfy them. Furthermore, they could be restrictive in high-dimensional contexts.

The next condition is sufficient to show the consistency of $\hat{\Psi}$, which appears in the asymptotic variance of our GMM estimator.

Assumption 14. $\int \left| m_j\left(y, \tilde{\theta}_\ell, \hat{\eta}_{j\ell}\right) - m_j\left(y, \theta_0, \hat{\eta}_{j\ell}\right)\right|^2 ||\hat{\boldsymbol{\kappa}}_{\boldsymbol{j\ell}}(\boldsymbol{z_j})||^2 F_0(dw) \xrightarrow{p} 0.$

Finally, as in any GMM setting, we need conditions for convergence of the Jacobian: $\frac{\partial}{\partial\theta}\hat{\psi}(\bar{\theta}) \xrightarrow{p} \Upsilon = \mathbb{E}\left[\frac{\partial}{\partial\theta}\psi\left(W, \theta_0, \eta_0, \boldsymbol{\kappa_0}\right)\right]$ for any $\bar{\theta} \xrightarrow{p} \theta_0$. To that end, we impose the following:

Assumption 15. *Υ exists and there is a neighborhood $\mathcal{N}$ of θ_0 and $||\cdot||$ such that*

i) $||T\left(\hat{\eta}_\ell - \eta_0\right)||_{L^2(Z)} ||\hat{\kappa}_\ell - \kappa_0||_{L^2(Z)} \xrightarrow{p} 0$*;*

ii) For all $||T\left(\eta - \eta_0\right)||_{L^2(Z)} ||\kappa - \kappa_0||_{L^2(Z)}$ (where we are considering each element of $\boldsymbol{\kappa}_j$) small enough, $\psi\left(W, \theta, \eta, \boldsymbol{\kappa}\right)$ is differentiable in θ on $\mathcal{N}$ with probability approaching one and there is a C and $d\left(W, \eta, \boldsymbol{\kappa}\right)$ such that for $\theta \in \mathcal{N}$ and for each $||T\left(\eta - \eta_0\right)||_{L^2(Z)} ||\kappa - \kappa_0||_{L^2(Z)}$ small enough

$$\left|\left|\frac{\partial\psi\left(W, \theta, \eta, \boldsymbol{\kappa}\right)}{\partial\theta} - \frac{\partial\psi\left(W, \theta_0, \eta, \boldsymbol{\kappa}\right)}{\partial\theta}\right|\right| \leq d\left(W, \eta, \boldsymbol{\kappa}\right) ||\theta - \theta_0||^{1/C}; \quad \mathbb{E}\left[d\left(W, \eta, \kappa\right)\right] < C;$$

iii) For each q and k, $\int \left|\frac{\partial\psi_q(w, \theta_0, \hat{\eta}_\ell, \hat{\boldsymbol{\kappa}}_{\boldsymbol{\ell}})}{\partial\theta_k} - \frac{\partial\psi_q(w, \theta_0, \eta_0, \boldsymbol{\kappa_0})}{\partial\theta_k}\right| F_0(dw) \xrightarrow{p} 0.$

Given the previous assumptions and findings, the following result, which shows the asymptotic normality of $\hat{\theta}$, can be obtained in relatively simple terms.

Theorem 4. *Let Assumptions 3, 4, 7, 9, 10, 11, 12, 13, and 14 hold. In addition, let $\hat{\theta} \xrightarrow{p} \theta_0$, $\hat{\Lambda} \xrightarrow{p} \Lambda$, and $\Upsilon^{'}\Lambda\Upsilon$ be non-singular. Then,*

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, V\right), \quad V = \left(\Upsilon^{'}\Lambda\Upsilon\right)^{-1}\Upsilon^{'}\Lambda\Psi\Lambda\Upsilon\left(\Upsilon^{'}\Lambda\Upsilon\right)^{-1}.$$

If Assumption 15 also holds, then $\hat{V} \xrightarrow{p} V$.

Theorem 4 implies that confidence intervals for $\hat{\theta}$ can be obtained straightforwardly in a standard way, using $\hat{V}$, and the usual quantiles of the standard normal distribution. This holds despite the convergence rates of nuisance parameters being slower than $\sqrt{n}$. Note that Theorem 4 relies on the consistency of $\hat{\theta}$. We provide sufficient conditions for this in Section E.

9 Monte Carlo

This section studies the performance of D-CMRs, introduced in Section 7, in samples of finite sizes. We have run several Monte Carlo experiments in the context of Example 2 to estimate the parameters of a production function, θ_{0p} (cf. Equation 3.4), and the parameter associated with the productivity process, $\theta_{0\omega}$ (cf. Equation 3.8), simulating data of a panel of n firms observed across T periods.

9.1 Data Generating Process

The data generating process (DGP) we work with is similar to the one considered by Ackerberg et al. (2014), Section 4.3. Only for this section, let us distinguish between variables in logs and in levels. Uppercase variables denote variables in logs while lowercase variables stand for variables in levels. In our experiments, firms are followed during three periods, i.e., $T = 3$. We consider a Cobb-Douglass production function in logs:

$$Y_{it} = \theta_{01} + \theta_{0k}K_{it} + \omega_{it} + \epsilon_{it},$$

where $\theta_{01} = 0$ and $\theta_{0k} = 1$. The law of motion of capital is given by

$$k_{it} = (1 - \delta)\, k_{i,t-1} + \mu_{it} i_{i,t-1},$$

where $1 - \delta = 0.9$, μ_{it} is a lognormal standard shock to the capital accumulation process, and i_{it} is the firm's investment decision. This decision is assumed to follow

$$I_{it} = \gamma_0 + \gamma_1 K_{it} + \gamma_2 \omega_{it} + \exp\left(-0.5K_{it} + 0.5\omega_{it}\right),$$

where $\gamma_0 = 0$, $\gamma_1 = -0.7$, and $\gamma_2 = 5$. We consider a large value for γ_2 to exacerbate the endogeneity bias due to the correlation between inputs and the anticipated productivity shock. We have specified an admittedly ad-hoc process for I_{it}. We do this for two reasons. First, we avoid solving potentially complicated firms' dynamic programs. Second, such a process, as we showed, is the source of the non-parametric component of this model, and thus we would like to model it flexibly to assess how our estimator behaves in highly nonlinear circumstances. Hence, we would like to have some degrees of freedom to make the relationship between I_{it} and ω_{it} complex.

Productivity is assumed to follow a normal AR(1) process with $\theta_{0\omega} = 0.7$. The variance of the innovation term in this process is specified such that the standard deviation of ω_{it} is $\sigma_\omega = 0.1$. The unanticipated productivity or measurement error in output is normal and iid over firms and time. The standard deviation of this shock σ_ϵ varies across the three periods of observed data such that $\sigma_{\epsilon_1} = 0.2$, $\sigma_{\epsilon_2} = 0.05$, and $\sigma_{\epsilon_3} = 0.1$. Finally, to avoid our GDP depending on the starting values of the variables, we focus on data coming from the steady-state distribution implied by the model. For this, we have simulated the data over one hundred periods and kept the last three.[22]

9.2 Results

Recall that **Step 0** in our algorithm to estimate OR-IVs involves choosing a vector of functions of the conditioning variables appearing in the CMRs implied by the model, (3.10)-(3.13). A given vector function $f(Z)$ will allow us to obtain one unconditional LR moment. In our simulations, we automatically construct four debiased moments, and thus we have to provide four such vectors of functions.

[22] The same idea has been followed by Ackerberg et al. (2014) and Ackerberg et al. (2015).

These are

$$\begin{aligned} f_1(Z) &= \left(K_{i1}, K_{i1}, K_{i2}, K_{i2}\right)', \\ f_2(Z) &= \left(I_{i1}, I_{i1}, I_{i2}, I_{i2}\right)', \\ f_3(Z) &= \left(K_{i1}, K_{i1}, I_{i2}, I_{i2}\right)', \\ f_4(Z) &= \left(K_{i1}, I_{i1}, I_{i2}, I_{i2}\right)'. \end{aligned}$$

Let us emphasize that each of the previous vectors has four elements since our model has been written in terms of four CMRs, (3.10)-(3.13). We have specified four of these vectors as we aim to construct four different orthogonal moments.

Based on these, we have run GMM using the identity matrix as the weighting matrix $\hat{\Lambda}$.[23] In all situations, the bases coincide, i.e., $\gamma_j = \tilde{\gamma}$, and β_j's are assumed to be constant across j, for simplicity. The estimation of the OR-IVs is based on regressors (5.14)-(5.17) and we obtain $\hat{\beta}_\ell$ using our algorithm outlined in Section 5.2.

We acknowledge that one limitation our overall procedure has is that it involves several choices by the user. Hence, we are interested in studying the performance of our proposed algorithm using different choices, with the hope that these do not play an important role, as our theoretical results indicate, at least for a reasonable sample size. These choices are λ_n, the number of folds L, the number of bases r, the type of basis $\tilde{\gamma}$, and the machine learner employed in the first stage.

Except for specific situations, we estimate η_0 with boosting letting $L = 4$, γ's are exponential bases, $r = 9$.[24] The tuning parameter λ_n is the recommendation by Belloni, Chen, Chernozhukov, and Hansen (2012) (p. 2380, BCCH below). This is such that $\lambda_n = \frac{1.1}{\sqrt{n-n_\ell}}\Phi^{-1}\left(1-\frac{c_2}{2r}\right)$, with $c_2 = 0.1/\log((n-n_\ell)\vee r)$. Also, we have chosen a smaller and larger λ_n. The smaller λ_n is such that $\lambda_n = \frac{1.01}{\sqrt{n-n_\ell}}\Phi^{-1}\left(1-\frac{c_2}{2r}\right)$, with $c_2 = 2/\log(\log(\log((n-n_\ell)\vee r)))$. The case with larger λ_n has $\lambda_n = \frac{1.3}{\sqrt{n-n_\ell}}\Phi^{-1}\left(1-\frac{c_2}{2r}\right)$, with $c_2 = 0.1/\log((n-n_\ell)\vee r)$. We also consider a scenario where $L = 6$. In a different experiment, we specify a larger number of coefficients such that $r = 25$. Additionally, we model γ's through Fourier basis.[25] Finally, in another situation, η_0 is estimated with Random Forest.[26] Taking everything together, we present seven different experiments. In a given experiment, three different sample sizes are considered. We based results on $1,000$ Monte Carlo repetitions.

Table 1 and 2 display the associated bias and the 95% coverage of D-CMRs for each of the three

[23] This can in principle be improved by considering some matrix that depends on the instruments Z. However, we have kept our choice for simplicity and to reduce computational time.

[24] We use the R-function `gbmt` for boosting. Default options were kept. In particular, weights are equal to 1, the model offset is a vector of zeros, the number of trees is 2000, interaction depth is 3, the minimum number of observations in a node is 10, shrinkage is set at 0.001, bag fraction is 0.5, and 50% of the sample is for training the model. Predictions were conducted using the first 500 iterations of the boosting sequence. A one-dimensional exponential basis is of the form $\tilde{\gamma}_k(V) = \exp(\alpha_k V)$, each with different rate parameter α_k. To create a multi-dimensional basis we use the tensor product among one-dimensional bases. For a discussion of the use of bases with R, we refer the reader to Ramsay et al. (2009).

[25] A one-dimensional Fourier basis is such that $\tilde{\gamma}_1(V) = 1$, $\tilde{\gamma}_2(V) = \sin(\alpha V)$, $\tilde{\gamma}_3(V) = \cos(\alpha V)$, $\tilde{\gamma}_4(V) = \sin(2\alpha V)$, $\tilde{\gamma}_5(V) = \cos(2\alpha V)$, $\cdots$, where $\alpha = 2\pi/K$ and K is the range of V.

[26] We use the R-function `ranger`. Default options were kept. In particular, the number of trees is 500, the minimal node size is 5, the minimal terminal node size is 1, sampling is with replacement, the splitting rule is based on variance, the significance threshold is 0.5, and the regularization factor is 1.

parameters of interest, θ_{01}, θ_{0k}, and $\theta_{0\omega}$, across the Monte Carlo repetitions. The tables indicate reasonably good performance for D-CMRs uniformly, regardless of the specific experiment we focus on. We observe that as the sample size increases, D-CMRs reports small bias, which is true for every parameter of the model. Hence, point estimates provide a good approximation to the true values. As expected, a smaller λ_n tends to yield a smaller bias. The number of folds seems irrelevant in terms of bias. In this example, increasing the number of bases does not necessarily produce better bias, but the difference with respect to a case with smaller r does not seem of practical relevance. The same holds for the specific choice for γ's and the machine learning tool employed. While the bias is always smaller for $\hat{\theta}_k$, we see that θ_{01} and $\theta_{0\omega}$ are more difficult to learn properly. Nevertheless, this issue disappears for larger n's.

More importantly, our procedure is able to control the size. As n becomes larger, the coverage gets closer to the nominal level. The three different choices of the tuning parameter produce almost the same coverage. While we observe some differences when L is larger, any apparent distinction disappears as n grows. A larger r causes a slight improvement in terms of coverage, however, the improvement disappears when n is sufficiently large. The choice of the specific basis seems innocuous for coverage. Random Forest tends to yield a coverage level slightly larger for parameters $\hat{\theta}_1$ and $\hat{\theta}_\omega$, possibly explained by the fact that the bias is shrinking faster than the variance as n grows. This might be prevented by exploring other choices of the tuning parameters used in the first stage when η_0 is estimated. The main takeaway is that we see relatively stable performance over the choices considered in this simulation example. These observations, then, are in line with our theoretical results. As we emphasized several times in the text, our setting is challenging from an econometric perspective as we are dealing with unknown operators (or an unknown Gram matrix). Yet, D-CMRs perform well for sample sizes that can be arguably regarded as small. In conclusion, our Monte Carlo results allow us to be confident about the good properties of our estimation strategy.

Table 1: Monte Carlo Results - Bias and 95% Coverage

			$n = 250$				
Est.	Smaller	Larger	λ_n	Larger	Larger	Fourier	Random
	λ_n	λ_n	(BCCH)	L	r	Basis	Forest
Bias ($\hat{\theta}_1$)	0.095	0.097	0.100	0.105	0.095	0.105	0.100
Cov95%	0.935	0.934	0.936	0.912	0.937	0.948	0.914
Bias ($\hat{\theta}_k$)	-0.031	-0.039	-0.041	-0.044	-0.036	-0.046	-0.042
Cov95%	0.912	0.913	0.906	0.894	0.910	0.925	0.918
Bias ($\hat{\theta}_\omega$)	-0.160	-0.162	-0.163	-0.165	-0.160	-0.166	-0.253
Cov95%	0.738	0.742	0.739	0.651	0.745	0.733	0.777
			$n = 500$				
Est.	Smaller	Larger	λ_n	Larger	Larger	Fourier	Random
	λ_n	λ_n	(BCCH)	L	r	Basis	Forest
Bias ($\hat{\theta}_1$)	0.048	0.061	0.059	0.060	0.059	0.071	0.035
Cov95%	0.943	0.939	0.947	0.927	0.941	0.959	0.963
Bias ($\hat{\theta}_k$)	-0.013	-0.029	-0.027	-0.027	-0.027	-0.040	-0.021
Cov95%	0.903	0.935	0.927	0.894	0.935	0.935	0.949
Bias ($\hat{\theta}_\omega$)	-0.081	-0.088	-0.087	-0.074	-0.087	-0.095	-0.103
Cov95%	0.926	0.922	0.922	0.886	0.922	0.919	0.970

NOTE: The table shows the bias and the 95% coverage of D-CMRs, across different specifications of our algorithm. Except for specific situations, we have estimated η_0 with boosting, $L = 4$, γ's are constructed from exponential basis, $r = 9$, and λ_n is the recommendation by Belloni, Chen, Chernozhukov, and Hansen (2012) (p. 2380, BCCH). All these choices are used in the specification appearing in the fourth column of the table. Smaller λ_n refers to the case $\lambda_n = \frac{1.01}{\sqrt{n-n_\ell}}\Phi^{-1}\left(1-\frac{c_2}{2r}\right)$, with $c_2 = 2/\log(\log(\log((n-n_\ell)\vee r)))$. Larger λ_n stands for the case $\lambda_n = \frac{1.3}{\sqrt{n-n_\ell}}\Phi^{-1}\left(1-\frac{c_2}{2r}\right)$, with $c_2 = 0.1/\log((n-n_\ell)\vee r)$. Larger L is when $L = 6$. Larger r reproduces the estimation with $r = 25$. Fourier Basis employs these for the γ's. We use random forest to estimate η_0 in the last column of the table. Results are based on $1,000$ Monte Carlo repetitions.

Table 2: Monte Carlo Results - Bias and 95% Coverage (continued)

	$n = 750$						
Est.	Smaller	Larger	λ_n	Larger	Larger	Fourier	Random
	λ_n	λ_n	(BCCH)	L	r	Basis	Forest
Bias ($\hat{\theta}_1$)	0.028	0.039	0.037	0.038	0.039	0.053	0.022
Cov95%	0.944	0.946	0.949	0.955	0.958	0.965	0.980
Bias ($\hat{\theta}_k$)	-0.002	-0.020	-0.017	-0.017	-0.020	-0.037	-0.018
Cov95%	0.880	0.929	0.925	0.924	0.930	0.944	0.945
Bias ($\hat{\theta}_\omega$)	-0.018	-0.025	-0.023	-0.012	-0.025	-0.033	-0.041
Cov95%	0.952	0.951	0.954	0.952	0.951	0.950	0.990

NOTE: The table shows the bias and the 95% coverage of D-CMRs, across different specifications of our algorithm. Except for specific situations, we have estimated η_0 with boosting, $L = 4$, γ's are constructed from exponential basis, $r = 9$, and λ_n is the recommendation by Belloni, Chen, Chernozhukov, and Hansen (2012) (p. 2380, BCCH). All these choices are used in the specification appearing in the fourth column of the table. Smaller λ_n refers to the case $\lambda_n = \frac{1.01}{\sqrt{n-n_\ell}}\Phi^{-1}\left(1-\frac{c_2}{2r}\right)$, with $c_2 = 2/\log(\log(\log((n-n_\ell)\vee r)))$. Larger λ_n stands for the case $\lambda_n = \frac{1.3}{\sqrt{n-n_\ell}}\Phi^{-1}\left(1-\frac{c_2}{2r}\right)$, with $c_2 = 0.1/\log((n-n_\ell)\vee r)$. Larger L is when $L = 6$. Larger r reproduces the estimation with $r = 25$. Fourier Basis employs these for the γ's. We use random forest to estimate η_0 in the last column of the table. Results are based on $1,000$ Monte Carlo repetitions.

10 Final Remarks

This paper has extended the construction of LR/orthogonal/debiased moments to general models defined by a finite number of CMRs, with possible different conditioning variables and endogenous regressors. As we have argued, our strategy exploits the CMRs implied by the model in a general way, and thus can be applied in a wide variety of settings. Hence, our approach will hopefully pave the way for the employment of machine learning techniques in contexts where the construction of LR has remained unexplored such as non-linear GMM settings, missing data models, production functions at the firm level, dynamic discrete choice models, simultaneous equations models, and many others. Conveniently, our construction is based on a Lasso-type program, making it straightforward to implement. We encourage researchers to use our recipe to convert typical "ad-hoc" choices of IVs into valid OR-IVs, leading to a LR moment, especially important in high-dimensional contexts.

Our theoretical results and Monte Carlo experiments motivate us to leverage our procedure further. In future versions of this work, we plan to use data from a panel of Chilean firms, from 1979 to 1986. This data has been extensively studied by the production function literature; see, e.g., Levinsohn and Petrin (2003), Ackerberg et al. (2015), and Gandhi et al. (2020).[27] Our motivation to work with this application is that production function estimation and measures of productivity play a central role in several empirical settings in economics, with important implications for policymaking. For instance, production functions have been used to study the effects of trade liberalization, exporting, foreign

[27] We are indebted to David A. Rivers and Salvador Navarro for sharing the data with us and answering our questions regarding the construction of the variables.

ownership, competition, importing intermediate goods, investment climate, and learning by doing (see Ackerberg et al., 2007, 2015, and references therein). To the best of our knowledge, no previous work has constructed debiased moments for all the interesting parameters in these settings using a semiparametric perspective. Hence, we are interested in determining if our strategy can uncover larger heterogeneity patterns among production functions than previously recognized. We leave such an exploration for future versions of this paper.

We recognize that our paper has its limitations. We have assumed that our limited number of LR moments are sufficient to identify θ_0. Additionally, taking identification for granted, we have ignored that efficiency can be improved by selecting other orthogonal moments. These are crucial matters that should not be overlooked. It is essential to explore the construction of debiased moments that are assured to preserve identification in models defined by a number of CMRs using modern tools, following, e.g., Muandet et al. (2020) and Zhang et al. (2021). Furthermore, as we have seen, several κ_0's might exist. This raises the question of whether it is possible to characterize a suitable notion of "optimality" among these OR-IVs. From an efficiency standpoint, it is well-known that the first best is the optimal IV (see, e.g., Chamberlain, 1992b). This special κ not only yields an estimator that achieves the efficient semiparametric bound, but it is also a valid OR-IV (Newey, 1990; van der Vaart, 1998). Such a choice, nonetheless, is difficult to estimate in general settings, which might explain why it has not been popular among applied researchers. It might be interesting to define a broader criterion that yields second-best choices of OR-IVs that are guaranteed to improve efficiency in estimation relative to other OR-IVs and whose computation is tractable in practice. Moreover, a more general theory for the estimation of OR-IVs can be derived. We have performed our construction exclusively for the space of sparse functions. It might be promising to develop a general framework for different functional spaces $\mathcal{G}_n$, including Reproducing Kernel Hilbert Spaces and Neural Networks. Finally, observe that the algorithm that this paper proposes only serves for debiasing structural parameters θ_0's. It might be important to reproduce the exercise for more general parameters, which include smooth functions of high-dimensional parameters, e.g., an average partial effect that depends on η_0. While the theoretical characterization of debiased moments for such parameters has already been derived by Argañaraz and Escanciano (2023), a suitable implementation routine for it remains to be explored. Hopefully, these ideas will be addressed in subsequent works.

Appendix

A Orthogonality Results

Proof of Proposition 1: The existence, linearity, and continuity of $S^*_{\theta_0,\eta_0}$ follows from Theorem 2.21 in Carrasco et al. (2007). For any function $g \in L^2(Z)$, by definition, $S^*_{\theta_0,\eta_0}$ satisfies

$$\langle S_{\theta_0,\eta_0} b, g\rangle_{L^2(Z)} = \left\langle b, S^*_{\theta_0,\eta_0} g\right\rangle_{\boldsymbol{B}}.$$

Then, the result of the proposition easily follows by writing:

$$\begin{aligned}\langle S_{\theta_0,\eta_0} b, g\rangle_{L^2(Z)} &= \sum_{j=1}^{J} \mathbb{E}\left[\mathbb{E}\left[b(X)'\tilde{\nu}_j\left(Y,\theta_0,\eta_0\right)\Big|\, Z_j\right] g_j(Z_j)\right] \\ &= \sum_{j=1}^{J} \mathbb{E}\left[b(X)'\tilde{\nu}_j\left(Y,\theta_0,\eta_0\right) g_j(Z_j)\right] \\ &= \mathbb{E}\left[b(X)' \sum_{j=1}^{J} \mathbb{E}\left[\tilde{\nu}_j\left(Y,\theta_0,\eta_0\right) g_j(Z_j)\right| X\right]\right], \end{aligned} \tag{A.1}$$

where the first equality holds under Assumption 2, and the second and third equality uses the law of iterated expectation. Hence, expression (A.1) implies

$$S^*_{\theta_0,\eta_0} g = \sum_{j=1}^{J} \mathbb{E}\left[\tilde{\nu}_j\left(Y,\theta_0,\eta_0\right) g_j(Z_j)\right| X\right]. \blacksquare$$

B Justification of the Algorithm for Estimation of OR-IVs

The justification of the proposed algorithm for estimation of OR-IVs follows by similar arguments as presented by Belloni et al. (2012) and Chernozhukov et al. (2022d). Particularly, we have applied Algorithm A.1 in Belloni et al. (2012) to our context. $\hat{D}_{jk\ell} \equiv \hat{D}_{l\ell}$ is the normalization term, which is constructed as the square root of the empirical second moment of the regressors of the problem times the corresponding residuals (a normalization of the first-order conditions of the unrestricted problem). The formula for λ_n is the same as in Chernozhukov et al. (2022d), which is the one recommended by Belloni et al. (2012).[28] As a result, the procedure that we propose based on tuning parameter λ_n is justified by a similar argument to Theorem 1 in Belloni et al. (2012).

Observe, nevertheless, that there exists a difference between our program and the type of Lasso problems considered by Belloni et al. (2012) and Chernozhukov et al. (2022d). As we emphasized in the main text, regressors are unknown in our case and thus we need to estimate them, while the aforementioned papers work under the standard situation where regressors are known. In any case,

[28] After we properly account for the fact that $\lambda_n = \tilde{\lambda}_n/n$, where $\tilde{\lambda}_n$ is the tuning parameter in Belloni et al. (2012).

the key condition in Theorem 1 in Belloni et al. (2012) ask for the asymptotic validity of the penalty loadings, i.e., $\hat{D}_{l\ell}$. Let $\hat{D}^0_{l\ell} = \hat{D}^0_{jk\ell}$ be the “ideal” penalty loadings, which are defined as

$$\hat{D}^0_{j'k\ell} = \left[\frac{1}{n-n_\ell} \sum_{i \notin I_\ell} \left\{ \sum_{j=1}^{J} \mathbb{E}\left[\left(\mathbb{E}\left[\tilde{\nu}_{j'} \left(Y, \theta_0, \eta_{0j'} \right) \gamma_{j'k} \left(Z_{jk} \right) \middle| X \right] \right)' \tilde{\nu}_j \left(Y, \theta_0, \eta_{0j} \right) \middle| Z_j \right] \epsilon^0_j \right\}^2 \right]^{1/2},$$

where ϵ^0_j is the $j-th$ entry of the vector $f(Z) - f^*(Z)$, i.e., the difference between the starting instrument and its orthogonal projection on $\overline{\mathcal{R}\left(S_{\theta_0,\eta_0} S^*_{\theta_0,\eta_0} \right)}$. Then, the ideal loadings are constructed from the “population” regressors and ϵ^0_j. Our estimated loadings are asymptotically valid if they obey

$$a\hat{D}^0_l \leq \hat{D}_l \leq b\hat{D}^0_l, \tag{B.1}$$

where $0 < a \leq 1 \leq b$ such that $a \xrightarrow{p} 1$ and $b \xrightarrow{p} b'$, with $b' \geq 1$ (cf. Equation (3.2) in Belloni et al. (2012)). Condition (B.1) is written in terms of the estimated regressors, as it involves $\hat{D}_l$. But, under a mild convergence condition, it can be written in terms of loading based on the population regressors, with probability approaching one. In this case, condition (B.1) will be analogous to the condition required by Theorem 1 in Belloni et al. (2012). Let us define

$$\tilde{D}^0_{j'k\ell} = \left[\frac{1}{n-n_\ell} \sum_{i \notin I_\ell} \left\{ \sum_{j=1}^{J} \mathbb{E}\left[\left(\mathbb{E}\left[\tilde{\nu}_{j'} \left(Y, \theta_0, \eta_{0j'} \right) \gamma_{j'k} \left(Z_{jk} \right) \middle| X \right] \right)' \tilde{\nu}_j \left(Y, \theta_0, \eta_{0j} \right) \middle| Z_j \right] \tilde{\epsilon}_{ji\ell} \right\}^2 \right]^{1/2},$$

$$\tilde{\epsilon}_{ji\ell} = f_j \left(Z_{ji} \right) - \sum_{j'=1}^{J} \sum_{k=1}^{r_{j'}} \hat{\beta}_{j'k\ell} \mathbb{E}\left[\left(\mathbb{E}\left[\tilde{\nu}_{j'} \left(Y_i, \theta_0, \eta_{0j} \right) \gamma_{j'k} \left(Z_{j'i} \right) \middle| X \right] \right)' \tilde{\nu}_j \left(Y_i, \theta_0, \eta_{0j} \right) \middle| Z_{ji} \right].$$

In addition, let us ignore the subscript associated with cross-fitting, and assume

Assumption 16. *There exists a neighborhood $\mathcal{N}$ of θ_0 and $||\cdot||$ such that for $\theta \in \mathcal{N}$ and for*

$$\left(\sum_{j=1}^{J} ||T_j \left(\eta_j - \eta_{0j} \right)||^2_2 \right)^{1/2}$$

small enough, there exists a function $h(W, \eta)$ and a C such that

$$\left| \hat{D}_l(\theta) - \tilde{D}_l \right| \leq h(W, h) \, ||\theta - \theta_0|| \, , \quad \mathbb{E}\left[h(W, \eta) \right] < C,$$

where $\hat{D}_l(\theta)$ is the estimated loading evaluated at θ.

Then, by the Conditional Markov inequality, Assumption 16 implies that, with probability approaching one, $h(W, \hat{\eta}) = O_p(1)$. Moreover, let $\bar{\theta} \xrightarrow{p} \theta_0$, with probability approaching one,

$$\left| \hat{D}_l(\bar{\theta}) - \tilde{D}_l \right| \leq h(W, \hat{\eta}) \left| \left| \bar{\theta} - \theta_0 \right| \right| = O_p(1) o_p(1) \xrightarrow{p} 0.$$

Hence, $\hat{D}_\ell(\bar{\theta}) \xrightarrow{p} \tilde{D}_l$ follows by the Conditional Markov inequality. This implies that with probability approaching one, condition (B.1) is equivalent to

$$a\hat{D}^0_l \leq \tilde{D}_l \leq b\hat{D}^0_l,$$

where notice that now $\tilde{D}_l$ depends on the population regressors and then we are in an analogous case to the one considered by Belloni et al. (2012).

C Optimization

C.1 Algorithm

Step 4 of the iterative algorithm above requires to solve

$$\min_{\beta \in \mathbb{R}^r} \sum_{j=1}^{J} \frac{1}{n-n_\ell} \left(\boldsymbol{f_{j\ell}} - \hat{\boldsymbol{M}}_{j\ell}\beta \right)' \left(\boldsymbol{f_{j\ell}} - \hat{\boldsymbol{M}}_{j\ell}\beta \right) + 2\lambda_n \left|\left| \hat{D}_\ell \beta \right|\right|_1, \tag{C.1}$$

where $\hat{D}_\ell$ is a diagonal matrix with elements $\hat{D}_{jk\ell} \equiv \hat{D}_{l\ell}$ along the main diagonal, with $l = 1, \cdots, r$. Hence, the first r_1 entries correspond to the regressors with $\gamma_1(Z_1)$, the next r_2 entries are the regressors with $\gamma_2(Z_2)$, and so on. To solve (C.1), we use an extension of the coordinate descent approach for Lasso (Friedman et al., 2007, 2010; Fu, 1998) to our particular objective function. To be precise, we implement a coordinate-wise descent algorithm with a soft-thresholding update. Let v_l denote the l^{th} element of a generic vector v and let e_l be a $r \times 1$ unit vector with 1 in the l^{th} coordinate and zeros elsewhere. This algorithm can be implemented as follows:

For $l = 1 : r$, do

Step 1: Compute loadings (which do not depend on β_k):

$$A_l = \frac{1}{n-n_\ell} \sum_{j=1}^{J} e_l' \hat{\boldsymbol{M}}_{\boldsymbol{j}}' \left(\boldsymbol{f_j} - \hat{\boldsymbol{M}}_{\boldsymbol{j}}\beta + \hat{\boldsymbol{M}}_{\boldsymbol{j}} e_l \beta_l \right)$$

$$B_l = \frac{1}{n-n_\ell} \sum_{j=1}^{J} e_l' \hat{\boldsymbol{M}}_{\boldsymbol{j}}' \hat{\boldsymbol{M}}_{\boldsymbol{j}} e_l.$$

Step 2: Update coordinate β_l:

$$\beta_l = \begin{cases} \frac{A_l + \hat{D}_l \lambda_n}{B_l} & \text{if} \quad A_l < -\hat{D}_l \lambda_n \\ 0 & \text{if} \quad A_l \in \left[-\hat{D}_l \lambda_n, \hat{D}_l \lambda_n \right] \\ \frac{A_l - \hat{D}_l \lambda_n}{B_l} & \text{if} \quad A_l > \hat{D}_l \lambda_n. \end{cases}$$

C.2 Justification

In this section, we justify the previous coordinate-wise soft-thresholding update. Observe that

$$\frac{\partial}{\partial \beta_l} \left[\sum_{j=1}^{J} \frac{1}{n-n_\ell} \left(\boldsymbol{f_{j\ell}} - \hat{\boldsymbol{M}}_{\boldsymbol{j\ell}}\beta \right)' \left(\boldsymbol{f_{j\ell}} - \hat{\boldsymbol{M}}_{\boldsymbol{j\ell}}\beta \right) \right] = -2A_l + 2B_l\beta_l,$$

where we note that neither A_l nor B_l depend on β_l. The subgradient of the penalty term is

$$\frac{\partial}{\partial \beta_l} 2 \left|\left| \hat{D}\beta \right|\right|_1 = \begin{cases} -2\hat{D}_l \lambda_n & \text{if} \quad \beta_l < 0 \\ \left[-2\hat{D}_l \lambda_n, 2\hat{D}_l \lambda_n \right] & \text{if} \quad \beta_l = 0 \\ 2\hat{D}_l \lambda_n & \text{if} \quad \beta_l > 0 \end{cases}$$

Therefore, ((1/2) of) the subgradient of the objective function of our program is

$$\frac{\partial}{\partial \beta_l}\frac{1}{2}\left[\sum_{j=1}^{J}\frac{1}{n-n_\ell}\left(\boldsymbol{f_{j\ell}}-\hat{\boldsymbol{M}}_{\boldsymbol{j\ell}}\beta\right)'\left(\boldsymbol{f_{j\ell}}-\hat{\boldsymbol{M}}_{\boldsymbol{j\ell}}\beta\right)+2\left|\left|\hat{D}\beta\right|\right|_1\right]=\begin{cases}-A_l+B_l\beta_l-\hat{D}_l\lambda_n & \text{if} \quad \beta_l<0\\ \left[-A_l-\hat{D}_l\lambda_n,-A_l+\hat{D}_l\lambda_n\right] & \text{if} \quad \beta_l=0\\ -A_l+B_l\beta_l+\hat{D}_l\lambda_n & \text{if} \quad \beta_l>0\end{cases}$$

Hence, equalizing those terms to zero and solving for β_l gives the element-wise update provided above.

Note that the first term of the objective function in (C.1) is differentiable and convex and the penalty term is the sum of convex functions. Hence, the whole objective function in (C.1) is a particular case of Equation 21 in Friedman et al. (2007), and thus the coordinate descent converges to the solution to (C.1) (Tseng, 2001).

D Asymptotic Results of OR-IVs

Lemma 5. *Let Assumptions 3 and 4 hold. Then,*

$$\left|\left|\hat{G}_{j\ell}-G_j\right|\right|_\infty=O_p\left(\varepsilon_n\right), \quad \varepsilon_n=\sqrt{\frac{\log(r)}{n}}.$$

Proof of Lemma 5: Let $\tilde{G}_{j\ell}=\frac{1}{n-n_\ell}\sum_{i\notin I_\ell}M_j\left(Z_{ji}\right)M_j\left(Z_{ji}\right)'$. Then, by the triangle inequality,

$$\left|\left|\hat{G}_{j\ell}-G_j\right|\right|_\infty\leq\left|\left|\hat{G}_{j\ell}-\tilde{G}_{j\ell}\right|\right|_\infty+\left|\left|\tilde{G}_{j\ell}-G_j\right|\right|_\infty.$$

We first show that $\left|\left|\tilde{G}_{j\ell}-G_j\right|\right|_\infty=O_p\left(\varepsilon_n\right)$. To prove this, we follow the proofs of Lemma A10 of Chernozhukov et al. (2022d) and Lemma D.1 of Bakhitov (2022). Let us define

$$T_{iqk}^{j}=M_{jq}\left(Z_{ji}\right)M_{jk}\left(Z_{ji}\right)-\mathbb{E}\left[M_{jq}\left(Z_{ji}\right)M_{jk}\left(Z_{ji}\right)\right], \quad U_{qk}^{j}=\frac{1}{n-n_\ell}\sum_{i\notin I_\ell}T_{iqk}^{j}.$$

where $M_{jk}\left(Z_{ji}\right)$ is the $k-th$ element of the vector $M_j\left(Z_{ji}\right)$. Note that in the previous displays, the elements just defined depend on ℓ, but we omitted this dependence to simplify the exposition. Then, for any constant C, we have

$$\begin{aligned}\mathbb{P}\left(\left|\left|\tilde{G}_{j\ell}-G_j\right|\right|_\infty\geq C\varepsilon_n\right)&\leq\sum_{q,k=1}^{r}\mathbb{P}\left(\left|U_{qk}^{j}\right|\geq C\varepsilon_n\right)\\&\leq r^2\max_{k,q}\mathbb{P}\left(\left|U_{qk}^{j}\right|\geq C\varepsilon_n\right).\end{aligned}$$

Note that $\mathbb{E}\left[T_{iqk}^{j}\right]=0$ and by Assumption 3,

$$\begin{aligned}\left|T_{iqk}^{j}\right|&\leq\left|M_{jq}\left(Z_{ji}\right)\right|\left|M_{jk}\left(Z_{ji}\right)\right|+\mathbb{E}\left[\left|M_{jq}\left(Z_{ji}\right)\right|\left|M_{jk}\left(Z_{ji}\right)\right|\right]\\&\leq 2c_j^2.\end{aligned}$$

The previous fact shows that T^j_{iqk} is a bounded random variable. Therefore, it is sub-Gaussian. Let $\left|\left|T^j_{iqk}\right|\right|_{\Psi_w}$ denote the sub-Gaussian norm. Then, $K_j = \frac{2c_j^2}{\log 2} \geq \left|\left|T^j_{iqk}\right|\right|_{\Psi_w}$. By Hoeffding's inequality (see Thereom 2.6.1 in Vershynin, 2018), there is a constant c such that

$$\begin{aligned} r^2 \max_{k,q} \mathbb{P}\left(\left|U^j_{qk}\right| \geq C\varepsilon_n\right) &\leq 2r^2 \exp\left(-\frac{cC^2\log(r)}{K_j^2}\right) \\ &= 2\exp\left(\log(r)\left[2 - \frac{cC^2}{K_j^2}\right]\right) \\ &\longrightarrow 0, \end{aligned}$$

for any C such that $C \geq K_j\sqrt{\frac{2}{c}}$. Hence, for C large enough, $\mathbb{P}\left(\left|\left|\tilde{G}_{j\ell} - G_j\right|\right|_\infty \geq C\varepsilon_n\right) \to 0$ as $r \to \infty$, as needed.

Next, let $\mathcal{W}^c_\ell$ contain the data for each $i \in I^c_\ell$. Then, each estimated element in the matrix $\hat{G}_{j\ell}$ depends on $\mathcal{W}^c_\ell$ only. Now, define

$$P^j_{i\ell qk} = \hat{M}_{j\ell q}\left(Z_{ji}\right)\hat{M}_{j\ell k}\left(Z_{ji}\right) - M_{jq}\left(Z_{ji}\right)M_{jk}\left(Z_{ji}\right), \quad Q^j_{\ell qk} = \frac{1}{n - n_\ell}\sum_{i \in I_\ell} P^j_{i\ell qk}.$$

Conditional on $\mathcal{W}^c_\ell$, by the the conditional Markov's inequality and the triangle inequality, we can write for any $C \geq 0$

$$\begin{aligned} \mathbb{P}\left(\left|\left|\hat{G}_{j\ell} - \tilde{G}_{j\ell}\right|\right|_\infty \geq C\varepsilon_n \middle| \mathcal{W}^c_\ell\right) &\leq \mathbb{P}\left(\max_{k,q}\left|Q^j_{\ell qk}\right| \geq C\varepsilon_n \middle| \mathcal{W}^c_\ell\right) \\ &\leq \mathbb{P}\left(\left|Q^j_{\ell q^* k^*}\right| \geq C\varepsilon_n \middle| \mathcal{W}^c_\ell\right) \\ &\leq \frac{1}{C\varepsilon_n}\mathbb{E}\left[\left|Q^j_{\ell q^* k^*}\right| \middle| \mathcal{W}^c_\ell\right] \\ &\leq \frac{1}{C\varepsilon_n}\mathbb{E}\left[\left|P^j_{i\ell q^* k^*}\right| \middle| \mathcal{W}^c_\ell\right] \\ &\leq \frac{1}{C\varepsilon_n}\mathbb{E}\left[\max_{k,q}\left|P^j_{i\ell qk}\right| \middle| \mathcal{W}^c_\ell\right] \\ &\longrightarrow 0, \end{aligned}$$

where $\max_{k,q}\left|Q^j_{\ell qk}\right| \equiv \left|Q^j_{\ell q^* k^*}\right|$, and the last display follows from Assumption 4. We have shown then

$$\left|\left|\hat{G}_{j\ell} - G_j\right|\right|_\infty \leq O_p\left(\varepsilon_n\right) + O_p\left(\varepsilon_n\right) = O_p\left(\varepsilon_n\right),$$

as needed. ■

Let us define β_* as

$$\beta_* \in \arg\min_{v \in \mathbb{R}^r} \left(\beta_0 - v\right)' \sum_{j=1}^J G_j\left(\beta_0 - v\right) + 2\varepsilon_n \sum_{k \in S^c_{\bar{\beta}}} \left|v_k\right|. \tag{D.1}$$

A maintained assumption throughout this work is that $||\beta_*||_1 = O_p(1)$.

Lemma 6. $\left|\left|\sum_{j=1}^{J} G_j\left(\beta_* - \beta_0\right)\right|\right|_\infty \leq \varepsilon_n$.

Proof of Lemma 6: The first order condition (sub-gradient of the objective function) for β_* implies that for $k \in S_{\bar{\beta}}$, we have $e_k^{'} \sum_{j=1}^{J} G_j\left(\beta_* - \beta_0\right) = 0$, where e_k is the $k-th$ column of an identity matrix I_r of dimension $r \times r$. For $k \in S_{\bar{\beta}}^c$, we have that $e_k^{'} \sum_{j=1}^{J} G_j\left(\beta_* - \beta_0\right) + \varepsilon_n \pi_k = 0$, where $\pi_k = sign\left(\beta_{*k}\right)$ if $\beta_{*k} \neq 0$ and $\pi_k \in [-1, 1]$ if $\beta_{*k} = 0$. Hence, in any case $\left|e_k^{'} \sum_{j=1}^{J} G_j\left(\beta_* - \beta_0\right)\right| \leq \varepsilon_n$. ■

Lemma 7. $\left(\beta_0 - \beta_*\right) \sum_{j=1}^{J} G_j\left(\beta_0 - \beta_*\right) \leq Cs\varepsilon_n^2$.

Proof of Lemma 7: By definition of β_*,

$$\begin{aligned}\left(\beta_0 - \beta_*\right)^{'} \sum_{j=1}^{J} G_j\left(\beta_0 - \beta_*\right) + 2\varepsilon_n \sum_{k \in S_{\bar{\beta}}^c} \left|\beta_{*,k}\right| &\leq \left(\beta_0 - \bar{\beta}\right)^{'} \sum_{j=1}^{J} G_j\left(\beta_0 - \bar{\beta}\right) + 2\varepsilon_n \sum_{k \in S_{\bar{\beta}}^c} \left|\bar{\beta}_k\right| \\ &= \left(\beta_0 - \bar{\beta}\right)^{'} \sum_{j=1}^{J} G_j\left(\beta_0 - \bar{\beta}\right)\end{aligned} \tag{D.2}$$

Let β_0 be the linear projection of f^* on $M = (M_1, \cdots, M_J)$ in the sense that

$$\sum_{J=1}^{J} \mathbb{E}\left[M_j\left(Z_j\right)\left(f_j^*\left(Z_j\right) - M_j\left(Z_j\right)^{'}\beta_0\right)\right] = 0.$$

Next, notice that by the triangle inequality

$$\begin{aligned}\left(\beta_0 - \bar{\beta}\right)^{'} \sum_{j=1}^{J} G_j\left(\beta_0 - \bar{\beta}\right) &= \sum_{j=1}^{J} \mathbb{E}\left[\left\{M_j\left(Z_j\right)^{'}\left(\beta_0 - \bar{\beta}\right)\right\}^2\right] \\ &= \sum_{j=1}^{J} \left|\left|M_j(Z_j)^{'}\beta_0 - M_j(Z_j)^{'}\bar{\beta}\right|\right|_2^2 \\ &\leq 2\sum_{j=1}^{J}\left(\left|\left|f_j^*(Z_j) - M_j(Z_j)^{'}\beta_0\right|\right|_2^2 + \left|\left|f_j^*(Z_j) - M_j(Z_j)^{'}\bar{\beta}\right|\right|_2^2\right) \\ &\leq 4\sum_{j=1}^{J} \left|\left|f_j^*(Z_j) - M_j(Z_j)^{'}\bar{\beta}\right|\right|_2^2 \\ &\leq Cs\varepsilon_n^2,\end{aligned} \tag{D.3}$$

where the last inequality follows from Assumption 5. The result then follows from Equation (D.2) and $\varepsilon_n \sum_{k \in S_{\bar{\beta}}^c} \left|\beta_{*,k}\right| \geq 0$. ■

Lemma 8. *Let S_{β_*} be the vector of indices of nonzero elements of β_*. Then, $s_* = \left|S_{\beta_*}\right| \leq Cs$.*

Proof of Lemma 8: For all $k \in S_{\beta_*} \backslash S_{\bar{\beta}}$, the first order conditions to (D.1) imply $\left|e_k^{'} \sum_{j=1}^{J} G_j\left(\beta_* - \beta_0\right)\right| = \varepsilon_n$. Therefore, it holds that

$$\sum_{k \in S_{\beta_*} \backslash S_{\bar{\beta}}} \left(e_k^{'} \sum_{j=1}^{J} G_j\left(\beta_* - \beta_0\right)\right)^2 = \varepsilon_n^2 \left|S_{\beta_*} \backslash S_{\bar{\beta}}\right|.$$

Furthermore, using Lemma 7 and the fact that the largest eigenvalue of $\sum_{j=1}^{J} G_j$ is bounded, we obtain

$$\begin{aligned}
\sum_{k\in S_{\beta_*}\backslash S_{\bar{\beta}}} \left(e_k^{'} \sum_{j=1}^{J} G_j \left(\beta_* - \beta_0\right) \right)^2 &\leq \sum_{k=1}^{r} \left(e_k^{'} \sum_{j=1}^{J} G_j \left(\beta_* - \beta_0\right) \right)^2 \\
&= \left(\beta_* - \beta_0\right)^{'} \sum_{j=1}^{J} G_j \left(\sum_{k=1}^{r} e_k e_k^{'} \right) \sum_{j=1}^{J} G_j \left(\beta_* - \beta_0\right) \\
&= \left(\beta_* - \beta_0\right)^{'} \left(\sum_{j=1}^{J} G_j \right)^2 \left(\beta_* - \beta_0\right) \\
&\leq \lambda_{max} \left(\sum_{j=1}^{J} G_j \right) \left\{ \left(\beta_* - \beta_0\right)^{'} \sum_{j=1}^{J} G_j \left(\beta_* - \beta_0\right) \right\} \\
&\leq Cs\varepsilon_n^2,
\end{aligned}$$

where $\lambda_{max}(A)$ denotes the maximum eigenvalue of an arbitrary matrix A. The previous result implies

$$\varepsilon_n^2 \left| S_{\beta_*} \backslash S_{\bar{\beta}} \right| \leq Cs\varepsilon_n^2.$$

Diving both sides of the previous expression by ε_n^2 yields $\left| S_{\beta_*} \backslash S_{\bar{\beta}} \right| \leq Cs$. Hence,

$$s_* = \left| S_{\bar{\beta}} \right| + \left| S_{\beta_*} \backslash S_{\bar{\beta}} \right| \leq s + Cs \leq Cs,$$

as needed ■.

Lemma 9. $\sum_{j=1}^{J} \mathbb{E}\left[\left(f_j^* \left(Z_j\right) - M_j \left(Z_j\right)^{'} \beta_* \right)^2 \right] \leq Cs\varepsilon_n^2.$

Proof of Lemma 9: By the triangle inequality and Assumption 5, we can write

$$\begin{aligned}
\sum_{j=1}^{J} \mathbb{E}\left[\left(f_j^* \left(Z_j\right) - M_j \left(Z_j\right)^{'} \beta_* \right)^2 \right] &\leq 2\sum_{j=1}^{J} \left|\left| f_j^* \left(Z_j\right) - M_j \left(Z_j\right)^{'} \bar{\beta} \right|\right|_2^2 + 2\sum_{j=1}^{J} \left|\left| M_j \left(Z_j\right)^{'} \bar{\beta} - M_j \left(Z_j\right)^{'} \beta_* \right|\right|_2^2 \\
&\leq Cs\varepsilon_n^2 + 4\sum_{j=1}^{J} \left|\left| M_j \left(Z_j\right)^{'} \bar{\beta} - M_j \left(Z_j\right)^{'} \beta_0 \right|\right|_2^2 \qquad \text{(D.4)} \\
&\quad + 4\sum_{j=1}^{J} \left|\left| M_j \left(Z_j\right)^{'} \beta_0 - M_j \left(Z_j\right)^{'} \beta_* \right|\right|_2^2 . \qquad \text{(D.5)}
\end{aligned}$$

Notice that by the result in (D.3),

$$\sum_{j=1}^{J} \left|\left| M_j \left(Z_j\right)^{'} \bar{\beta} - M_j \left(Z_j\right)^{'} \beta_0 \right|\right|_2^2 = \left(\beta_0 - \bar{\beta}\right)^{'} \sum_{j=1}^{J} G_j \left(\beta_0 - \bar{\beta}\right) \leq Cs\varepsilon_n^2. \qquad \text{(D.6)}$$

Moreover, by Lemma 7,

$$\sum_{j=1}^{J} \left|\left| M_j \left(Z_j\right)^{'} \beta_0 - M_j \left(Z_j\right)^{'} \beta_* \right|\right|_2^2 = \left(\beta_0 - \beta_*\right) \sum_{j=1}^{J} G_j \left(\beta_0 - \beta_*\right) \leq Cs\varepsilon_n^2, \qquad \text{(D.7)}$$

Plugging (D.6) into (D.4) and (D.7) into (D.5) yields the desired result. ■

Lemma 10. $\left|\left|\sum_{j=1}^{J}\hat{G}_j\beta_* - \sum_{j=1}^{J}G_j\beta_*\right|\right|_\infty = O_p\left(\varepsilon_n\right).$

Proof of Lemma 10: It can be easily verified that by definition of $||\cdot||_\infty$ and $||\cdot||_1$, we have

$$\begin{aligned}\left|\left|\sum_{j=1}^{J}\hat{G}_j\beta_* - \sum_{j=1}^{J}G_j\beta_*\right|\right|_\infty &= \left|\left|\sum_{j=1}^{J}\left(\hat{G}_j - G_j\right)\beta_*\right|\right|_\infty \\ &\leq \left|\left|\sum_{j=1}^{J}\left(\hat{G}_j - G_j\right)\right|\right|_\infty ||\beta_*||_1 \\ &= O_p\left(\varepsilon_n\right),\end{aligned}$$

as needed. ■

Lemma 11. *Let* $\Delta = \hat{\beta} - \beta_*$*. For any* $\hat{S}$ *such that* $\beta_{*,\hat{S}^c} = 0$ *with probability 1, with probability approaching 1,*

$$\Delta^{'}\sum_{j=1}^{J}\hat{G}_j\Delta \leq 3\lambda_n\left|\left|\Delta\right|\right|_1, \quad \left|\left|\Delta_{\hat{S}^c}\right|\right|_1 \leq 3\left|\left|\Delta_{\hat{S}}\right|\right|_1.$$

Proof of Lemma 11: By definition of $\hat{\beta}$, we have

$$\sum_{j=1}^{J}\left(\hat{\beta}^{'}\hat{G}_j\hat{\beta} - 2\hat{F}_j^{'}\hat{\beta}\right) + 2\lambda_n\left|\left|\hat{\beta}\right|\right|_1 \leq \sum_{j=1}^{J}\left(\beta_*^{'}\hat{G}_j\beta_* - 2\hat{F}_j^{'}\beta_*\right) + 2\lambda_n\left|\left|\beta_*\right|\right|_1.$$

Using $\hat{\beta} = \Delta + \beta_*$ in the previous expression and re-arranging terms, we obtain

$$\Delta^{'}\sum_{j=1}^{J}\hat{G}_j\Delta + 2\lambda_n\left|\left|\beta_* + \Delta\right|\right|_1 \leq 2\lambda_n\left|\left|\beta_*\right|\right| + 2\sum_{j=1}^{J}\left(\hat{F}_j - \hat{G}_j\beta_*\right)^{'}\Delta. \tag{D.8}$$

By definition of β_0, $\sum_{j=1}^{J}G_j\beta_0 - \sum_{j=1}^{J}F_j = 0$. Then, by Assumption 7, Lemma 6, Lemma 10, and the triangle inequality, we have

$$\begin{aligned}\left|\left|\sum_{j=1}^{J}\left(\hat{G}_j\beta_* - \hat{F}_j\right)\right|\right|_\infty &\leq \left|\left|\sum_{j=1}^{J}\left(\hat{G}_j\beta_* - G_j\beta_*\right)\right|\right|_\infty + \left|\left|\sum_{j=1}^{J}\left(G_j\beta_* - \hat{F}_j\right)\right|\right|_\infty \\ &\leq \left|\left|\sum_{j=1}^{J}\left(\hat{G}_j\beta_* - G_j\beta_*\right)\right|\right|_\infty + \left|\left|\sum_{j=1}^{J}\left(F_j - \hat{F}_j\right)\right|\right|_\infty + \left|\left|\sum_{j=1}^{J}\left(G_j\beta_* - F_j\right)\right|\right|_\infty \\ &\leq O_p\left(\varepsilon_n\right) + O_p\left(\varepsilon_n\right) + \left|\left|\sum_{j=1}^{J}\left(G_j\beta_* - F_j\right)\right|\right|_\infty \\ &\leq O_p\left(\varepsilon_n\right) + \left|\left|\sum_{j=1}^{J}\left(G_j\beta_0 - F_j\right)\right|\right|_\infty + \left|\left|\sum_{j=1}^{J}G_j\left(\beta_* - \beta_0\right)\right|\right|_\infty \\ &= O_p\left(\varepsilon_n\right).\end{aligned}$$

Therefore, by the Hölder's inequality, we have that $\left|\sum_{j=1}^{J}\left(\hat{F}_j - \hat{G}\beta_*\right)'\Delta\right| \leq \left|\left|\sum\left(\hat{F}_j - \hat{G}\beta_*\right)\right|\right|_\infty ||\Delta||_1 = O_p(\varepsilon_n)||\Delta||_1$. Recall that $\varepsilon_n = o(\lambda_n)$, and then we can write

$$\begin{aligned}\Delta'\sum_{j=1}^{J}\hat{G}_j\Delta + 2\lambda_n||\beta_* + \Delta||_1 &\leq 2\lambda_n||\beta_*||_1 + O_p(\varepsilon_n)||\Delta||_1\\ &\leq 2\lambda_n||\beta_*||_1 + \lambda_n||\Delta||_1, \qquad \text{(D.9)}\end{aligned}$$

with probability approaching 1. Moreover, by the triangle inequality, $||\beta_*||_1 \leq ||\beta_* + \Delta|| + ||\Delta||_1$. Plugging this into (D.9) results in $\Delta'\sum_{j=1}^{J}\hat{G}_j\Delta \leq 3\lambda_n||\Delta||_1$, and the first result of the lemma is obtained.

Furthermore, as $\Delta'\sum_{j=1}^{J}\hat{G}_j\Delta \geq 0$, it also follows from (D.9) that

$$2\lambda_n||\beta_* + \Delta||_1 \leq 2\lambda_n||\beta_*||_1 + \lambda_n||\Delta||_1. \qquad \text{(D.10)}$$

From the fact that $\beta_{*,\hat{S}^c} = 0$, it follows that $||\beta_* + \Delta||_1 = \left|\left|\beta_{*,\hat{S}} + \Delta_{\hat{S}}\right|\right|_1 + ||\Delta_{\hat{S}^c}||_1$ and $||\beta_*||_1 = \left|\left|\beta_{*,\hat{S}}\right|\right|_1$. Dividing both sides of (D.10) by λ_n and substituting the previous conclusions yields

$$\begin{aligned}2\left|\left|\beta_{*,\hat{S}} + \Delta_{\hat{S}}\right|\right|_1 + 2||\Delta_{\hat{S}^c}||_1 &\leq 2\left|\left|\beta_{*,\hat{S}}\right|\right|_1 + ||\Delta||_1\\ &= 2\left|\left|\beta_{*,\hat{S}}\right|\right|_1 + ||\Delta_{\hat{S}}||_1 + ||\Delta_{\hat{S}^c}||_1\\ &\leq 2\left(\left|\left|\beta_{*,\hat{S}} - \Delta_{\hat{S}}\right|\right|_1 + ||\Delta_{\hat{S}}||_1\right) + ||\Delta_{\hat{S}}||_1 + ||\Delta_{\hat{S}^c}||_1\\ &= 2\left|\left|\beta_{*,\hat{S}} - \Delta_{\hat{S}}\right|\right|_1 + 3||\Delta_{\hat{S}}||_1 + ||\Delta_{\hat{S}^c}||_1,\end{aligned}$$

where the second equality follows from the reverse triangle inequality. Subtracting $2\left|\left|\beta_{*,\hat{S}} + \Delta_{\hat{S}}\right|\right|_1 + ||\Delta_{\hat{S}^c}||_1$ from both sides in the previous displays yields

$$||\Delta_{\hat{S}^c}||_1 \leq 3||\Delta_{\hat{S}}||_1,$$

as needed. ■

Lemma 12. *$||\Delta||_2 \leq c\lambda_n\sqrt{s}$.*

Proof of Lemma 12: let N denote the indices corresponding to the largest $|S_{\beta_*}|$ entries in $\Delta_{S^c_{\beta_*}}$, so that $N \subseteq S^c_{\beta_*}$, $|N| = |S_{\beta_*}|$, and $|\Delta_k| \geq |\Delta_q|$ for any $k \in N$ and $q \in S^c_{\beta_*}\backslash N$. For $\hat{S} = S_{\beta_*} \cup N$ it follows from Assumption 6, Lemma 8, Lemma 11, and the Cauchy–Schwarz inequality that with probability approaching 1,

$$\begin{aligned}||\Delta_{\hat{S}}||_2^2 &\leq C\Delta'\sum_{j=1}^{J}\hat{G}_j\Delta\\ &\leq C\lambda_n||\Delta||_1\\ &= C\lambda_n\left(||\Delta_{\hat{S}}||_1 + ||\Delta_{\hat{S}^c}||_1\right)\\ &\leq C\lambda_n||\Delta_{\hat{S}}||_1\\ &\leq C\lambda_n\sqrt{s_*}||\Delta_{\hat{S}}||_2\\ &\leq C\lambda_n\sqrt{s}||\Delta_{\hat{S}}||_2.\end{aligned}$$

Then dividing through by $||\Delta_{\hat{S}}||_2$ then gives, with probability approaching 1,

$$||\Delta_{\hat{S}}||_2 \leq C\lambda_n\sqrt{s}. \tag{D.11}$$

By Lemma 6.9 of Bühlmann and Van De Geer (2011), Lemma 11, (D.11), and the Cauchy–Schwarz inequality

$$||\Delta_{\hat{S}^c}||_2 \leq \left(\left|\hat{S}\right|\right)^{-1/2}||\Delta_{\hat{S}^c}||_1 \leq 3\left(\left|\hat{S}\right|\right)^{-1/2}||\Delta_{\hat{S}}||_1 \leq 3\left(\left|\hat{S}\right|\right)^{-1/2}\left(\left|\hat{S}\right|\right)^{1/2}||\Delta_{\hat{S}}||_2 \leq C\lambda_n\sqrt{s}.$$

Hence, by the triangle inequality, with probability approaching one,

$$||\Delta||_2 \leq ||\Delta_{\hat{S}}||_2 + ||\Delta_{\hat{S}^c}||_2 \leq C\lambda_n\sqrt{s}. \blacksquare$$

Lemma 13. $\sum_{j=1}^J \mathbb{E}\left[\left(\left(M_j(Z_j) - \hat{M}_j(Z_j)\right)'\hat{\beta}\right)^2 \middle| \mathcal{W}_\ell^c\right] = O_p\left(s\lambda_n^2\right).$

Proof of Lemma 13: By the triangle inequality,

$$\begin{aligned}\sum_{j=1}^J \mathbb{E}\left[\left(\left(M_j(Z_j) - \hat{M}_j(Z_j)\right)'\hat{\beta}\right)^2 \middle| \mathcal{W}_\ell^c\right] \leq{}& 2\sum_{j=1}^J \mathbb{E}\left[\left(\left(M_j(Z_j) - \hat{M}_j(Z_j)\right)'\left(\hat{\beta}-\beta_*\right)\right)^2 \middle| \mathcal{W}_\ell^c\right] \\ &+ 2\sum_{j=1}^J \mathbb{E}\left[\left(\left(M_j(Z_j) - \hat{M}_j(Z_j)\right)'\beta_*\right)^2 \middle| \mathcal{W}_\ell^c\right]\end{aligned}$$

Let us provide a bound for the first term on the right-hand side above. By Assumption 8 and Lemma 12, with probability approaching 1

$$\begin{aligned}\sum_{j=1}^J \mathbb{E}\left[\left(\left(M_j(Z_j) - \hat{M}_j(Z_j)\right)'\left(\hat{\beta}-\beta_*\right)\right)^2 \middle| \mathcal{W}_\ell^c\right] &\leq \left(\hat{\beta}-\beta_*\right)' B\left(\hat{\beta}-\beta_*\right) \\ &\leq \lambda_{max}(B)\,||\Delta||_2^2 \\ &\leq Cs\varepsilon_n^2\lambda_n^2 \\ &\leq Cs\lambda_n^2\end{aligned} \tag{D.12}$$

By the same token, by Assumption 8, we write

$$\begin{aligned}\sum_{j=1}^J \mathbb{E}\left[\left(\left(M_j(Z_j) - \hat{M}_j(Z_j)\right)'\beta_*\right)^2 \middle| \mathcal{W}_\ell^c\right] &\leq \lambda_{max}(B)\,||\beta_*||_2^2 \\ &\leq C\varepsilon_n^2 \\ &\leq Cs\lambda_n^2,\end{aligned} \tag{D.13}$$

where the last inequality follows from $\varepsilon_n = o(\lambda_n)$. The results in (D.12) and (D.13) implies that

$$\sum_{j=1}^J \mathbb{E}\left[\left(\left(M_j(Z_j) - \hat{M}_j(Z_j)\right)'\hat{\beta}\right)^2 \middle| \mathcal{W}_\ell^c\right] \leq Cs\lambda_n^2,$$

as needed. $\blacksquare$

Lemma 14. $\sum_{j=1}^{J}\mathbb{E}\left[\left(f_j^*\left(Z_j\right)-\hat{M}_j\left(Z_j\right)^{\prime}\hat{\beta}\right)^2\middle|\mathcal{W}_\ell^c\right]=O_p\left(s\lambda_n^2\right)$.

Proof of Lemma 14: By the triangle inequality and Lemma 9,

$$\begin{aligned}\sum_{j=1}^{J}\mathbb{E}\left[\left(f_j^*\left(Z_j\right)-\hat{M}_j\left(Z_j\right)^{\prime}\hat{\beta}\right)^2\middle|\mathcal{W}_\ell^c\right]&\leq 2\sum_{j=1}^{J}\mathbb{E}\left[\left(f_j^*\left(Z_j\right)-\hat{M}_j\left(Z_j\right)^{\prime}\beta_*\right)^2\right]\\&\quad+2\sum_{j=1}^{J}\mathbb{E}\left[\left(M_j(Z_j)^{\prime}\beta_*-\hat{M}_j\left(Z_j\right)^{\prime}\hat{\beta}\right)^2\middle|\mathcal{W}_\ell^c\right]\\&\leq Cs\varepsilon_n^2+2\sum_{j=1}^{J}\mathbb{E}\left[\left(M_j(Z_j)^{\prime}\beta_*-\hat{M}_j\left(Z_j\right)^{\prime}\hat{\beta}\right)^2\middle|\mathcal{W}_\ell^c\right].\end{aligned}\tag{D.14}$$

Next, by the triangle inequality

$$\sum_{j=1}^{J}\mathbb{E}\left[\left(M_j(Z_j)^{\prime}\beta_*-\hat{M}_j\left(Z_j\right)^{\prime}\hat{\beta}\right)^2\middle|\mathcal{W}_\ell^c\right]\leq 2\sum_{j=1}^{J}\mathbb{E}\left[\left(M_j(Z_j)^{\prime}\left(\beta_*-\hat{\beta}\right)\right)^2\right]\tag{D.15}$$

$$+2\sum_{j=1}^{J}\mathbb{E}\left[\left(\left(M_j(Z_j)-\hat{M}_j(Z_j)\right)^{\prime}\hat{\beta}\right)^2\middle|\mathcal{W}_\ell^c\right]\tag{D.16}$$

We now find a bound for (D.15). Since the maximum eigenvalue of $\sum_{j=1}^{J}G_j$ is bounded, and using Lemma 12, we have

$$\sum_{j=1}^{J}\mathbb{E}\left[\left(M_j(Z_j)^{\prime}\left(\beta_*-\hat{\beta}\right)\right)^2\right]\leq\lambda_{max}\left(\sum_{j=1}^{J}G_j\right)||\Delta||_2^2\leq Cs\lambda_n^2.\tag{D.17}$$

Furthermore, by Lemma 13 we know that

$$\sum_{j=1}^{J}\mathbb{E}\left[\left(\left(M_j\left(Z_j\right)-\hat{M}_j\left(Z_j\right)\right)^{\prime}\hat{\beta}\right)^2\middle|\mathcal{W}_\ell^c\right]\leq Cs\lambda_n^2,\tag{D.18}$$

Plugging the results in (D.17) and (D.18) into (D.15) and (D.16), respectively yields

$$\sum_{j=1}^{J}\mathbb{E}\left[\left(M_j(Z_j)^{\prime}\beta_*-\hat{M}_j\left(Z_j\right)^{\prime}\hat{\beta}\right)^2\middle|\mathcal{W}_\ell^c\right]\leq Cs\lambda_n^2.$$

Using the last displays and that $\varepsilon_n=o\left(\lambda_n\right)$ in (D.14) gives the desired result. ■

Proof of Theorem 2: By Lemma 14,

$$\begin{aligned}||\kappa_0\left(Z\right)-\hat{\kappa}\left(Z\right)||_{L^2(Z)}^2&=\sum_{j=1}^{J}\mathbb{E}\left[\left(f_j^*\left(Z_j\right)-\hat{M}_j\left(Z_j\right)^{\prime}\hat{\beta}\right)^2\middle|\mathcal{W}_\ell^c\right]\\&\leq Cs\lambda_n^2,\end{aligned}$$

as needed. ■

E Asymptotic Results of the Parameter of Interest

Lemma 15. *Let Assumptions 3, 4, and 7 hold. In addition, suppose that $\varepsilon_n = o(\lambda_n)$. Then,*

$$\left|\left|\hat{\beta}\right|\right|_1 = O_p(1).$$

Proof of Lemma 15: We follow the proof of Lemma D.9 in Bakhitov (2022). Notice that Expression (D.9) in the proof of Lemma 11 implies

$$2\lambda_n \left|\left|\hat{\beta}\right|\right|_1 \leq 2\lambda_n \left|\left|\beta_*\right|\right|_1 + \lambda_n \left|\left|\hat{\beta} - \beta_*\right|\right|_1.$$

Next, let us divide by $2\lambda_n$ throughout, then by the triangle inequality

$$\begin{aligned}\left|\left|\hat{\beta}\right|\right|_1 &\leq \left|\left|\beta_*\right|\right|_1 + \frac{1}{2}\left|\left|\hat{\beta} - \beta_*\right|\right|_1 \\ &\leq \left|\left|\beta_*\right|\right|_1 + \frac{1}{2}\left(\left|\left|\hat{\beta}\right|\right| + \left|\left|\beta_*\right|\right|_1\right),\end{aligned}$$

with probability approaching one. Subtracting $\left|\left|\hat{\beta}\right|\right|/2$ from both sides in the previous display and multiplying by 2 yields

$$\left|\left|\hat{\beta}\right|\right|_1 \leq 3\left|\left|\beta_*\right|\right|_1 = O_p(1),$$

as needed. ■

Lemma 16. *Let Assumptions 3, 4, 7, 10, 11, and 12 hold. In addition, suppose that $\varepsilon_n = o(\lambda_n)$. Then,*

$$i) \int \left|\left|\hat{\Delta}_\ell(w)\right|\right|^2 F_0(dw) \xrightarrow{p} 0, \quad \textit{and} \quad ii) \sqrt{n}\int \hat{\Delta}_\ell(w) F_0(dw) \xrightarrow{p} 0.$$

Proof of Lemma 16: First, we show *i)*. Lemma 15 and Assumption 10 imply that $\sup_{z_j} |\hat{\kappa}_{j\ell}| = O_p(1)$ a.s., for any $\hat{\kappa}_{j\ell}$ in $\hat{\boldsymbol{\kappa}}_{\boldsymbol{j\ell}}$. Next, observe that by the triangle inequality and Assumption 9 *i)*, we have

$$\begin{aligned}\int \left|\left|\hat{\Delta}_\ell(w)\right|\right|^2 F_0(dw) &= \int \left|\left|\sum_{j=1}^{J}\left(m_j\left(y,\theta_0,\hat{\eta}_{j\ell}\right) - m_j\left(y,\theta_0,\eta_{0j}\right)\right)\left(\hat{\boldsymbol{\kappa}}_{\boldsymbol{j\ell}}(\boldsymbol{Z_j}) - \boldsymbol{\kappa}_{\boldsymbol{0j}}(\boldsymbol{Z_j})\right)\right|\right|^2 F_0(dw) \\ &\leq \sum_{j=1}^{J}\int \left|m_j\left(y,\theta_0,\hat{\eta}_{j\ell}\right) - m_j\left(y,\theta_0,\eta_{0j}\right)\right|^2 \left|\left|\hat{\boldsymbol{\kappa}}_{\boldsymbol{j\ell}}(\boldsymbol{Z_j}) - \boldsymbol{\kappa}_{\boldsymbol{0j}}(\boldsymbol{Z_j})\right|\right|^2 F_0(dw) \\ &\leq O_p(1)\sum_{j=1}^{J}\int \left|m_j\left(y,\theta_0,\hat{\eta}_{j\ell}\right) - m_j\left(y,\theta_0,\eta_{0j}\right)\right|^2 F_0(dw) \\ &\xrightarrow{p} 0.\end{aligned}$$

Second, let us show *ii).* By the Cauchy-Schwarz inequality,

$$\begin{aligned}\left|\left|\sqrt{n}\int \hat{\Delta}_{\ell}(w)F_0(dw)\right|\right| \leq \sqrt{n}&\left|\left(\int\sum_{j=1}^{J}\mathbb{E}\left[m_j\left(y,\theta_0,\hat{\eta}_{j\ell}\right)-m_j\left(y,\theta_0,\eta_{0j}\right)\right| Z_j,\mathcal{W}_{\ell}^{c}\right]^2 F_0(dZ_j)\right)^{1/2}\right| \\ &\left|\left|\left(\int\sum_{j=1}^{J}\left(\boldsymbol{\hat{\kappa}_{j\ell}(Z_j)}-\boldsymbol{\kappa_{0j}(Z_j)}\right)^2 F_0(dZ_j)\right)^{1/2}\right|\right|\end{aligned} \tag{E.1}$$

By Assumption 12, with probability approaching one,

$$\begin{aligned}\left|\left(\int\sum_{j=1}^{J}\mathbb{E}\left[m_j\left(y,\theta_0,\hat{\eta}_{j\ell}\right)-m_j\left(y,\theta_0,\eta_{0j}\right)\right| Z_j,\mathcal{W}_{\ell}^{c}\right]^2 F_0(dZ_j)\right)^{1/2}\right| &\leq C\left|\left|T\left(\eta-\eta_0\right)\right|\right|_{L^2(Z)} \\ &= O_p\left(\mu_n^{\eta}\right).\end{aligned} \tag{E.2}$$

Using (E.2) and Theorem 2 in (E.1) yields with probability approaching one,

$$\left|\left|\sqrt{n}\int \hat{\Delta}_{\ell}(w)F_0(dw)\right|\right| = O_p\left(\sqrt{n}\mu_n^{\eta}\mu_n^{\kappa}\right)\to 0,$$

by Assumption 11 *ii).* Then, the conclusion follows by the Conditional Markov inequality. ■

Lemma 17. *Let Assumption 13 hold. Then, there is a* $C>0$ *such that for* $\left|\left|T\left(\eta-\eta_0\right)\right|\right|_{L^2(Z)}$ *small enough,*

$$\left|\left|\overline{\psi}\left(\theta_0,\eta,\boldsymbol{\kappa_0}\right)\right|\right| \leq C\left|\left|T\left(\eta-\eta_0\right)\right|\right|_{L^2(Z)}^2.$$

Proof of Lemma 17: The result follows from Proposition 7.3.3 of Luenberger (1997). ■

To prove Lemma 3, let $\boldsymbol{\imath_q}$ be a $q-$dimensional vector of ones and define

$$\begin{aligned}g\left(W_i,\theta,\eta\right) &= \sum_{j=1}^{J} m_j\left(Y_i,\theta,\eta\right)\boldsymbol{\imath_q}, \\ \phi\left(W_i,\theta,\eta,\boldsymbol{\kappa}\right) &= \sum_{j=1}^{J} m_j\left(Y_i,\theta,\eta\right)\left(\boldsymbol{\kappa_j}\left(\boldsymbol{Z_{ji}}\right)-\boldsymbol{\imath_q}\right).\end{aligned}$$

Then, we can write

$$\psi\left(W_i,\theta,\eta,\boldsymbol{\kappa}\right) = g\left(W_i,\theta,\eta\right)+\phi\left(W_i,\theta,\eta,\boldsymbol{\kappa}\right). \tag{E.3}$$

Notice that, by using representation (E.3), we have written the LR functions as the sum of two terms $g+\phi$ as in Chernozhukov et al. (2022a) (Equation (2.3)). Two differences are worth mentioning. First, instead of having the Riesz representer entering in ϕ, we have the OR-IVs. Second, g and ϕ, by construction, are evaluated at the same θ. In particular, $\psi\left(W_i,\theta_0,\eta,\boldsymbol{\kappa}\right)=g\left(W_i,\theta_0,\eta\right)+\phi\left(W_i,\theta_0,\eta,\boldsymbol{\kappa}\right)$. Chernozhukov et al. (2022a) allow for g and ϕ to be evaluated at different θ's as ϕ has mean zero when evaluated at the true nuisance parameters value, for any $\theta\in\Theta$. In our case, that is true only at $\theta=\theta_0$.

Proof of Lemma 3: To show the result we will verify the conditions of Lemma 8 of Chernozhukov et al. (2022a) and restrict g and ϕ to be always evaluated at the same θ.

First, note that, by Assumption 9, $\mathbb{E}\left[||\psi\left(W,\theta_0,\eta_0,\boldsymbol{\kappa_0}\right)||^2\right] < \infty$. Moreover, by the triangle inequality,

$$\int ||g\left(y,\theta_0,\hat{\eta}_\ell\right) - g\left(y,\theta_0,\eta_0\right)||\, F_0(dw) \leq ||\boldsymbol{\imath_q}||^2 \sum_{j=1}^{J} \int \left|m_j\left(y,\theta_0,\hat{\eta}_{j\ell}\right) - m_j\left(y,\theta_0,\eta_{0j}\right)\right|^2 F_0(dw). \quad \text{(E.4)}$$

Hence, (E.4) and Assumption 9 *i)* imply Assumption 1 *(i)* of Chernozhukov et al. (2022a). Similarly, by the triangle inequality and Assumption 9 *(i)*,

$$\int ||\phi\left(w,\theta_0,\hat{\eta}_\ell,\boldsymbol{\kappa_0}\right) - \phi\left(w,\theta_0,\eta_0,\boldsymbol{\kappa_0}\right)||^2 F_0(dw) \leq 2\sum_{j=1}^{J} \int ||m_j\left(y,\theta_0,\hat{\eta}_{j\ell}\right)\boldsymbol{\kappa_{0j}(z_j)} - m_j\left(y,\theta_0,\eta_{0j}\right)\boldsymbol{\kappa_{0j}(z_j)}||^2 + o_p(1). \quad \text{(E.5)}$$

Then, Assumption 9 *ii)* imply Assumption 1 *(ii)* of Chernozhukov et al. (2022a). By the triangle inequality, we can show

$$\int ||\phi\left(w,\theta_0,\eta_0,\boldsymbol{\hat{\kappa}_\ell}\right) - \phi\left(w,\theta_0,\eta_0,\boldsymbol{\kappa_0}\right)||^2 F_0(dw) \leq \sum_{j=1}^{J} \int \left|m_j\left(y,\theta_0,\eta_{0j}\right)\right|^2 ||\boldsymbol{\hat{\kappa}_{j\ell}(z_j)} - \boldsymbol{\kappa_0(z_j)}||^2 F_0(dw). \quad \text{(E.6)}$$

Therefore, Assumption 9 *iii)* imply Assumption 1 *(iii)* of Chernozhukov et al. (2022a).[29]

Observe that

$$\hat{\Delta}_\ell(w) = \phi\left(w,\theta_0,\hat{\eta}_\ell,\boldsymbol{\hat{\kappa}_\ell}\right) - \phi\left(w,\theta_0,\eta_0,\boldsymbol{\hat{\kappa}_\ell}\right) - \phi\left(w,\theta_0,\hat{\eta}_\ell,\boldsymbol{\kappa_0}\right) + \phi\left(w,\theta_0,\eta_0,\boldsymbol{\kappa_0}\right).$$

Then, Lemma 16 implies Assumption 2 *(i)* of Chernozhukov et al. (2022a).

By Lemma 15 and Assumption 10, we have $\int \phi\left(w,\theta_0,\eta_0,\boldsymbol{\hat{\kappa}_\ell}\right) F_0(dw) = 0$ with probability approaching one. Furthermore, $\left|\left|\bar{\psi}\left(\theta_0,\hat{\eta}_\ell,\boldsymbol{\kappa_0}\right)\right|\right| \leq C\,||T\left(\hat{\eta}_\ell - \eta_0\right)||^2_{L^2(Z)}$, with probability approaching one. Hence, $\sqrt{n}\left|\left|\bar{\psi}\left(\theta_0,\hat{\eta}_\ell,\boldsymbol{\kappa_0}\right)\right|\right| \xrightarrow{p} 0$, by Assumption 11 *i)*. These results verify Assumption 3 *(i)* and *(iv)* of Chernozhukov et al. (2022a). Then, all the conditions of Lemma 8 of Chernozhukov et al. (2022a) hold in our context and the result of Lemma 3 can be obtained. ■

Let $\Psi = \mathbb{E}\left[\psi\left(W,\theta_0,\eta_0,\boldsymbol{\kappa_0}\right)\psi\left(W,\theta_0,\eta_0,\boldsymbol{\kappa_0}\right)'\right]$. We next show that

Lemma 18. *Let Assumptions 9 and 14 hold. Then,* $\hat{\Psi} \xrightarrow{p} \Psi$.

Proof of Lemma 18: This proof follows similarly to the proof of Lemma E1 of Chernozhukov et al.

[29] Assumption 1 *(iii)* of Chernozhukov et al. (2022a) is a convergence condition for estimators $\hat{\theta}_\ell$ and $\boldsymbol{\hat{\kappa}_\ell}$, but since we are restricting g and ϕ to be evaluated at $\theta = \theta_0$, we only need a convergence condition for $\boldsymbol{\kappa_0}$ and have $\hat{\theta}_\ell = \theta_0$.

(2022a). For each $i \in I_\ell$, let $\hat{\Delta}_\ell\left(W_i\right)$ be as in the main text. Additionally, let

$$\hat{R}_{1\ell i} = g\left(W_i, \theta_0, \hat{\eta}_\ell\right) - g\left(W_i, \theta_0, \eta_0\right) = \sum_{j=1}^{J}\left(m_j\left(Y_i, \theta_0, \hat{\eta}_{j\ell}\right) - m_j\left(Y_i, \theta_0, \eta_{0j}\right)\right)\boldsymbol{\imath_q},$$

$$\hat{R}_{2\ell i} = \phi\left(W_i, \theta_0, \hat{\eta}_\ell, \boldsymbol{\kappa_0}\right) - \phi\left(W_i, \theta_0, \eta_0, \boldsymbol{\kappa_0}\right) = \sum_{j=1}^{J}\left(m_j\left(Y_i, \theta_0, \hat{\eta}_\ell\right) - m_j\left(Y_i, \theta_0, \eta_0\right)\right)\left(\boldsymbol{\kappa_{0j}}\left(\boldsymbol{Z_{ji}}\right) - \boldsymbol{\imath_q}\right),$$

$$\hat{R}_{3\ell i} = \phi\left(W_i, \theta_0, \eta_0, \hat{\boldsymbol{\kappa}}_{\boldsymbol{\ell}}\right) - \phi\left(W_i, \theta_0, \eta_0, \boldsymbol{\kappa_0}\right) = \sum_{j=1}^{J} m_j\left(Y_i, \theta_0, \eta_0\right)\left(\hat{\boldsymbol{\kappa}}_{\boldsymbol{j\ell}}\left(\boldsymbol{Z_{ji}}\right) - \boldsymbol{\kappa_{0j}}\left(\boldsymbol{Z_{ji}}\right)\right),$$

$$\hat{R}_{4\ell i} = \sum_{j=1}^{J}\left(m_j\left(Y_i, \tilde{\theta}_\ell, \hat{\eta}_{j\ell}\right) - m_j\left(Y_i, \theta_0, \hat{\eta}_{j\ell}\right)\right)\hat{\boldsymbol{\kappa}}_{\boldsymbol{j\ell}}\left(\boldsymbol{Z_{ji}}\right).$$

By Assumption 9, $\mathbb{E}\left[\left|\left|\hat{R}_{k\ell i}\right|\right|^2 \middle| \mathcal{W}_\ell^c\right] \xrightarrow{p} 0$, $k = 1, 2, 3$. Similarly, by Assumption 14, $\mathbb{E}\left[\left|\left|\hat{R}_{4\ell i}\right|\right|^2 \middle| \mathcal{W}_\ell^c\right] \xrightarrow{p} 0$. Also, by Lemma 16 *i)* $\mathbb{E}\left[\left|\left|\hat{\Delta}_\ell(W)\right|\right|^2 \middle| \mathcal{W}_\ell^c\right] \xrightarrow{p} 0$. Then, it follows for $\psi_i = \psi\left(W_i, \theta_0, \eta_0, \boldsymbol{\kappa_0}\right)$,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i \in I_\ell}\left|\left|\hat{\psi}_{i\ell} - \psi_i\right|\right|^2 \middle| \mathcal{W}_\ell^c\right] &\leq \mathbb{E}\left[\left|\left|\hat{\psi}_{i\ell} - \psi_i\right|\right|^2 \middle| \mathcal{W}_\ell^c\right] \\
&\leq C\left(\sum_{k=1}^{4}\mathbb{E}\left[\left|\left|\hat{R}_{k\ell i}\right|\right|^2 \middle| \mathcal{W}_\ell^c\right] + \mathbb{E}\left[\left|\left|\hat{\Delta}_\ell(W_i)\right|\right|^2 \middle| \mathcal{W}_\ell^c\right]\right) \\
&\xrightarrow{p} 0.
\end{aligned}$$

Hence, by the Conditional Markov inequality $\frac{1}{n}\sum_{i \in I_\ell}\left|\left|\hat{\psi}_{i\ell} - \psi_i\right|\right|^2 \xrightarrow{p} 0$. Let $\tilde{\Psi} = \frac{1}{n}\sum_{i=1}^{n}\psi_i\psi_i'$. Then, by the triangle inequality and the Cauchy-Schwarz inequality,

$$\begin{aligned}
\left|\left|\hat{\Psi} - \tilde{\Psi}\right|\right| &\leq \sum_{\ell=1}^{L}\frac{1}{n}\sum_{i \in I_\ell}\left(\left|\left|\hat{\psi}_{i\ell} - \psi_i\right|\right|^2 + 2\left|\left|\psi_i\right|\right|\left|\left|\hat{\psi}_{i\ell} - \psi_i\right|\right|\right) \\
&\leq \underbrace{\sum_{\ell=1}^{L}\frac{1}{n}\sum_{i \in I_\ell}\left|\left|\hat{\psi}_{i\ell} - \psi_i\right|\right|^2}_{o_p(1)} + 2\underbrace{\sum_{\ell=1}^{L}\left(\frac{1}{n}\sum_{i \in I_\ell}\left|\left|\psi_i\right|\right|^2\right)^{1/2}}_{O_p(1)}\underbrace{\left(\frac{1}{n}\sum_{i \in I_\ell}\left|\left|\hat{\psi}_{i\ell} - \psi_i\right|\right|^2\right)^{1/2}}_{o_p(1)} \\
&= o_p(1)\left(O_p(1) + 1\right) \xrightarrow{p} 0.
\end{aligned}$$

Moreover, by the law of large numbers, $\tilde{\Psi} \xrightarrow{p} \Psi$. Hence, the conclusion of the lemma follows by the triangle inequality. ■

Lemma 19. *Let Assumption 15 hold and* $\bar{\theta} \xrightarrow{p} \theta_0$. *Then,* $\frac{\partial\hat{\psi}(\bar{\theta})}{\partial\theta} \xrightarrow{p} \Upsilon$.

Proof of Lemma 19: We follow the proof of Lemma E2 of Chernozhukov et al. (2022a). Let

$\hat{\Upsilon}_\ell = \frac{1}{n_\ell}\sum_{i\in I_\ell}\frac{\partial\psi(W_i,\bar{\theta},\hat{\eta}_\ell,\hat{\boldsymbol{\kappa}}_{\boldsymbol{\ell}})}{\partial\theta}$ and $\tilde{\Upsilon}_\ell = \frac{1}{n_\ell}\sum_{i\in I_\ell}\frac{\partial\psi(W_i,\theta_0,\hat{\eta}_\ell,\hat{\boldsymbol{\kappa}}_{\boldsymbol{\ell}})}{\partial\theta}$. Notice that by Assumption 15 *ii)*,

$$\mathbb{E}\left[\frac{1}{n_\ell}\sum_{i\in I_\ell} d\left(W_i,\hat{\eta}_\ell,\hat{\boldsymbol{\kappa}}_{\boldsymbol{\ell}}\right)\middle|\,\mathcal{W}_\ell^c\right] = \mathbb{E}\left[d\left(W_i,\hat{\eta}_\ell,\hat{\boldsymbol{\kappa}}_{\boldsymbol{\ell}}\right)\middle|\,\mathcal{W}_\ell^c\right] < C,$$

with probability approaching one. Then, by the Conditional Markov inequality, $\frac{1}{n_\ell}\sum_{i\in I_\ell} d\left(W_i,\hat{\eta}_\ell,\hat{\boldsymbol{\kappa}}_{\boldsymbol{\ell}}\right) = O_p(1)$. Next, by Assumption 15 *i)* and *ii)*, and the triangle inequality, with probability approaching one,

$$\begin{aligned}\left|\left|\hat{\Upsilon}_\ell - \tilde{\Upsilon}_\ell\right|\right| &\leq \frac{1}{n_\ell}\sum_{i\in I_\ell} d\left(W_i,\hat{\eta}_\ell,\hat{\boldsymbol{\kappa}}_{\boldsymbol{\ell}}\right)\left|\left|\bar{\theta}-\theta_0\right|\right|^{1/C}\\ &= O_p(1)o_p(1)\xrightarrow{p} 0.\end{aligned}$$

Then, $\hat{\Upsilon}_\ell - \tilde{\Upsilon}_\ell \xrightarrow{p} 0$ follows by the Conditional Markov inequality. Finally, let $\bar{\Upsilon}_\ell = \frac{1}{n_\ell}\sum_{i\in I_\ell}\frac{\partial\psi(W_i,\theta_0,\eta_0,\boldsymbol{\kappa}_{\mathbf{0}})}{\partial\theta}$. Similarly then, using Assumption 15 *iii)*, we have that $\tilde{\Upsilon}_\ell - \bar{\Upsilon}_\ell \xrightarrow{p} 0$. What is more, by the law of large numbers, $\bar{\Upsilon}_\ell \xrightarrow{p} \Upsilon$. Hence, the conclusion follows by the triangle inequality. ■

Proof of Theorem 4: Based on the results of Lemmas 3-19, the proof can be derived using standard asymptotic arguments as in, e.g., the proof of Proposition 21.20 in Ruud (2000). ■

The result in Theorem 4 relies on the consistency of $\hat{\theta}$. We now proceed by establishing the consistency of our estimator.

Theorem 20. *If i)* $\hat{\Lambda} \xrightarrow{p} \Lambda$*, where* Λ *is a positive definite matrix; ii)* $\mathbb{E}\left[\psi\left(W,\theta,\eta_0,\boldsymbol{\kappa}_0\right)\right] = 0$ *if and only if* $\theta = \theta_0$*; iii)* Θ *is compact; iv)* $\int ||m_j\left(y,\theta,\hat{\eta}_{j\ell}\right)\hat{\boldsymbol{\kappa}}_{\boldsymbol{j\ell}}\boldsymbol{(z_j)} - m_j\left(y,\theta,\eta_{0j}\right)\boldsymbol{\kappa}_{\mathbf{0}\boldsymbol{j}}\boldsymbol{(z_j)}||\, F_0(dw) \xrightarrow{p} 0$ *and* $\mathbb{E}\left[||m_j\left(Y,\theta,\eta_0\right)\boldsymbol{\kappa}_{\mathbf{0}\boldsymbol{j}}\boldsymbol{(Z_j)}||\right] < \infty$ *for all* $\theta\in\Theta$*; v) There is a* $C > 0$ *and* $d\left(W,\eta,\boldsymbol{\kappa}\right)$ *such that for each* $||T\left(\eta-\eta_0\right)||_{L^2(Z)}\,||\kappa-\kappa_0||_{L^2(Z)}$ *small enough and all* $\tilde{\theta}$*,* $\theta\in\Theta$*,*

$$\left|\left|\psi\left(W,\tilde{\theta},\eta,\boldsymbol{\kappa}\right) - \psi\left(W,\theta,\eta,\boldsymbol{\kappa}\right)\right|\right| \leq d\left(W,\eta,\boldsymbol{\kappa}\right)\left|\left|\tilde{\theta}-\theta\right|\right|^{1/C},\quad \mathbb{E}\left[d\left(W,\eta,\boldsymbol{\kappa}\right)\right] < C.$$

Then, $\hat{\theta}\xrightarrow{p}\theta$*.*

Proof of Theorem 20: We follow the proof of Theorem A3 of Chernozhukov et al. (2022a). Observe that, by the triangle inequality and Assumption *iv)*,

$$\begin{aligned}\int ||\psi\left(w,\theta,\hat{\eta}_\ell,\hat{\boldsymbol{\kappa}}_{\boldsymbol{\ell}}\right) - \psi\left(w,\theta,\eta_0,\boldsymbol{\kappa}_{\mathbf{0}}\right)||\,F_0(dw) &\leq \sum_{j=1}^{J}\int ||m_j\left(y,\theta,\hat{\eta}_{j\ell}\right)\hat{\boldsymbol{\kappa}}_{\boldsymbol{j\ell}}\boldsymbol{(z_j)} - m_j\left(y,\theta,\eta_{0j}\right)\boldsymbol{\kappa}_{\mathbf{0}\boldsymbol{j}}\boldsymbol{(z_j)}||\,F_0(dw)\\ &\xrightarrow{p} 0.\end{aligned}$$

It follows then that $\hat{\psi}(\theta)\xrightarrow{p}\bar{\psi}(\theta) = \mathbb{E}\left[\psi\left(W,\theta,\eta_0,\boldsymbol{\kappa}_{\mathbf{0}}\right)\right]$ for all $\theta\in\Theta$. Next, by *v)*, with probability approaching one,

$$
\begin{aligned}
\left|\left|\hat{\psi}\left(\tilde{\theta}\right)-\hat{\psi}\left(\theta\right)\right|\right| &\leq \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\left|\left|\psi\left(W_i,\tilde{\theta},\hat{\eta}_\ell,\hat{\boldsymbol{\kappa}}_{\boldsymbol{\ell}}\right)-\psi\left(W_i,\theta,\hat{\eta}_\ell,\hat{\boldsymbol{\kappa}}_{\boldsymbol{\ell}}\right)\right|\right| \\
&\leq \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell} d\left(W_i,\hat{\eta}_\ell,\hat{\boldsymbol{\kappa}}_{\boldsymbol{\ell}}\right)\left|\left|\tilde{\theta}-\theta\right|\right|^{1/C} \\
&= \hat{M}\left|\left|\tilde{\theta}-\theta\right|\right|^{1/C}.
\end{aligned}
$$

Note, by the conditional Markov inequality, $\hat{M}=O_p(1)$. Then, by Corollary 2.2 of Newey (1991), we have $\sup_{\theta\in\Theta}\left|\left|\hat{\psi}(\theta)-\bar{\psi}(\theta)\right|\right|\xrightarrow{p}0$. Moreover, observe that condition *v)* also implies that $\bar{\psi}(\theta)$ is continuous on Θ. Finally, note that the second condition in *iv)* implies that $\mathbb{E}\left[||\psi\left(W,\theta,\eta_0,\boldsymbol{\kappa}_{\boldsymbol{0}}\right)||\right]<\infty$ for all $\theta\in\Theta$, by the triangle inequality. The conclusion then follows similarly to the proof of Theorem 2.6 of Newey and McFadden (1994). ■

F Additional Monte Carlo Details

As we stated in the main text, to obtain our estimator $\hat{\theta}=\left(\hat{\theta}_1,\hat{\theta}_k,\hat{\theta}_\omega\right)'$, we use GMM based on four debiased moments. These can be written as

$$
\begin{aligned}
\psi\left(W,\theta_0,\eta_0\right) = &\left(Y_1-\eta_{01}\left(I_1,K_1\right)\right)\boldsymbol{\kappa_{01}}\left(\boldsymbol{Z_1}\right)+\left(Y_2-\theta_{01}-\theta_{0k}K_2-\theta_{0\omega}\left(\eta_{01}\left(Z_1\right)-\theta_{01}-\theta_{0k}K_1\right)\right)\boldsymbol{\kappa_{02}}\left(\boldsymbol{Z_1}\right) \\
&+\left(Y_2-\eta_{02}\left(I_2,K_2\right)\right)\boldsymbol{\kappa_{03}}\left(\boldsymbol{Z_2}\right)+\left(Y_3-\theta_{01}-\theta_{0k}K_3-\theta_{0\omega}\left(\eta_{02}\left(Z_2\right)-\theta_{01}-\theta_{0k}K_2\right)\right)\boldsymbol{\kappa_{04}}\left(\boldsymbol{Z_2}\right).
\end{aligned}
$$

Notice that our GMM program involves a three-dimensional non-linear search. To increase the reliability of our results, we have reduced the dimension of the problem such that we see θ_{01} and $\theta_{0\omega}$ as functions of θ_{0k}. In this way, we only search over the dimension θ_{0k}. We have accomplished this as follows. Notice

$$\eta_{0t}\left(Z_t\right)=\theta_{01}+\theta_{0k}K_t+\omega_t\left(I_t,K_t\right),$$

which implies that

$$\theta_{01}+\omega_t\left(I_t,K_t\right)=\eta_{0t}\left(Z_t\right)-\theta_{0k}K_t. \tag{F.1}$$

As ω_t follows an AR(1) process, we have

$$\omega_t=\theta_{0\omega}\omega_{t-1}+\epsilon_t^\omega,\qquad \mathbb{E}\left[\epsilon_t^\omega\middle|\,\omega_{t-1}\right]=0. \tag{F.2}$$

Plugging (F.1) into (F.2) and re-arranging terms yields

$$\eta_{0t}\left(Z_t\right)-\theta_{0k}K_t=\tilde{c}+\theta_{0\omega}\left(\eta_{0,t-1}\left(Z_{t-1}\right)-\theta_{0k}K_{t-1}\right)+\epsilon_t^\omega,\quad \tilde{c}=\theta_{01}\left(1-\theta_{0\omega}\right).$$

Hence, for a given value of θ_{0k}, we can identify $\theta_{0\omega}$ as the slope in a linear regression of $\eta_{0t}-\theta_{0k}K_t$ on $\eta_{0,t-1}-\theta_{0k}K_{t-1}$. The parameter θ_{01} can also be identified from this regression equation by using the equality $\theta_{01}=\tilde{c}/(1-\theta_{0\omega})$, provided that $\theta_{0\omega}\neq 1$. As $\theta_{01}=0$ in our Monte Carlo experiments, we directly consider $\tilde{c}=\theta_{01}$. Then, in our non-linear search, we impose these restrictions and minimize the GMM objective function based on ψ, treating it as a function of θ_{0k} only.

References

Ackerberg, Daniel, C Lanier Benkard, Steven Berry, and Ariel Pakes (2007): "Econometric Tools for Analyzing Market Outcomes," *Handbook of Econometrics*, 6, 4171–4276.

Ackerberg, Daniel, Xiaohong Chen, Jinyong Hahn, and Zhipeng Liao (2014): "Asymptotic Efficiency of Semiparametric Two-Step GMM," *Review of Economic Studies*, 81 (3), 919–943.

Ackerberg, Daniel A, Kevin Caves, and Garth Frazer (2015): "Identification Properties of Recent Production Function Estimators," *Econometrica*, 83 (6), 2411–2451.

Ai, Chunrong and Xiaohong Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71 (6), 1795–1843.

——— (2012): "The Semiparametric Efficiency Bound for Models of Sequential Moment Restrictions Containing Unknown Functions," *Journal of Econometrics*, 170 (2), 442–457.

Andersen, Torben G and Bent E Sørensen (1996): "GMM Estimation of a Stochastic Volatility Model: A Monte Carlo Study," *Journal of Business & Economic Statistics*, 14 (3), 328–352.

Argañaraz, Facundo and Juan Carlos Escanciano (2023): "On the Existence and Information of Orthogonal Moments For Inference," *arXiv preprint arXiv:2303.11418*.

Athey, Susan, Guido W. Imbens, and Stefan Wager (2018): "Approximate Residual Balancing: Debiased Inference of Average Treatment Effects in High Dimensions," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80 (4), 597–623.

Bakhitov, Edvard (2022): "Automatic Debiased Machine Learning in Presence of Endogeneity," *Working Paper, https://edbakhitov. com/assets/pdf/jmp_edbakhitov.pdf.*

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen (2012): "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica*, 80 (6), 2369–2429.

Belloni, Alexandre and Victor Chernozhukov (2013): "Least Squares After Model Selection in High-Dimensional Sparse Models," *Bernoulli*, 19 (2), 521–547.

Belloni, Alexandre, Victor Chernozhukov, Iván Fernández-Val, and Christian Hansen (2017): "Program Evaluation and Causal Inference with High-Dimensional Data," *Econometrica*, 85 (1), 233–298.

Bennett, Andrew, Nathan Kallus, Xiaojie Mao, Whitney Newey, Vasilis Syrgkanis, and Masatoshi Uehara (2022): "Inference On Strongly Identified Functionals of Weakly Identified Functions," *arXiv preprint arXiv:2208.08291*.

Bickel, Peter J (1982): "On Adaptive Estimation," *The Annals of Statistics*, 647–671.

Bickel, Peter J., Ya'acov Ritov, and Alexandre B. Tsybakov (2009): "Simultaneous Analysis of Lasso and Dantzig Selector," *The Annals of Statistics*, 37 (4), 1705 – 1732.

Bierens, Herman J (1990): "A Consistent Conditional Moment Test of Functional Form," *Econometrica: Journal of the Econometric Society*, 1443–1458.

Bonhomme, Stéphane (2012): "Functional Differencing," *Econometrica*, 80 (4), 1337–1385.

Bradic, Jelena, Victor Chernozhukov, Whitney K Newey, and Yinchu Zhu (2022): "Minimax Semiparametric Learning with Approximate Sparsity," *arXiv preprint arXiv:1912.12213*.

Bravo, Francesco, Juan Carlos Escanciano, and Ingrid Van Keilegom (2020): "Two-Step Semiparametric Empirical Likelihood Inference," *The Annals of Statistics*, 48 (1), 1 – 26.

Brown, Bryan W and Whitney K Newey (1998): "Efficient Semiparametric Estimation of Expectations," *Econometrica*, 66 (2), 453–464.

Bühlmann, Peter and Sara Van De Geer (2011): *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Science & Business Media.

Carrasco, Marine and Jean-Pierre Florens (2000): "Generalization of GMM to a Continuum of Moment Conditions," *Econometric Theory*, 16 (6), 797–834.

Carrasco, Marine, Jean-Pierre Florens, and Eric Renault (2007): *Chapter 77 Linear Inverse Problems in Structural Econometrics Estimation Based On Spectral Decomposition and Regularization*, vol. 6 of *Handbook of Econometrics*, Elsevier.

Cha, Jooyoung, Harold D. Chiang, and Yuya Sasaki (2023): "Inference in High-Dimensional Regression Models without the Exact or Lp Sparsity," *The Review of Economics and Statistics*, 1–32.

Chamberlain, Gary (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of econometrics*, 34 (3), 305–334.

——— (1992a): "Comment: Sequential Moment Restrictions in Panel Data," *Journal of Business & Economic Statistics*, 10 (1), 20–26.

——— (1992b): "Efficiency Bounds for Semiparametric Regression," *Econometrica: Journal of the Econometric Society*, 567–596.

Chen, Xiaohong and Demian Pouzo (2009): "Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals," *Journal of Econometrics*, 152 (1), 46–60.

Chen, Xiaohong and Yin Jia Jeff Qiu (2016): "Methods for Nonparametric and Semiparametric Regressions with Endogeneity: A Gentle Guide," *Annual Review of Economics*, 8, 259–290.

Chen, Xiaohong and Halbert White (1999): "Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators," *IEEE Transactions on Information Theory*, 45 (2), 682–691.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins (2018): "Double/debiased Machine Learning for Treatment and Structural Parameters," *The Econometrics Journal*, 21, C1–C68.

Chernozhukov, Victor, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins (2022a): "Locally Robust Semiparametric Estimation," *Econometrica*, 90 (4), 1501–1535.

Chernozhukov, Victor, Whitney Newey, and James Robins (2022b): "De-Biased Machine Learning of Global and Local Parameters Using Regularized Riesz Representers," *arXiv preprint arXiv:1802.08667*.

Chernozhukov, Victor, Whitney Newey, and Rahul Singh (2022c): "De-Biased Machine Learning of Global and Local Parameters Using Regularized Riesz Representers," *arXiv preprint arXiv:1802.08667*.

Chernozhukov, Victor, Whitney Newey, Rahul Singh, and Vasilis Syrgkanis (2020): "Adversarial Estimation of Riesz Representers," *arXiv preprint arXiv:2101.00009*.

Chernozhukov, Victor, Whitney K. Newey, Victor Quintas-Martinez, and Vasilis Syrgkanis (2021): "Automatic Debiased Machine Learning Via Neural Nets for Generalized Linear Regression," *arXiv preprint arXiv:2104.14737*.

Chernozhukov, Victor, Whitney K Newey, and Rahul Singh (2022d): "Automatic Debiased Machine Learning of Causal and Structural Effects," *Econometrica*, 90 (3), 967–1027.

Dikkala, Nishanth, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis (2020): "Minimax Estimation of Conditional Moment Models," *Advances in Neural Information Processing Systems*, 33, 12248–12262.

Farrell, Max H (2015): "Robust Inference On Average Treatment Effects with Possibly More Covariates Than Observations," *Journal of Econometrics*, 189 (1), 1–23.

Farrell, Max H., Tengyuan Liang, and Sanjog Misra (2021a): "Deep Neural Networks for Estimation and Inference," *Econometrica*, 89 (1), 181–213.

——— (2021b): "Deep Learning for Individual Heterogeneity: An Automatic Inference Framework," *arXiv preprint 2010.14694*.

Friedman, Jerome, Trevor Hastie, Holger Höfling, and Robert Tibshirani (2007): "Pathwise Coordinate Optimization," *The Annals of Applied Statistics*, 1 (2), 302 – 332.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani (2010): "Regularization Paths for Generalized Linear Models Via Coordinate Descent," *Journal of Statistical Software*, 33 (1), 1.

FU, WENJIANG J (1998): "Penalized Regressions: the Bridge Versus the Lasso," *Journal of Computational and Graphical Statistics*, 7 (3), 397–416.

GANDHI, AMIT, SALVADOR NAVARRO, AND DAVID A RIVERS (2020): "On the Identification of Gross Output Production Functions," *Journal of Political Economy*, 128 (8), 2973–3016.

GOLD, DAVID, JOHANNES LEDERER, AND JING TAO (2020): "Inference for High-Dimensional Instrumental Variables Regression," *Journal of Econometrics*, 217 (1), 79–111.

GRAHAM, BRYAN S (2011): "Efficiency Bounds for Missing Data Models with Semiparametric Restrictions," *Econometrica*, 79 (2), 437–452.

HOTZ, V JOSEPH AND ROBERT A MILLER (1993): "Conditional Choice Probabilities and the Estimation of Dynamic Models," *The Review of Economic Studies*, 60 (3), 497–529.

HRISTACHE, MARIAN AND VALENTIN PATILEA (2016): "Semiparametric Efficiency Bounds for Conditional Moment Restriction Models with Different Conditioning Variables," *Econometric Theory*, 32 (4), 917–946.

——— (2017): "Conditional Moment Models with Data Missing At Random," *Biometrika*, 104 (3), 735–742.

ICHIMURA, HIDEHIKO AND WHITNEY K NEWEY (2022): "The Influence Function of Semiparametric Estimators," *Quantitative Economics*, 13 (1), 29–61.

KLAASSEN, CHRIS AJ (1987): "Consistent Estimation of the Influence Function of Locally Asymptotically Linear Estimators," *The Annals of Statistics*, 15 (4), 1548–1562.

LEVINSOHN, JAMES AND AMIL PETRIN (2003): "Estimating Production Functions Using Inputs to Control for Unobservables," *The Review of Economic Studies*, 70 (2), 317–341.

LUENBERGER, DAVID G (1997): *Optimization By Vector Space Methods*, John Wiley & Sons.

LUO, YE, MARTIN SPINDLER, AND JANNIS KÜCK (2022): "High-Dimensional L_2Boosting: Rate of Convergence," *arXiv preprint arXiv:1602.08927*.

MUANDET, KRIKAMOL, WITTAWAT JITKRITTUM, AND JONAS KÜBLER (2020): "Kernel Conditional Moment Test Via Maximum Moment Restriction," in *Conference on Uncertainty in Artificial Intelligence*, PMLR, 41–50.

NEKIPELOV, DENIS, VIRA SEMENOVA, AND VASILIS SYRGKANIS (2022): "Regularised Orthogonal Machine Learning for Nonlinear Semiparametric Models," *The Econometrics Journal*, 25 (1), 233–255.

NEWEY, WHITNEY K (1990): "Semiparametric Efficiency Bounds," *Journal of applied econometrics*, 5 (2), 99–135.

——— (1991): "Uniform Convergence in Probability and Stochastic Equicontinuity," *Econometrica: Journal of the Econometric Society*, 1161–1167.

——— (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica: Journal of the Econometric Society*, 1349–1382.

Newey, Whitney K and Daniel McFadden (1994): "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics*, 4, 2111–2245.

Newey, Whitney K and James L Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71 (5), 1565–1578.

Newey, Whitney K and Richard J Smith (2004): "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72 (1), 219–255.

Neyman, J. (1959): "Optimal Asymptotic Tests of Composite Statistical Hypothesis," *Probability and Statistics: The Harald Cramer Volume*, 213–234.

Olley, G. Steven and Ariel Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64 (6), 1263–1297.

Petrin, Amil, Brian P Poi, and James Levinsohn (2004): "Production Function Estimation in Stata Using Inputs to Control for Unobservables," *The Stata Journal*, 4 (2), 113–123.

Ramsay, James, Giles Hooker, and Spencer Graces (2009): *Functional Data Analysis with R and MATLAB*, Springer-Verlag New York.

Robins, James, Lingling Li, Eric Tchetgen, and Aad van der Vaart (2008): "Higher Order Influence Functions and Minimax Estimation of Nonlinear Functionals," *Probability and statistics: essays in honor of David A. Freedman*, 2, 335–421.

Ruud, P.A. (2000): *An Introduction to Classical Econometric Theory*, Oxford University Press.

Sasaki, Yuya and Takuya Ura (2023): "Estimation and Inference for Policy Relevant Treatment Effects," *Journal of Econometrics*, 234 (2), 394–450.

Schmidt-Hieber, Johannes (2020): "Nonparametric Regression Using Deep Neural Networks with ReLU Activation Function," *The Annals of Statistics*, 48 (4), 1875 – 1897.

Singh, Rahul, Maneesh Sahani, and Arthur Gretton (2019): "Kernel Instrumental Variable Regression," *Advances in Neural Information Processing Systems*, 32.

Syrgkanis, Vasilis and Manolis Zampetakis (2020): "Estimation and Inference with Trees and Forests in High Dimensions," *arXiv preprint arXiv:2007.03210*.

Tseng, Paul (2001): "Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization," *Journal of Optimization Theory and Applications*, 109, 475–494.

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

Vershynin, Roman (2018): *High-dimensional Probability: An Introduction with Applications in Data Science*, vol. 47, Cambridge University Press.

Wooldridge, Jeffrey M (2009): "On Estimating Firm-Level Production Functions Using Proxy Variables to Control for Unobservables," *Economics letters*, 104 (3), 112–114.

——— (2010): *Econometric Analysis of Cross Section and Panel Data*, MIT press.

Zhang, Rui, Krikamol Muandet, Bernhard Schölkopf, and Masaaki Imaizumi (2021): "Instrument Space Selection for Kernel Maximum Moment Restriction," *arXiv preprint arXiv:2106.03340*.

Zheng, Wenjing and Mark J. van der Laan (2010): "Asymptotic Theory for Cross-Validated Targeted Maximum Likelihood Estimation," .