Double machine learning for treatment and causal parameters

Victor Chernozhukov
Denis Chetverikov
Mert Demirer
Esther Duflo
Christian Hansen
Whitney Newey

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP49/16

Double Machine Learning for Treatment and Causal Parameters

by

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey

Abstract. Most modern supervised statistical/machine learning (ML) methods are explicitly designed to solve prediction problems very well. Achieving this goal does not imply that these methods automatically deliver good estimators of causal parameters. Examples of such parameters include individual regression coefficients, average treatment effects, average lifts, and demand or supply elasticities. In fact, estimators of such causal parameters obtained via naively plugging ML estimators into estimating equations for such parameters can behave very poorly. For example, the resulting estimators may formally have inferior rates of convergence with respect to the sample size n caused by regularization bias. Fortunately, this regularization bias can be removed by solving auxiliary prediction problems via ML tools. Specifically, we can form an efficient score for the target low-dimensional parameter by combining auxiliary and main ML predictions. The efficient score may then be used to build an efficient estimator of the target parameter which typically will converge at the fastest possible $1/\sqrt{n}$ rate and be approximately unbiased and normal, allowing simple construction of valid confidence intervals for parameters of interest. The resulting method thus could be called a "double ML" method because it relies on estimating primary and auxiliary predictive models. Such double ML estimators achieve the fastest rates of convergence and exhibit robust good behavior with respect to a broader class of probability distributions than naive "single" ML estimators. In order to avoid overfitting, following [3], our construction also makes use of the K-fold sample splitting, which we call *cross-fitting*. The use of sample splitting allows us to use a very broad set of ML predictive methods in solving the auxiliary and main prediction problems, such as random forests, lasso, ridge, deep neural nets, boosted trees, as well as various hybrids and aggregates of these methods (e.g. a hybrid of a random forest and lasso). We illustrate the application of the general theory through application to the leading cases of estimation and inference on the main parameter in a partially linear regression model and estimation and inference on average treatment effects and average treatment effects on the treated under conditional random assignment of the treatment. These applications cover randomized control trials as a special case. We then use the methods in an empirical application which estimates the effect of 401(k) eligibility on accumulated financial assets.

Key words: Neyman, orthogonalization, cross-fit, double machine learning, debiased machine learning, orthogonal score, efficient score, post-machine-learning and post-regularization inference, random forest, lasso, deep learning, neural nets, boosted trees, efficiency, optimality.

Date: July, 2016.

We would like to acknowledge the research support from the National Science Foundation. We also thank participants of the MIT Stochastics and Statistics seminar, the Kansas Econometrics conference, Royal Economic Society Annual Conference, The Hannan Lecture at the Australasian Econometric Society meeting, The CORE 50th Anniversary Conference, the World Congress of Probability and Statistics 2016, the Joint Statistical Meetings 2016, The New England Day of Statistics Conference, CEMMAP's Masterclass on Causal Machine Learning, and St. Gallen's summer school on "Big Data", for many useful comments and questions.

1. Introduction and Motivation

We develop a series of results for obtaining root-n consistent estimation and valid inferential statements about a low-dimensional parameter of interest, θ_0, in the presence of an infinite-dimensional nuisance parameter, η_0. The parameter of interest will typically be a causal parameter or treatment effect parameter, and we consider settings in which the nuisance parameter will be estimated using modern machine learning (ML) methods such as random forests, lasso or post-lasso, boosted regression trees, or their hybrids.

As a lead example consider the partially linear model

$$Y = D\theta_0 + g_0(Z) + U, \quad \mathrm{E}[U \mid Z, D] = 0, \tag{1.1}$$

$$D = m_0(Z) + V, \qquad \mathrm{E}[V \mid Z] = 0, \tag{1.2}$$

where Y is the outcome variable, D is the policy/treatment variable of interest,[1] Z is a vector of other covariates or "controls", and U and V are disturbances. The first equation is the main equation, and θ_0 is the target parameter that we would like to estimate. If D is randomly assigned conditional on controls Z, θ_0 has the interpretation of the treatment effect (TE) parameter or "lift" parameter in business applications. The second equation keeps track of confounding, namely the dependence of the treatment/policy variable on covariates/controls. This equation will be important for characterizing regularization bias, which is akin to omitted variable bias. The confounding factors Z affect the policy variable D via the function $m_0(Z)$ and the outcome variable via the function $g_0(Z)$.

A conventional ML approach to estimation of θ_0 would be, for example, to construct a sophisticated ML estimator for $D\widehat{\theta}_0 + \widehat{g}_0(Z)$ for learning the regression function $D\theta_0 + g_0(Z)$.[2] Suppose, for the sake of clarity, that $\widehat{g}_0$ is obtained using an auxiliary sample and that, given this $\widehat{g}_0$, the final estimate of θ_0 is obtained using the main sample of observations enumerated by $i = 1, ..., n$:

$$\widehat{\theta}_0 = \Big(\frac{1}{n}\sum_{i=1}^{n} D_i^2\Big)^{-1} \frac{1}{n}\sum_{i=1}^{n} D_i(Y_i - \widehat{g}_0(Z_i)). \tag{1.3}$$

The estimator $\widehat{\theta}_0$ will generally have an "inferior", slower than $1/\sqrt{n}$, rate of convergence; namely,

$$|\sqrt{n}(\widehat{\theta}_0 - \theta_0)| \to_P \infty. \tag{1.4}$$

As we explain this below, the driving force behind this "inferior" behavior is the bias in learning $g_0(Z)$. Figure 1 provides a numerical illustration of this phenomenon for a conventional ML estimator based on a random forest in a simple computational experiment.

[1] We consider the case where D is a scalar for simplicity; extension to the case where D is a vector of fixed, finite dimension is accomplished by introducing an equation like (1.2) for each element of the vector.

[2] For instance, we could use iterative methods that alternate between the random forest to find an estimator for g_0 and least squares to find an estimator for θ_0.

We can decompose the scaled estimation error as

$$\sqrt{n}(\widehat{\theta}_0 - \theta_0) = \underbrace{\Big(\frac{1}{n}\sum_{i=1}^{n} D_i^2\Big)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i U_i}_{:=a} + \underbrace{\Big(\frac{1}{n}\sum_{i=1}^{n} D_i^2\Big)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i(g_0(Z_i) - \widehat{g}_0(Z_i))}_{:=b}.$$

The first term is well-behaved under mild conditions, obeying

$$a \rightsquigarrow N(0, \bar{\Sigma}), \quad \bar{\Sigma} = (\mathrm{E}D^2)^{-1}\mathrm{E}U^2D^2(\mathrm{E}D^2)^{-1}.$$

The term b is not centered however and will in general diverge:

$$|b| \to_P \infty.$$

Heuristically, b is the sum of n non-centered terms divided by $\sqrt{n}$, where each term contains the estimation error $\widehat{g}_0(Z_i) - g_0(Z_i)$. Thus, we expect b to be of stochastic order $\sqrt{n}n^{-\varphi}$, where $n^{-\varphi}$ is the rate of convergence of $\widehat{g}$ to g_0 in the root mean squared error sense and where we will have $\varphi < 1/2$ in the nonparametric case. More precisely, we can use that we are using an auxiliary sample to estimate g_0 to approximate the b up to a vanishing term by

$$b' = (\mathrm{E}D^2)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_0(Z_i)B_g(Z_i),$$

where $B_g(z) = g_0(z) - \mathrm{E}\widehat{g}_0(z)$ is the bias for estimating $g_0(z)$ for z in the support of Z. In typical scenarios, the use of regularization forces the order of the squared bias to be balanced with the variance to achieve an optimal root-mean square rate. The rate can not be faster than $1/\sqrt{n}$ and will often be of order $n^{-\varphi}$ for $0 < \varphi < 1/2$.[3]These rates mean that generically the term b' diverges, $|b'| \to \infty$, since $\mathrm{E}[|b'|] \gtrsim \sqrt{n}n^{-\varphi} \to \infty$. The estimator $\widehat{\theta}_0$ will thus have an "inferior" rate of convergence, diverging when scaled by $\sqrt{n}$, as claimed in (1.4).

Now consider a second construction that employs an "orthogonalized" formulation obtained by directly partialing out the effect of Z from both Y and D. Specifically consider the regression model implied by the partially linear model (1.1)-(1.2):

$$W = V\theta_0 + U,$$

where $V = D - m_0(Z)$ and $W = Y - \ell_0(Z)$, where $\ell_0(Z) = \mathrm{E}[Y|Z] = m_0(Z)\theta_0 + g_0(Z)$. Regression functions ℓ_0 and m_0 can easily be directly estimated using supervised ML methods. Specifically, we can construct $\widehat{\ell}_0$ and $\widehat{m}_0$, using an auxiliary sample, use these estimators to form $\widehat{W} = Y - \widehat{\ell}_0(Z)$ and $\widehat{V} = V - \widehat{m}_0(Z)$, and then obtain an "orthogonalized" or "double ML" estimator

$$\check{\theta}_0 = \Big(\frac{1}{n}\sum_{i=1}^{n}\widehat{V}_i^2\Big)^{-1}\frac{1}{n}\sum_{i=1}^{N}\widehat{V}_i\widehat{W}_i. \tag{1.5}$$

[3] In some cases, it is possible to give up on the optimal rate of convergence for the estimator $\widehat{g}_0$ and use under-smoothing. That is, bias squared and variance are not balanced, and the order of the bias is set to $o(n^{-1/2})$ which makes b' negligible. However, under-smoothing is typically possible only in classical low-dimensional settings.

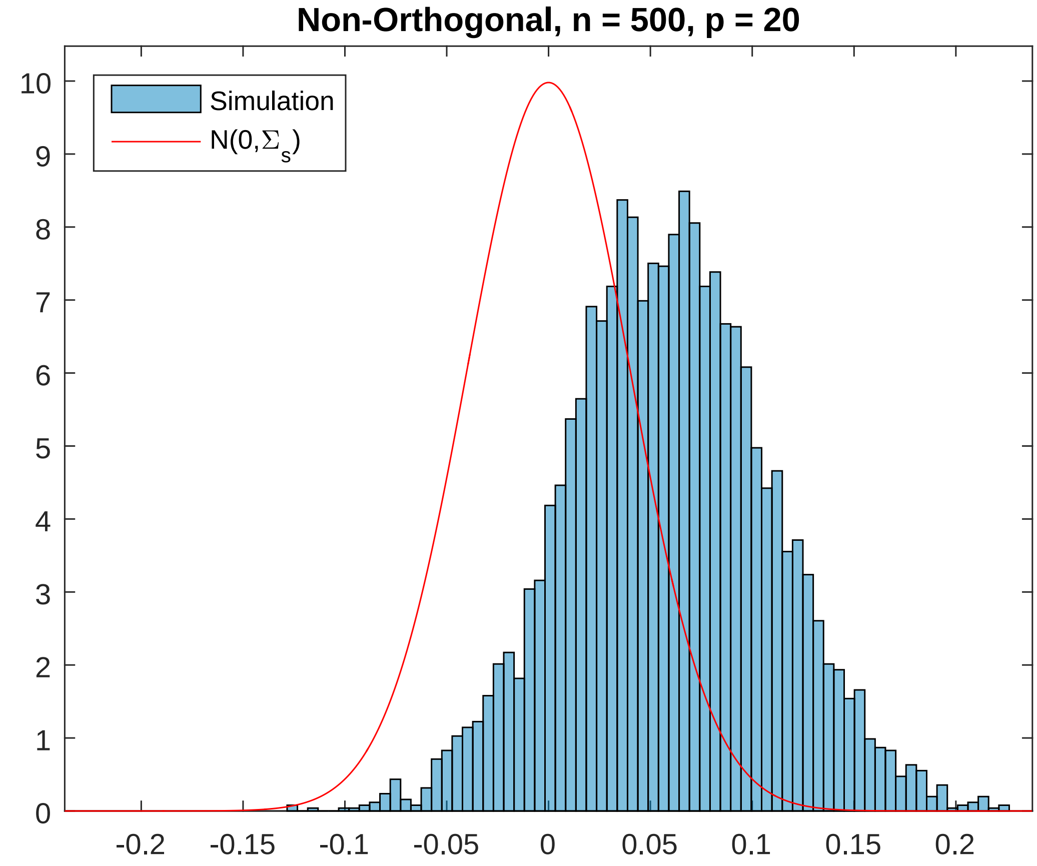

Figure 1. Behavior of a conventional (non-orthogonal) ML estimator $\widehat{\theta}_0$ in the partially linear model example in a simple simulation experiment where we learn g_0 using a random forest. The g_0 in this experiment is a very smooth function of a small number of variables, so the experiment is seemingly favorable to the use of random forests a priori. The histogram shows the simulated distribution of the centered estimator, $\widehat{\theta}_0 - \theta_0$. The estimator is badly biased, shifted much to the right relative to the true value θ_0. The distribution of the estimator (approximated by the blue histogram) is substantively different from a normal approximation (shown by the red curve) derived under the assumption that the bias is negligible.

In contrast to the previous estimator, this estimator will be root-n consistent and approximately Gaussian under a very mild set of conditions:

$$\sqrt{n}(\check{\theta}_0 - \theta_0) \rightsquigarrow N(0,\Sigma), \quad \Sigma = (\mathrm{E}V^2)^{-1}\mathrm{E}U^2V^2(\mathrm{E}V^2)^{-1}.$$

$\sqrt{n}$ asymptotic normality means that the estimator $\check{\theta}_0$ provides both superior point estimation properties relative to the naive $\widehat{\theta}_0$ and can be used in the construction of valid confidence intervals. Figure 2 shows the (relatively) good properties of the double ML estimator in a simple computational experiment. Here we use random forests to learn function ℓ_0 and m_0.

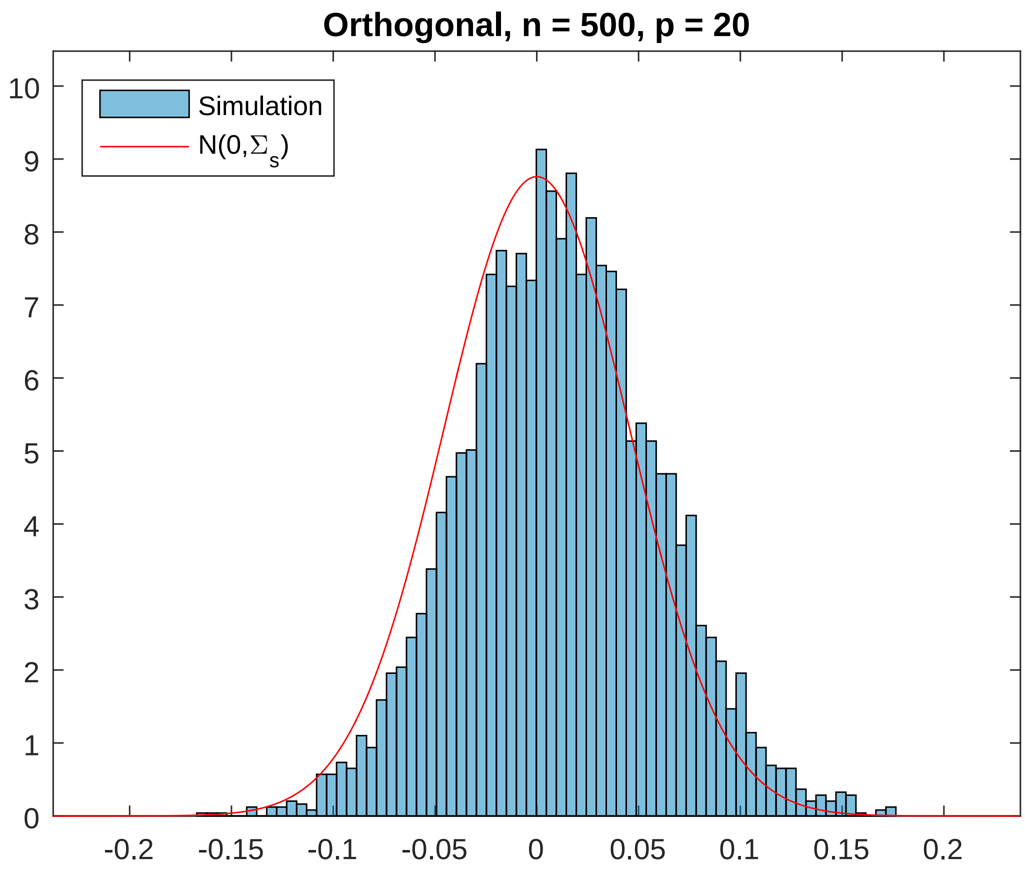

Figure 2. Behavior of the orthogonal, double ML estimator $\check{\theta}_0$ in the partially linear model example in a simple experiment where we learn ℓ_0 and m_0 using random forests. Note that the simulated data are exactly the same as those underlying Figure 1. The histogram shows the simulated distribution of the centered estimator, $\widehat{\theta}_0 - \theta_0$. The estimator is approximately unbiased, concentrates around θ_0, and is approximately normal.

Let us provide the explanation. We can decompose the scaled estimation error into three components:

$$\sqrt{n}(\check{\theta}_0 - \theta_0) = a^* + b^* + c^*.$$

First, the leading term a^* is approximately Gaussian under mild conditions:

$$a^* = \Big(\frac{1}{n}\sum_{i=1}^n V_i^2\Big)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n V_i U_i \rightsquigarrow N(0, \Sigma);$$

Second, the term

$$b^* = (\mathrm{E}D^2)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n (\widehat{m}_0(Z_i) - m_0(Z_i))(\widehat{\ell}_0(Z_i) - \ell_0(Z_i)),$$

now depends on the product of estimation errors, and so can vanish under a broad range of data-generating processes; for example, when $\widehat{m}_0$ and $\widehat{\ell}_0$ are consistent for m_0 and ℓ_0 at the $o(n^{-1/4})$ rate. Indeed, heuristically this term can be upper-bounded by $\sqrt{n}n^{-(\varphi_m+\varphi_\ell)}$, where $n^{-\varphi_m}$ and $n^{-\varphi_\ell}$ are the rates of convergence of $\widehat{m}_0$ to m_0 and $\widehat{\ell}_0$ to ℓ_0. It is often possible to have $\varphi_m + \varphi_l > 1/2$, for example, it suffices to have $\varphi_m > 1/4$ and $\varphi_\ell > 1/4$, as mentioned above. A more exact analysis is possible when we use different parts of the auxiliary sample to estimate m_0 and ℓ_0. In this case b^* can be approximated by

$$(\mathrm{E}D^2)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n B_g(Z_i)B_\ell(Z_i),$$

where $B_\ell(z) = \ell_0(z) - \mathrm{E}\widehat{\ell}_0(z)$ is the bias for estimating $\ell_0(z)$ and $B_m(z) = m_0(z) - \mathrm{E}\widehat{m}_0(z)$ is the bias for estimating $m(z)$. If the $L^2(P)$ norms of the biases are of order $o(n^{-1/4})$, which is an attainable rate in a wide variety of cases, then

$$\mathrm{E}|c^*| \leqslant \sqrt{n}\sqrt{\mathrm{E}B_m^2(Z)\mathrm{E}B_\ell^2(Z)} \leqslant \sqrt{n}o(n^{-1/2}) \to 0.$$

Third, the term c^* is the remainder term, which obeys

$$c^* = o_P(1)$$

and sample splitting plays a key role in driving this term to zero. Indeed, c^* contains expressions like

$$\Big(\frac{1}{n}\sum_{i=1}^n V_i^2\Big)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n V_i(\widehat{g}_0(Z_i) - g_0(Z_i)).$$

If we use sample splitting, conditional on the auxiliary sample, the key part of c^*, $\frac{1}{\sqrt{n}}\sum_{i=1}^n V_i(\widehat{g}_0(Z_i) - g_0(Z_i))$ has mean zero and variance

$$(\mathrm{E}V^2)\mathbb{E}_n(\widehat{g}_0(Z_i) - g_0(Z_i))^2 \to 0,$$

so that $c^* = o_P(1)$. If we do not use sample splitting, the key part is bounded by

$$\sup_{g\in\mathcal{G}_n}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n V_i(g(Z_i) - g_0(Z_i))\Big|,$$

where $\mathcal{G}_n$ is the smallest class of functions that contains the estimator $\widehat{g}$ with high probability. The function classes $\mathcal{G}_n$ are not Donsker and their entropy is growing with n, making it difficult to show that the term in the display above vanishes. Nonetheless, if $\mathcal{G}_n$'s entropy does not increase with n too rapidly, Belloni et al. [12] have proven that the terms like the one above and c^* more generally do vanish. However, verification of the entropy condition is so far only available for certain classes of machine learning methods, such as Lasso and Post-Lasso, and is likely to be difficult for practical versions of the methods that often employ data-driven tuning and cross-validation. It is also likely to be difficult for various hybrid methods, for example, the hybrid where we fit Random Forest after taking out the "smooth trend" in the data by Lasso.

Now we turn to a generalization of the orthogonalization principle above. The first "conventional" estimator $\widehat{\theta}_0$ given in (1.3) can be viewed as a solution for estimating equations

$$\frac{1}{n}\sum_{i=1}^{n} \varphi(W, \widehat{\theta}_0, \widehat{g}_0) = 0,$$

where φ is a known "score" function and $\widehat{g}_0$ is the estimator of the nuisance parameter g_0. In the partially linear model above, the score function is $\varphi(W, \theta, g) = (Y - \theta D - g(Z))D$. It is easy to see that this score function φ is sensitive to biased estimation of g. Specifically, the Gateauax derivative operator with respect to g does not vanish:

$$\partial_g \mathrm{E}\varphi(W, \theta_0, g)\Big|_{g=g_0} \neq 0.$$

The proofs of the general results in the next section show that this term's vanishing is a key to establishing good behavior of an estimator for θ_0.

By contrast the orthogonalized or double ML estimator $\check{\theta}_0$ given in (1.5) solves

$$\frac{1}{n}\sum_{i=1}^{n} \psi(W, \check{\theta}_0, \widehat{\eta}_0) = 0.$$

where ψ is the orthogonalized or debiased "score" function and $\widehat{\eta}_0$ is the estimator of the nuisance parameter η_0. In the partially linear model (1.1)-(1.2), the estimator uses the score function $\psi(W, \theta, \eta) = ((Y - \ell(Z) - \theta(D - m(Z)))(D - m(Z))$, with the nuisance parameter being $\eta = (\ell, m)$. It is easy to see that these score functions ψ are not sensitive to biased estimation of η_0. Specifically, the Gateuax derivative operator with respect to η vanishes in this case:

$$\partial_\eta \mathrm{E}\psi(W, \theta_0, \eta)\Big|_{\eta=\eta_0} = 0.$$

The proofs of the general results in the next section show that this property is the key to generating estimators with desired properties.

The basic problem outlined above is clearly related to the traditional semiparametric estimation framework which focuses on obtaining $\sqrt{n}$-consistent and asymptotically normal estimates for low-dimensional components with nuisance parameters estimated by conventional nonparametric estimators such as kernels or series. See, for example, the important work by [14], [50], [40], [54], [2], [41], [49], [38], [15], [19], [53], and [1]. The major point of departure from the present work and this traditional work is that we allow for the use of modern ML methods, a.k.a. machine learning methods, for modeling and fitting the non-parametric (or high-dimensional) components of the model for modern, high-dimensional data. As noted above, considering ML estimators requires us to accommodate estimators whose realizations belong to function classes $\mathcal{G}_n$ that are not Donsker and have entropy that grows with n. Conditions employed in the traditional semiparametric literature rule out this setting which necessitates the use of a different set of tools and development of new results. The framework we consider based on modern ML methods also expressly allows for data-driven choice of an approximating model for the high-dimensional component which addresses a crucial problem that arises in empirical work.

We organize there rest of the paper as follows. In Section 2, we present general theory for orthogonalized or "double" ML estimators. We present a formal application of the general results to estimation of average treatment effects (ATE) in partially linear model and in a fully heterogeneous effect model in Section 3. In Section 4 we present a sample application where we apply double ML methods to study the impact of 401(k) eligibility on accumulated assets. In an appendix, we define some additional notation and present proofs.

Notation. The symbols P and E denote probability and expectation operators with respect to a generic probability measure. If we need to signify the dependence on a probability measure P, we use P as a subscript in P_P and E_P. Note also that we use capital letters such as W to denote random elements and use the corresponding lower case letters such as w to denote fixed values that these random elements can take. In what follows, we use $\|\cdot\|_{P,q}$ to denote the $L^q(P)$ norm; for example, we denote

$$\|f(W)\|_{P,q} := \left(\int |f(w)|^q dP(w)\right)^{1/q}.$$

For a differentiable map $x \mapsto f(x)$, mapping $\mathbb{R}^d$ to $\mathbb{R}^k$, we use $\partial_{x'} f$ to abbreviate the partial derivatives $(\partial/\partial x')f$, and we correspondingly use the expression $\partial_{x'} f(x_0)$ to mean $\partial_{x'} f(x) \mid_{x=x_0}$, etc. We use x' to denote the transpose of a column vector x.

2. A General Approach to Post-Regularized Estimation and Inference Based on Orthogonalized Estimating Equations

2.1. Generic Construction of Orthogonal (Double ML) Estimators and Confidence Regions.

Here we formally introduce the model and state main results under high-level conditions. We are interested in the true value θ_0 of the low-dimensional target (causal) parameter $\theta \in \Theta$, where Θ is a convex subset of $\mathbb{R}^{d_\theta}$. We assume that θ_0 satisfies the moment conditions

$$\mathrm{E}_P[\psi_j(W, \theta_0, \eta_0)] = 0, \quad j = 1, \dots, d_\theta, \tag{2.1}$$

where $\psi = (\psi_1, \dots, \psi_{d_\theta})'$ is a vector of known score functions, W is a random element taking values in a measurable space $(\mathcal{W}, \mathcal{A}_\mathcal{W})$ with law determined by a probability measure $P \in \mathcal{P}_n$, and η_0 is the true value of the nuisance parameter $\eta \in T$ for some convex set T equipped with a norm $\|\cdot\|_e$. We assume that the score functions $\psi_j \colon \mathcal{W} \times \Theta \times T \to \mathbb{R}$ are measurable once we equip Θ and T with their Borel σ-fields. We assume that a random sample $(W_i)_{i=1}^N$ from the distribution of W is available for estimation and inference. As explained below in detail, we employ sample-splitting and assume that n observations are used to estimate θ_0 and the other $N - n$ observations are used to estimate η_0. The set of probability measures $\mathcal{P}_n$ is allowed to depend on n and in particular to expand as n gets large. Note that our formulation allows the nuisance parameter η to be infinite-dimensional; that is, η can be a function or a vector of functions.

As discussed in the introduction, we require the following orthogonality condition for the score ψ. If we start with a model with score φ that does not satisfy this orthogonality condition, we first transform it into a score ψ that satisfies this condition as described in the next section.

Definition 2.1 (Neyman orthogonality or unbiasedness condition). *The score $\psi = (\psi_1, \ldots, \psi_{d_\theta})'$ obeys the orthogonality condition with respect to $\mathcal{T} \subset T$ if the following conditions hold: The Gateaux derivative map*

$$\mathrm{D}_{r,j}[\eta - \eta_0] := \partial_r \Big\{ \mathrm{E}_P \Big[\psi_j(W, \theta_0, \eta_0 + r(\eta - \eta_0)) \Big] \Big\}$$

exists for all $r \in [0, 1)$, $\eta \in \mathcal{T}$, and $j = 1, \ldots, d_\theta$ and vanishes at $r = 0$; namely, for all $\eta \in \mathcal{T}$ and $j = 1, \ldots, d_\theta$,

$$\partial_\eta \mathrm{E}_P \psi_j(W, \theta_0, \eta)\Big|_{\eta = \eta_0} [\eta - \eta_0] := \mathrm{D}_{0,j}[\eta - \eta_0] = 0. \tag{2.2}$$

Estimation will be carried out using the finite-sample analog of the estimating equations (2.1). We assume that the true value η_0 of the nuisance parameter η can be estimated by $\widehat{\eta}_0$ using a part of the data $(W_i)_{i=1}^N$. Different structured assumptions on T allow us to use different machine-learning tools for estimating η_0. For instance,

1) smoothness of η_0 calls for the use of adaptive kernel estimators with bandwidth values obtained, for example, using the Lepski method;
2) approximate sparsity for η_0 with respect to some dictionary calls for the use of forward selection, lasso, post-lasso, or some other sparsity-based technique;
3) well-approximability of η_0 by trees calls for the use of regression trees and random forests.

Sample Splitting

In order to set up estimation and inference, we use sample splitting. We assume that n observations with indices $i \in I \subset \{1, \ldots, N\}$ are used for estimation of the target parameter θ, and the other $\pi n = N - n$ observations with indices $i \in I^c$ are used to provide estimator

$$\widehat{\eta}_0 = \widehat{\eta}_0(I^c)$$

of the true value η_0 of the nuisance parameter η. The parameter $\pi = \pi_n$ determines the portion of the entire data that is used for estimation of the nuisance parameter η. We assume that π is bounded away from zero, that is, $\pi \geqslant \pi_0 > 0$ for some fixed constant π_0, so that πn is at least of the same order as n, though we could allow for $\pi \to 0$ as $N \to \infty$ in principle. We assume that I and I^c form a random partition of the set $\{1, ..., N\}$. We conduct asymptotic analysis with respect to n increasing to ∞.

We let $\mathbb{E}_n$, and when needed $\mathbb{E}_{n,I}$, denote the empirical expectation with respect to the sample $(W_i)_{i \in I}$:

$$\mathbb{E}_n \psi(W) := \mathbb{E}_{n,I}[\psi(W)] = \frac{1}{n} \sum_{i \in I} \psi(W_i).$$

Generic Estimation

The true value η_0 of the nuisance parameter η is estimated by $\widehat{\eta}_0 = \widehat{\eta}_0(I^c)$ using the sample $(W_i)_{i\in I^c}$. The true value θ_0 of the target parameter θ is estimated by

$$\check{\theta}_0 = \widehat{\theta}_0(I, I^c)$$

using the sample $(W_i)_{i\in I}$. We construct the estimator $\check{\theta}_0$ of θ_0 as an approximate ϵ_n-solution in Θ to a sample analog of the moment conditions (2.1), that is,

$$\Big\| \mathbb{E}_{n,I}[\psi(W, \check{\theta}_0, \widehat{\eta}_0)\Big\| \leqslant \inf_{\theta\in\Theta} \Big\| \mathbb{E}_{n,I}[\psi(W, \theta, \widehat{\eta}_0)]\Big\| + \epsilon_n, \quad \epsilon_n = o(\delta_n n^{-1/2}), \tag{2.3}$$

where $(\delta_n)_{n\geqslant 1}$ is some sequence of positive constants converging to zero.

Let ω, c_0, and C_0 be some strictly positive (and finite) constants, and let $n_0 \geqslant 3$ be some positive integer. Also, let $(B_{1n})_{n\geqslant 1}$ and $(B_{2n})_{n\geqslant 1}$ be some sequences of positive constants, possibly growing to infinity, where $B_{1n} \geqslant 1$ and $B_{2n} \geqslant 1$ for all $n \geqslant 1$. Denote

$$J_0 := \partial_{\theta'} \Big\{ \mathrm{E}_P[\psi(W, \theta, \eta_0)] \Big\}\Big|_{\theta=\theta_0}. \tag{2.4}$$

The quantity J_0 measures the degree of identifiability of θ_0 by the moment conditions (2.1). In typical cases, the singular values of J_0 will be bounded from above and away from zero.

We are now ready to state our main regularity conditions.

Assumption 2.1 (**Moment condition problem**). *For all $n \geqslant n_0$ and $P \in \mathcal{P}_n$, the following conditions hold. (i) The true parameter value θ_0 obeys (2.1), and Θ contains a ball of radius $C_0 n^{-1/2}\log n$ centered at θ_0. (ii) The map $(\theta,\eta)\mapsto \mathrm{E}_P[\psi(W,\theta,\eta)]$ is twice continuously Gateaux-differentiable on $\Theta\times\mathcal{T}$. (iii) The score ψ obeys the near orthogonality condition given in Definition 2.1 for the set $\mathcal{T}\subset T$. (iv) For all $\theta\in\Theta$, we have $\|\mathrm{E}_P[\psi(W,\theta,\eta_0)]\| \geqslant 2^{-1}\|J_0(\theta-\theta_0)\|\wedge c_0$, where singular values of J_0 are between c_0 and C_0. (v) For all $r\in[0,1)$, $\theta\in\Theta$, and $\eta\in\mathcal{T}$,*

(a) $\mathrm{E}_P[\|\psi(W,\theta,\eta) - \psi(W,\theta_0,\eta_0)\|^2] \leqslant C_0(\|\theta-\theta_0\| \vee \|\eta-\eta_0\|_e)^{\omega}$,
(b) $\|\partial_r \mathrm{E}_P\left[\psi(W,\theta,\eta_0 + r(\eta-\eta_0))\right]\| \leqslant B_{1n}\|\eta-\eta_0\|_e$,
(c) $\|\partial_r^2 \mathrm{E}_P[\psi(W,\theta_0 + r(\theta-\theta_0), \eta_0 + r(\eta-\eta_0))]\| \leqslant B_{2n}(\|\theta-\theta_0\|^2 \vee \|\eta-\eta_0\|_e^2)$.

Assumption 2.1 is mild and standard in moment condition problems. Assumption 2.1(i) requires θ_0 to be sufficiently separated from the boundary of Θ. Assumption 2.1(ii) is rather weak because it only requires differentiability of the function $(\theta,\eta)\mapsto \mathrm{E}_P[\psi(W,\theta,\eta)]$ and does not require differentiability of the function $(\theta,\eta)\mapsto\psi(W,\theta,\eta)$. Assumption 2.1(iii) is discussed above. Assumption 2.1(iv) implies sufficient identifiability of θ_0. Assumptions 2.1(v) is a smoothness condition.

Next, we state conditions related to the estimator $\widehat{\eta}_0$. Let $(\Delta_n)_{n\geqslant 1}$ and $(\tau_{\pi n})_{n\geqslant 1}$ be some sequences of positive constants converging to zero. Also, let $a > 1$, $v > 0$, $K > 0$, and $q > 2$ be some constants.

Assumption 2.2 (**Quality of estimation of nuisance parameter and score regularity**). *For all $n \geqslant n_0$ and $P \in \mathcal{P}_n$, the following conditions hold. (i) With probability at least $1 - \Delta_n$, we have $\widehat{\eta}_0 \in \mathcal{T}$. (ii) For all $\eta \in \mathcal{T}$, we have $\|\eta - \eta_0\|_e \leqslant \tau_{\pi n}$. (iii) The true value η_0 of the nuisance parameter η satisfies $\eta_0 \in \mathcal{T}$. (iv) For all $\eta \in \mathcal{T}$, the function class $\mathcal{F}_{1,\eta} = \{\psi_j(\cdot, \theta, \eta) \colon j = 1, ..., d_\theta, \theta \in \Theta\}$ is suitably measurable and its uniform entropy numbers obey*

$$\sup_Q \log N(\epsilon \|F_{1,\eta}\|_{Q,2}, \mathcal{F}_{1,\eta}, \|\cdot\|_{Q,2}) \leqslant v \log(a/\epsilon), \quad \textit{for all } 0 < \epsilon \leqslant 1 \tag{2.5}$$

where $F_{1,\eta}$ is a measurable envelope for $\mathcal{F}_{1,\eta}$ that satisfies $\|F_{1,\eta}\|_{P,q} \leqslant K$. (v) For all $\eta \in \mathcal{T}$ and $f \in \mathcal{F}_{1,\eta}$, we have $c_0 \leqslant \|f\|_{P,2} \leqslant C_0$. (vi) The estimation rate $\tau_{\pi n}$ satisfies (a) $n^{-1/2} \leqslant C_0 \tau_{\pi n}$, (b) $(B_{1n}\tau_{\pi n})^{\omega/2} + n^{-1/2+1/q} \leqslant C_0 \delta_n$, and (c) $n^{1/2} B_{1n}^2 B_{2n} \tau_{\pi n}^2 \leqslant C_0 \delta_n$.

Assumption 2.2 states requirements on the quality of estimation of η_0 as well as imposes some mild assumptions on the score ψ. The estimator $\widehat{\eta}_0$ has to converge to η_0 at the rate $\tau_{\pi n}$, which needs to be faster than $n^{-1/4}$, with a more precise requirement stated above. Note that if $\pi \to 0$, the requirements on the quality of $\widehat{\eta}_0$ become more stringent. This rate condition is widely used in traditional semi-parametric estimation which employs classical nonparametric estimators for $\widehat{\eta}_0$. The new generation of machine learning methods are often able to perform much better than the classical methods, and so the requirement may be more easily satisfied by these methods. Suitable measurability, required in Assumption 2.2(iv), is a mild regularity condition that is satisfied in all practical cases. Assumption 2.2(vi) is a set of growth conditions.

Theorem 2.1 (**Uniform Bahadur Representation and Approximate Normality**). *Under Assumptions 2.1 and 2.2, the estimator $\check{\theta}_0$ defined by equation (2.3), obeys*

$$\sqrt{n}\Sigma_0^{-1/2}(\check{\theta}_0 - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i \in I} \bar{\psi}(W_i) + O_P(\delta_n) \rightsquigarrow N(0, I),$$

uniformly over $P \in \mathcal{P}_n$, where $\bar{\psi}(\cdot) := -\Sigma_0^{-1/2} J_0^{-1} \psi(\cdot, \theta_0, \eta_0)$ and

$$\Sigma_0 := J_0^{-1} \mathrm{E}_P[\psi(W, \theta_0, \eta_0)\psi(W, \theta_0, \eta_0)'](J_0^{-1})'.$$

The result establishes that the estimator based on the orthogonalized scores achieves the root-n rate of convergence and is approximately normally distributed. It is noteworthy that this convergence result, both rate and distributional approximation, holds uniformly with respect to P varying over an expanding class of probability measures $\mathcal{P}_n$. This means that the convergence holds under any sequence of probability distributions $\{P_n\}$ with $P_n \in \mathcal{P}_n$ for each n, which in turn implies that the results are robust with respect to perturbations of a given P along such sequences. The same property can be shown to fail for methods not based on orthogonal scores. The result can be used for standard construction of confidence regions which are uniformly valid over a large, interesting class of models.

An estimator based on sample-splitting does not use the full sample by construction. However, there will be no asymptotic loss in efficiency from sample splitting if it is possible to send $\pi_n \searrow 0$

while satisfying Assumption 2.1. Such a sequence requires the number of observations $\pi_n n$ used for producing $\widehat{\eta}_0$ to be small compared to n while also requiring the estimator of the nuisance parameter to remain of sufficient quality given this small number of observations.

Corollary 2.1 (**Achieving no efficiency loss from sample-splitting**). *(1) If $\pi_n \to 0$ and the conditions of Theorem 2.1 continue to hold, the sample-splitting estimator obeys*

$$\sqrt{N}\Sigma_0^{-1/2}(\check{\theta}_0 - \theta_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \bar{\psi}(W_i) + o_P(1) \rightsquigarrow N(0, I),$$

uniformly over $P \in \mathcal{P}_n$. That is, it is as asymptotically efficient as using all N observations.

Below we explore alternative ways of sample splitting.

2.2. **Achieving Full Efficiency by Cross-Fitting.** Here we exploit the use of cross-fitting, following [3], for preventing loss of efficiency.

Full Efficiency by 2-fold Cross-Fitting

We could proceed with a random 50-50 split of $\{1, ..., N\}$ into I and I^c. This means that the ratio of the sizes of the samples I^c and I is $\pi = 1$. Indeed, the size of I is n, the size of I^c is also n, and the total sample size is $N = 2n$. We may then construct an estimator $\check{\theta}_0(I, I^c)$ that employs the nuisance parameter estimator $\widehat{\eta}_0(I^c)$, as before. Then, we reverse the roles of I and I^c and construct an estimator $\check{\theta}_0(I^c, I)$ that employs the nuisance parameter estimator $\widehat{\eta}_0(I)$. The two estimators may then be aggregated into the final estimator:

$$\tilde{\theta}_0 = \check{\theta}_0(I, I^c)/2 + \check{\theta}(I, I^c)/2. \tag{2.6}$$

This 2-fold cross-fitting generalizes to the K-fold cross-fitting, which is subtly different from (and hence should not be confused with) cross-validation. This approach can be thought as a "leave-a-block out" approach.

Full Efficiency by K-fold Cross-Fitting

We could proceed with a K-fold random split $I_k, k = 1, ..., K$ of the entire sample $\{1, ..., N\}$, so that $\pi = K - 1$. In this case, the size of each split I_k is $n = N/K$, the size of $I_k^c = \cup_{m \neq k} I_m$ is $N \cdot [(K-1)/K]$, and the total sample size is N. We may then construct K estimators

$$\check{\theta}_0(I_k, I_k^c), \quad k = 1, ..., K,$$

that employ the nuisance parameter estimators $\widehat{\eta}_0(I_k^c)$. The K estimators may then be aggregated into

$$\tilde{\theta}_0 = \frac{1}{K}\sum_{k=1}^{K} \check{\theta}_0(I_k, I_k^c). \tag{2.7}$$

The following is an immediate corollary of Theorem 2.1 that shows that resulting estimators entail no loss from sample-splitting asymptotically under assumptions of the theorem.

Theorem 2.2 (**Achieving no efficiency loss by K-fold cross-fitting**). *Under the conditions of Theorem 2.1, the aggregated estimator $\tilde{\theta}_0$ defined by equation (2.7), obeys*

$$\sqrt{N}\Sigma_0^{-1/2}(\tilde{\theta}_0 - \theta_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\bar{\psi}(W_i) + o_P(1) \rightsquigarrow N(0,I),$$

uniformly over $P \in \mathcal{P}_n$.

If the score function turns out to be efficient for estimating θ_0, then the resulting estimator is also efficient in the semi-parametric sense.

Corollary 2.2 (**Semi-parametric efficiency**). *If the score function ψ is the efficient score for estimating θ_0 at a given $P \in \mathcal{P} \subset \mathcal{P}_n$, in the semi-parametric sense as defined in* [55]*, then the large sample variance Σ of $\tilde{\theta}_0$ reaches the semi-parametric efficiency bound at this P relative to the model $\mathcal{P}$.*

Note that efficient scores are automatically orthogonal with respect to the nuisance parameters by construction as discussed below. It should be noted though that orthogonal scores do not have to be efficient scores. For instance, the scores discussed in the introduction for the partially linear are only efficient in the homoscedastic model.

2.3. **Construction of score functions satisfying the orthogonality condition.** Here we discuss several methods for generating orthogonal scores in a wide variety of settings, including the classical Neyman's construction.

1) Orthogonal Scores for Likelihood Problems with Finite-Dimensional Nuisance Parameters. In likelihood settings with finite-dimensional parameters, the construction of orthogonal equations was proposed by Neyman [42] who used them in construction of his celebrated $C(\alpha)$-statistic.[4]

Suppose that the log-likelihood function associated to observation W is $(\theta,\beta) \mapsto \ell(W,\theta,\beta)$, where $\theta \in \Theta \subset \mathbb{R}^d$ is the target parameter and $\beta \in T \subset \mathbb{R}^{p_0}$ is the nuisance parameter. Under regularity conditions, the true parameter values θ_0 and β_0 obey

$$\mathrm{E}[\partial_\theta \ell(W,\theta_0,\beta_0)] = 0, \quad \mathrm{E}[\partial_\beta \ell(W,\theta_0,\beta_0)] = 0. \tag{2.8}$$

Note that the original score function $\varphi(W,\theta,\beta) = \partial_\theta \ell(W,\theta,\beta)$ in general does not possess the orthogonality property. Now consider the new score function

$$\psi(W,\theta,\eta) = \partial_\theta \ell(W,\theta,\beta) - \mu \partial_\beta \ell(W,\theta,\beta), \tag{2.9}$$

[4] The $C(\alpha)$-statistic, or the orthogonal score statistic, has been explicitly used for testing and estimation in high-dimensional sparse models in [10]. The discussion of Neyman's construction here draws on [28].

where the nuisance parameter is

$$\eta = (\beta', \operatorname{vec}(\mu)')' \in T \times \mathcal{D} \subset \mathbb{R}^p, \quad p = p_0 + dp_0,$$

μ is the $d \times p_0$ *orthogonalization* parameter matrix whose true value μ_0 solves the equation

$$J_{\theta\beta} - \mu J_{\beta\beta} = 0 \text{ (i.e., } \mu_0 = J_{\theta\beta} J_{\beta\beta}^{-1}),$$

and

$$J = \begin{pmatrix} J_{\theta\theta} & J_{\theta\beta} \\ J_{\beta\theta} & J_{\beta\beta} \end{pmatrix} = \partial_{(\theta',\beta')} \mathrm{E}\Big[\partial_{(\theta',\beta')'} \ell(W, \theta, \beta)\Big]\Big|_{\theta=\theta_0;\ \beta=\beta_0}.$$

Provided that μ_0 is well-defined, we have by (2.8) that

$$\mathrm{E}[\psi(W, \theta_0, \eta_0)] = 0,$$

where $\eta_0 = (\beta_0', \operatorname{vec}(\mu_0)')'$. Moreover, it is trivial to verify that under standard regularity conditions the score function ψ obeys the orthogonality condition (2.2) exactly, that is,

$$\partial_\eta \mathrm{E}[\psi(W, \theta_0, \eta)]\Big|_{\eta=\eta_0} = 0.$$

Note that in this example, μ_0 not only creates the necessary orthogonality but also creates the *efficient score* for inference on the target parameter θ, as emphasized by Neyman [42].

2) Orthogonal Scores for Likelihood Problems with Infinite-Dimensional Nuisance Parameters. Neyman's construction can be extended to semi-parametric models where the nuisance parameter β is a function. In this case, the original score functions $(\theta, \beta) \mapsto \partial_\theta \ell(W, \theta, \beta)$ corresponding to the log-likelihood function $(\theta, \beta) \mapsto \ell(W, \theta, \beta)$ associated to observation W can be transformed into efficient score functions ψ that obey the orthogonality condition by projecting the original score functions onto the orthocomplement of the tangent space induced by the nuisance parameter β; see Chapter 25 of [55] for a detailed description of this construction. By selecting elements of the orthocomplement, we generate scores that are orthogonal but not necessarily efficient. The projection gives the unique score that is efficient. Note that the projection may create additional nuisance parameters, so that the new nuisance parameter η could be of larger dimension than β. Other relevant references include [56], [33], [8], and [10]. The approach is related to Neyman's construction in the sense that the score ψ arising in this model is actually the Neyman's score arising in a one-dimensional least favorable parametric subfamily; see Chapter 25 of [55] for details.

3) Orthogonal Scores for Conditional Moment Problems with Infinite-Dimensional Nuisance Parameters. Next, consider a conditional moment restrictions framework studied by Chamberlain [18]:

$$\mathrm{E}[\varphi(W, \theta_0, h_0(X)) \mid X] = 0,$$

where X and W are random vectors with X being a sub-vector of W, $\theta \in \Theta \subset \mathbb{R}^d$ is a finite-dimensional parameter whose true value θ_0 is of interest, h is a functional nuisance parameter mapping the support of X into a convex set $V \subset \mathbb{R}^l$ whose true value is h_0, and φ is a known function with values in $\mathbb{R}^k$ for $k \geqslant d + l$. This framework is of interest because it covers a rich variety of models without having to explicitly rely on the likelihood formulation.

Here we would like to build a (generalized) score function $(\theta, \eta) \mapsto \psi(W, \theta, \eta)$ for estimating θ_0, the true value of parameter θ, where η is a new nuisance parameter with true value η_0 that obeys the orthogonality condition (2.2). To this end, let $v \mapsto \mathrm{E}[\varphi(W, \theta_0, v) \mid X]$ be a function mapping $\mathbb{R}^l$ into $\mathbb{R}^k$ and let

$$\gamma(X, \theta_0, h_0) = \partial_{v'} \mathrm{E}[\varphi(W, \theta_0, v) \mid X]|_{v=h_0(X)}$$

be a $k \times l$ matrix of its derivatives. We will set $\eta = (h, \beta, \Sigma)$ where β is a function mapping the support of X into the space of $d \times k$ matrices, $\mathbb{R}^{d\times k}$, and Σ is the function mapping the support of X into the space of $k \times k$ matrices, $\mathbb{R}^{k\times k}$. Define the true value β_0 of β as

$$\beta_0(X) = A(X)(I_{k\times k} - \Pi(X)),$$

where $A(X)$ is a $d \times k$ matrix of measurable transformations of X, $I_{k\times k}$ is the $k \times k$ identity matrix, and $\Pi(X) \neq I_{k\times k}$ is a $k \times k$ non-identity matrix with the property:

$$\Pi(X)\Sigma_0^{-1/2}(X)\gamma(X, \theta_0, h_0) = \Sigma_0^{-1/2}(X)\gamma(X, \theta_0, h_0), \tag{2.10}$$

where Σ_0 is the true value of parameter Σ. For example, $\Pi(X)$ can be chosen to be an orthogonal projection matrix:

$$\Pi(X) = \Big[\Sigma_0(X)^{-1/2}\gamma(X, \theta_0, h_0) \left(\gamma(X, \theta_0, h_0)'\Sigma_0(X)^{-1}\gamma(X, \theta_0, h_0)\right)^{-1} \\ \times \gamma(X, \theta_0, h_0)'\Sigma_0(X)^{-1/2} \Big].$$

Then an orthogonal score for the problem above can be constructed as

$$\psi(W, \theta, \eta) = \beta(X)\Sigma^{-1/2}(X)\varphi(Z, \theta, h(X)), \quad \eta = (h, \beta, \Sigma).$$

It is straightforward to check that under mild regularity conditions the score function ψ satisfies $\mathrm{E}[\psi(W, \theta_0, \eta_0)] = 0$ for $\eta_0 = (h_0, \varphi_0, \Sigma_0)$ and also obeys the exact orthogonality condition. Furthermore, by setting

$$A(X) = \left(\partial_{\theta'} \mathrm{E}[\varphi(W, \theta, h_0(X) \mid X]|_{\theta=\theta_0}\right)', \quad \Sigma_0(X) = \mathrm{E}\Big[\varphi(W, \theta_0, h_0(X))\varphi(W, \theta_0, h_0(X))'|X\Big],$$

and using $\Pi(X)$ suggested above, we obtain the efficient score ψ that yields an estimator of θ_0 achieving the semi-parametric efficiency bound, as calculated by Chamberlain [18].

3. Application to Estimation of Treatment Effects

3.1. **Treatment Effects in the Partially Linear Model.**

Here we revisit the partially linear model

$$Y = D\theta_0 + g_0(Z) + \zeta, \quad \mathrm{E}[\zeta \mid Z, D] = 0, \tag{3.1}$$

$$D = m_0(Z) + V, \qquad \mathrm{E}[V \mid Z] = 0. \tag{3.2}$$

If D is as good as randomly assigned conditional on covariates, then θ_0 measures the average treatment effect of D on potential outcomes.

The first approach, which we described in the introduction, employs the score function

$$\psi(W, \theta, \eta) := \{Y - \ell(Z) - \theta(D - m(Z))\}(D - m(Z)), \quad \eta = (\ell, m), \tag{3.3}$$

where ℓ and m are P-square-integrable functions mapping the support of Z to $\mathbb{R}$.

It is easy to see that θ_0 is a solution to

$$\mathrm{E}_P \psi(W, \theta_0, \eta_0) = 0,$$

and the orthogonality condition holds:

$$\partial_\eta \mathrm{E}_P \psi(W, \theta_0, \eta)\Big|_{\eta=\eta_0} = 0, \quad \eta_0 = (\ell_0, m_0),$$

where $\ell_0(Z) = \mathrm{E}_P[Y|Z]$. This approach represents a generalization of the approach of [7, 8] considered for the case of Lasso without cross-fitting. As mentioned, this generalization opens up the use of a much broader collection of machine learning methods, much beyond Lasso.

The second approach, which is first-order equivalent to the first, is to employ the score function

$$\psi(W, \theta, \eta) := \{Y - D\theta - g(Z)\}(D - m(Z)), \quad \eta = (g, m), \tag{3.4}$$

where g and m are P-square-integrable functions mapping the support of Z to $\mathbb{R}$. It is easy to see that θ_0 is a solution to

$$\mathrm{E}_P \psi(W, \theta_0, \eta_0) = 0,$$

and the orthogonality condition holds:

$$\partial_\eta \mathrm{E}_P \psi(W, \theta_0, \eta)\Big|_{\eta=\eta_0} = 0, \quad \eta_0 = (g_0, m_0).$$

This approach can be seen as "debiasing" the score function $(Y - D\theta - g(Z))D$, which does not possess the orthogonality property unless $m_0(Z) = 0$. This second approach represents a generalization of the approach of [31, 52, 57] considered for the case of Lasso-type methods without cross-fitting. Like above, our generalization allows for the use of broad collection machine learning methods, much beyond Lasso-type methods.

Algorithm 1 (**Double ML Estimation and Inference on ATE in the Partially Linear Model.**). We describe two estimators of θ_0 based on the use of score functions (3.3) and (3.3). Let K be a fixed integer. We construct a K-fold random partition of the entire sample $\{1, ..., N\}$ into equal parts $(I_k)_{k=1}^K$ each of size $n := N/K$, and construct the K estimators

$$\check{\theta}_0(I_k, I_k^c), \quad k = 1, ..., K,$$

where each estimator $\check{\theta}_0(I_k, I_k^c)$ is the root θ of the equation:

$$\frac{1}{n} \sum_{i \in I_k} \psi(W, \theta, \widehat{\eta}_0(I_k^c)) = 0,$$

for the score ψ defined in (3.3) or (3.4); for the case with the score function given by (3.3), the estimator employs the nuisance parameter estimators

$$\widehat{\eta}_0(I_k^c) := (\widehat{\ell}_0(Z; I_k^c), \widehat{m}_0(Z; I_k^c)),$$

based upon machine learning estimators of $\ell_0(Z)$ and $m_0(Z)$ using auxiliary sample I_k^c; and, for the case with the score function given by (3.4), the estimator employs the nuisance parameter estimators

$$\widehat{\eta}_0(I_k^c) := (\widehat{g}_0(Z; I_k^c), \widehat{m}_0(Z; I_k^c)),$$

based upon machine learning estimators of $g_0(Z)$ and $m_0(Z)$ using auxiliary sample I_k^c.

We then average the K estimators to obtain the final estimator:

$$\tilde{\theta}_0 = \frac{1}{K} \sum_{k=1}^{K} \check{\theta}_0(I_k, I_k^c). \tag{3.5}$$

The approximate standard error for this estimator is given by $\widehat{\sigma}/\sqrt{N}$, where

$$\widehat{\sigma}^2 = \Big(\frac{1}{N} \sum_{i=1}^{N} \widehat{V}_i^2 \Big)^{-2} \frac{1}{N} \sum_{i=1}^{N} \widehat{V}_i^2 \widehat{\zeta}_i^2,$$

where $\widehat{V}_i := D_i - \widehat{m}(Z_i, I_{k(i)}^c)$, $\widehat{\zeta}_i := (Y_i - \widehat{\ell}_0(Z_i, I_{k(i)}^c) - (D_i - \widehat{m}_0(Z_0, I_{k(i)}^c))\tilde{\theta}_0$ or $\widehat{\zeta}_i := Y_i - D_i \tilde{\theta}_0 - \widehat{g}_0(Z_0, I_{k(i)}^c)$, and $k(i) := \{k \in \{1, ..., K\} : i \in I_k\}$. The approximate $(1 - \alpha) \times 100\%$ confidence interval is given by:

$$[\tilde{\theta}_0 \pm \Phi^{-1}(1 - \alpha/2)\widehat{\sigma}/\sqrt{N}].$$

Let $(\delta_n)_{n=1}^{\infty}$ and $(\Delta_n)_{n=1}^{\infty}$ be sequences of positive constants approaching 0 as before. Let c and C be fixed positive constants and $K \geqslant 2$ be a fixed integer, and let $q > 4$.

Assumption 3.1. *Let $\mathcal{P}$ be the collection of probability laws P for the triple (Y, D, Z) such that: (i) equations (3.1)-(3.2) hold; (ii) the true parameter value θ_0 is bounded, $|\theta_0| \leqslant C$; (iii) $\|X\|_{P,q} \leqslant C$ for $X \in \{Y, D, g_0(Z), \ell_0(Z), m_0(Z)\}$; (iv) $\|V\|_{P,2} \geqslant c$ and $\|\zeta\|_{P,2} \geqslant c$; and (v) the ML estimators of the nuisance parameters based upon a random subset I_k^c of $\{1, ..., N\}$ of size $N - n$, for $n = N/K$, obey the condition: $\|\widehat{\eta}_0(Z, I_k^c) - \eta_0(Z, I_k^c)\|_{P,2} \leqslant \delta_n n^{-1/4}$ for all $n \geqslant 1$ with P-probability no less than $1 - \Delta_n$.*

Comment 3.1. The only non-primitive condition is the assumption on the rate of estimating the nuisance parameters. These rates of convergence are available for most often used ML methods and are case-specific, so we do not restate conditions that are needed to reach these rates. The conditions are not the tightest possible, but we choose to present the simple ones, so that results below follows as a special case of the general theorem of the previous section. We can easily obtain more refined conditions by doing customized proofs. ■

The following theorem follows as a corollary to the results in the previous section.

Theorem 3.1 (Estimation and Inference on Treatment Effects in the Partially Linear Model)**.** *Suppose Assumption 3.1 holds. Then, as $N \to \infty$, both of the two double ML estimators $\tilde{\theta}_0$, constructed in Algorithm 1 above, are first-order equivalent and obey*

$$\sigma^{-1}\sqrt{N}(\tilde{\theta}_0 - \theta_0) \rightsquigarrow N(0,1),$$

uniformly over $P \in \mathcal{P}$, where $\sigma^2 = [\mathrm{E}_P V^2]^{-1}\mathrm{E}_P[V^2\zeta^2][\mathrm{E}_P V^2]^{-1}$, and the result continues to hold if σ^2 is replaced by $\widehat{\sigma}^2$. Furthermore, the confidence regions based upon Double ML estimator $\check{\theta}_0$ have the uniform asymptotic validity:

$$\lim_{N\to\infty} \sup_{P\in\mathcal{P}} \left| \mathrm{P}_P\left(\theta_0 \in [\tilde{\theta}_0 \pm \Phi^{-1}(1-\alpha/2)\widehat{\sigma}/\sqrt{N}]\right) - (1-\alpha)\right| = 0.$$

Comment 3.2. Note that under conditional homoscedasticity, namely $\mathrm{E}[\zeta^2|Z] = \mathrm{E}[\zeta^2]$, the asymptotic variance σ^2 reduces to $\mathrm{E}[V^2]^{-1}\mathrm{E}[\zeta^2]$, which is the semi-parametric efficiency bound for the partially linear model. ■

3.2. **Treatment Effects in the Interactive Model.** We next consider estimation of average treatment effects (ATE) when treatment effects are fully heterogeneous and the treatment variable is binary, $D \in \{0,1\}$. We consider vectors (Y, D, Z) such that

$$Y = g_0(D,Z) + \zeta, \quad \mathrm{E}[\zeta \mid Z, D] = 0, \tag{3.6}$$

$$D = m_0(Z) + \nu, \quad \mathrm{E}[\nu \mid Z] = 0. \tag{3.7}$$

Since D is not additively separable, this model is more general than the partially linear model for the case of binary D. A common target parameter of interest in this model is the average treatment effect (ATE),

$$\mathrm{E}[g_0(1,Z) - g_0(0,Z)].$$

Another common target parameter is the average treatment effect for the treated (ATTE)

$$\mathrm{E}[g_0(1,Z) - g_0(0,Z)|D = 1].$$

The confounding factors Z affect the policy variable via the propensity score $m(Z)$ and the outcome variable via the function $g_0(D,Z)$. Both of these functions are unknown and potentially complicated, and we can employ machine learning methods to learn them.

We proceed to set up moment conditions with scores obeying orthogonality conditions. For estimation of the ATE, we employ

$$\begin{aligned} \psi(W,\theta,\eta) &:= (g(1,Z) - g(0,Z)) + \frac{D(Y - g(1,Z))}{m(Z)} - \frac{(1-D)(Y-g(0,Z))}{1-m(Z)} - \theta, \\ \eta(Z) &:= (g(0,Z), g(1,Z), m(Z)), \quad \eta_0(Z) := (g_0(0,Z), g_0(1,Z), m_0(Z))', \end{aligned} \tag{3.8}$$

where $\eta(Z)$ is the nuisance parameter consisting of P-square integrable functions mapping the support of Z to $\mathbb{R} \times \mathbb{R} \times (\varepsilon, 1-\varepsilon)$, with the true value of this parameter denoted by $\eta_0(Z)$, where $\varepsilon > 0$ is a constant.

For estimation of ATTE, we use the score

$$\begin{aligned} \psi(W, \theta, \eta) &= \frac{D(Y - g(0,Z))}{m} - \frac{m(Z)(1-D)(Y-g(0,Z))}{(1-m(Z))m} - \theta\frac{D}{m}, \\ \eta(Z) &:= (g(0,Z), g(1,Z), m(Z), m), \quad \eta_0(Z) = (g_0(0,Z), g_0(1,Z), m_0(Z), \mathrm{E}[D]),' \end{aligned} \tag{3.9}$$

where $\eta(Z)$ is the nuisance parameter consisting of three P-square integrable functions mapping the support of Z to $\mathbb{R} \times \mathbb{R} \times (\varepsilon, 1-\varepsilon)$ and a constant $m \in (\varepsilon, 1-\varepsilon)$, with the true value of this parameter denoted by $\eta_0(Z)$.

It can be easily seen that true parameter values θ_0 for ATE and ATTE obey

$$\mathrm{E}_P \psi(W, \theta_0, \eta_0) = 0,$$

for the respective scores and that the scores have the required orthogonality property:

$$\partial_\eta \mathrm{E}_P \psi(W, \theta_0, \eta)\Big|_{\eta=\eta_0} = 0.$$

Algorithm 2 (**Double ML Estimation and Inference on ATE and ATTE in the Interactive Model.**). We describe the estimator of θ_0 next. Let K be a fixed integer. We construct a K-fold random partition of the entire sample $\{1, ..., N\}$ into equal parts $(I_k)_{k=1}^K$ each of size $n := N/K$, and construct the K estimators

$$\check{\theta}_0(I_k, I_k^c), \quad k = 1, ..., K,$$

that employ the machine learning estimators

$$\widehat{\eta}_0(I_k^c) = \left(\widehat{g}_0(0, Z; I_k^c),\ \widehat{g}_0(1, Z; I_k^c),\ \widehat{m}_0(Z; I_k^c),\ \frac{1}{N-n}\sum_{i \in I_k^c} D_i\right)',$$

of the nuisance parameters

$$\eta_0(Z) = (g_0(0,Z), g_0(1,Z), m_0(Z), \mathrm{E}[D]),$$

and where each estimator $\check{\theta}_0(I_k, I_k^c)$ is defined as the root θ of the corresponding equation:

$$\frac{1}{n}\sum_{i \in I_k} \psi(W, \theta, \widehat{\eta}_0(I_k^c)) = 0,$$

for the score ψ defined in (3.8) for ATE and in (3.9) for ATTE.

We then average the K estimators to obtain the final estimator:

$$\tilde{\theta}_0 = \frac{1}{K}\sum_{k=1}^K \check{\theta}_0(I_k, I_k^c). \tag{3.10}$$

The approximate standard error for this estimator is given by $\widehat{\sigma}/\sqrt{N}$, where

$$\widehat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} \widehat{\psi}_i^2$$

where $\widehat{\psi}_i := \psi(W_i, \tilde{\theta}_0, \widehat{\eta}_0(I^c_{k(i)}))$, and $k(i) := \{k \in \{1, ..., K\} : i \in I_k\}$. The approximate $(1-\alpha) \times 100\%$ confidence interval is given by

$$[\tilde{\theta}_0 \pm \Phi^{-1}(1-\alpha/2)\widehat{\sigma}/\sqrt{N}].$$

Let $(\delta_n)_{n=1}^{\infty}$ and $(\Delta_n)_{n=1}^{\infty}$ be sequences of positive constants approaching 0, as before, and c, ε, C and $q > 4$ be fixed positive constants, and K be a fixed integer.

Assumption 3.2 (TE in the Heterogenous Model)**.** *Let $\mathcal{P}$ be the set of probability distributions P for (Y, D, Z) such that (i) equations (3.6)-(3.7) hold, with $D \in \{0, 1\}$, (ii) the moment conditions hold: $\|g\|_{P,q} \leqslant C$, $\|Y\|_{P,q} \leqslant C$, $P(\varepsilon \leqslant m_0(Z) \leqslant 1-\varepsilon) = 1$, and $\|\zeta^2\|_{P,2} \geqslant c$, and (ii) the ML estimators of the nuisance parameters based upon a random subset I_k^c of $\{1, ..., N\}$ of size $N-n$, for $n = N/K$, obey the condition: $\|\widehat{g}_0(D, Z, I_k^c) - g_0(D, Z, I_k^c)\|_{P,2} + \|\widehat{m}_0(Z, I_k^c) - m_0(Z, I_k^c)\|_{P,2} \leqslant \delta_n n^{-1/4}$ and $\|\widehat{m}_0(Z, I_k^c) - m_0(Z, I_k^c)\|_{P,\infty} \leqslant \delta_n$, for all $n \geqslant 1$ with P-probability no less than $1 - \Delta_n$.*

Comment 3.3. The only non-primitive condition is the assumption on the rate of estimating the nuisance parameters. These rates of convergence are available for most often used ML methods and are case-specific, so we do not restate conditions that are needed to reach these rates. The conditions are not the tightest possible, but we chose to present the simple ones, since the results below follow as a special case of the general theorem of the previous section. We can easily obtain more refined conditions by doing customized proofs. ■

The following theorem follows as a corollary to the results in the previous section.

Theorem 3.2 (**Double ML Inference on ATE and ATT**)**.** *(1) Suppose that the ATE, $\theta_0 = \mathrm{E}[g_0(1, Z) - g_0(0, Z)]$, is the target parameter and we use the estimator $\tilde{\theta}_0$ and other notations defined above. (2) Alternatively, suppose that the ATTE, $\theta_0 = \mathrm{E}[g_0(1, Z) - g_0(0, Z) \mid D = 1]$, is the target parameter and we use the estimator $\tilde{\theta}_0$ and other notations above. Consider the set $\mathcal{P}$ of data generating defined in Assumption 3.2. Then uniformly in $P \in \mathcal{P}$, the Double ML estimator $\tilde{\theta}_0$ concentrates around θ_0 with the rate $1/\sqrt{N}$ and is approximately unbiased and normally distributed:*

$$\sigma^{-1}\sqrt{N}(\check{\theta}_0 - \theta_0) \rightsquigarrow N(0, 1), \quad \sigma^2 = \mathrm{E}_P[\psi^2(W, \theta_0, \eta_0(Z))], \tag{3.11}$$

uniformly over $P \in \mathcal{P}$, and the result continues to hold if σ^2 is replaced by $\widehat{\sigma}^2$. Moreover, the confidence regions based upon Double ML estimator $\check{\theta}_0$ have uniform asymptotic validity:

$$\lim_{N\to\infty} \sup_{P \in \mathcal{P}} \left| \mathrm{P}_P\left(\theta_0 \in [\tilde{\theta}_0 \pm \Phi^{-1}(1-\alpha/2)\widehat{\sigma}/\sqrt{N}]\right) - (1-\alpha) \right| = 0.$$

The scores ψ are the efficient scores, so both estimators are asymptotically efficient, reaching the semi-parametric efficiency bound of [30].

4. Empirical Example

To illustrate the methods developed in the preceding sections, we consider the estimation of the effect of 401(k) eligibility on accumulated assets. The key problem in determining the effect of 401(k) eligibility is that working for a firm that offers access to a 401(k) plan is not randomly assigned. To overcome the lack of random assignment, we follow the strategy developed in [45] and [46]. In these papers, the authors use data from the 1991 Survey of Income and Program Participation and argue that eligibility for enrolling in a 401(k) plan in this data can be taken as exogenous after conditioning on a few observables of which the most important for their argument is income. The basic idea of their argument is that, at least around the time 401(k)'s initially became available, people were unlikely to be basing their employment decisions on whether an employer offered a 401(k) but would instead focus on income and other aspects of the job. Following this argument, whether one is eligible for a 401(k) may then be taken as exogenous after appropriately conditioning on income and other control variables related to job choice.

A key component of the argument underlying the exogeneity of 401(k) eligibility is that eligibility may only be taken as exogenous after conditioning on income and other variables related to job choice that may correlate with whether a firm offers a 401(k). [45] and [46] and many subsequent papers adopt this argument but control only linearly for a small number of terms. One might wonder whether such specifications are able to adequately control for income and other related confounds. At the same time, the power to learn about treatment effects decreases as one allows more flexible models. The principled use of flexible machine learning tools offers one resolution to this tension. The results presented below thus complement previous results which rely on the assumption that confounding effects can adequately be controlled for by a small number of variables chosen *ex ante* by the researcher.

In the example in this paper, we use the same data as in [27]. We use net financial assets - defined as the sum of IRA balances, 401(k) balances, checking accounts, U.S. saving bonds, other interest-earning accounts in banks and other financial institutions, other interest-earning assets (such as bonds held personally), stocks, and mutual funds less non-mortgage debt - as the outcome variable, Y, in our analysis. Our treatment variable, D, is an indicator for being eligible to enroll in a 401(k) plan. The vector of raw covariates, Z, consists of age, income, family size, years of education, a married indicator, a two-earner status indicator, a defined benefit pension status indicator, an IRA participation indicator, and a home ownership indicator.

We report estimates of the average treatment effect (ATE) of 401(k) eligibility on net financial assets both in the partially linear model as in (1.1) and allowing for heterogeneous treatment effects using the approach outlined in Section 3.2 in Table 1. All results are based on sample-splitting as discussed in Section 2.1 using a 50-50 split. We report results based on four different methods for estimating the nuisance functions used in forming the orthogonal estimating equations. We consider three tree-based methods, labeled "Random Forest", "Regression Tree", and "Boosting", and one ℓ_1-penalization based method, labeled "Lasso". For "Regression Tree," we fit a single CART tree to estimate each nuisance function with penalty parameter chosen by 10-fold cross-validation. The results in column "Random Forest" are obtained by estimating each nuisance function with a random forest using default settings as in [17]. Results in "Boosting" are obtained using boosted regression trees with default settings from [48]. "Lasso" results use conventional lasso regression with penalty parameter chosen by 10-fold cross-validation to estimate conditional expectations of net financial assets and use ℓ_1-penalized logistic regression with penalty parameter chosen by 10-fold cross-validation when estimating conditional expectations of 401(k) eligibility. For the three tree-based methods, we use the raw set of covariates as features. For the ℓ_1-penalization based method, we use a set of 275 potential control variables formed from the raw set of covariates and all second order terms, i.e. squares and first-order interactions.

Turning to the results, it is first worth noting that the estimated effect ATE of 401(k) eligibility on net financial assets is \$19,559 with an estimated standard error of 1413 when no control variables are used. Of course, this number is not a valid estimate of the causal effect of 401(k) eligibility on financial assets if there are neglected confounding variables as suggested by [45] and [46]. When we turn to the estimates that flexibly account for confounding reported in Table 1, we see that they are substantially attenuated relative to this baseline that does not account for confounding, suggesting much smaller causal effects of 401(k) eligiblity on financial asset holdings. It is interesting and reassuring that the results obtained from the different flexible methods are broadly consistent with each other. This similarity is consistent with the theory that suggests that results obtained through the use of orthogonal estimating equations and any sensible method of estimating the necessary nuisance functions should be similar. Finally, it is interesting that these results are also broadly consistent with those reported in the original work of [45] and [46] which used a simple intuitively motivated functional form, suggesting that this intuitive choice was sufficiently flexible to capture much of the confounding variation in this example.

Appendix A. Proofs

In this appendix, we use C to denote a strictly positive constant that is independent of n and $P \in \mathcal{P}_n$. The value of C may change at each appearance. Also, the notation $a_n \lesssim b_n$ means that $a_n \leqslant C b_n$ for all n and some C. The notation $a_n \gtrsim b_n$ means that $b_n \lesssim a_n$. Moreover, the notation $a_n = o(1)$ means that there exists a sequence $(b_n)_{n \geqslant 1}$ of positive numbers such that (i) $|a_n| \leqslant b_n$ for all n, (ii) b_n is independent of $P \in \mathcal{P}_n$ for all n, and (iii) $b_n \to 0$ as $n \to \infty$. Finally, the notation $a_n = O_P(b_n)$ means that for all $\epsilon > 0$, there exists C such that $\mathrm{P}_P(a_n > C b_n) \leqslant 1 - \epsilon$ for all n. Using this notation allows us to avoid repeating "uniformly over $P \in \mathcal{P}_n$" many times in the proofs.

Table 1. Estimated ATE of 401(k) Eligibility on Net Financial Assets

	Random Forest	Lasso	Regression Tree	Boosting
A. Interactive Model				
ATE	8133	8734	8073	8405
	(1483)	(1168)	(1219)	(1193)
B. Partially Linear Model				
ATE	8845	9314	8805	8612
	(1204)	(1352)	(1379)	(1338)

Estimated ATE and heteroscedasticity robust standard errors (in parentheses) from a linear model (Panel B) and heterogeneous effect model (Panel A) based on orthogonal estimating equations. Column labels denote the method used to estimate nuisance functions. Further details about the methods are provided in the main text.

Define the empirical process $\mathbb{G}_n(\psi(W))$ as a linear operator acting on measurable functions $\psi : \mathcal{W} \to \mathbb{R}$ such that $\|\psi\|_{P,2} < \infty$ via,

$$\mathbb{G}_n(\psi(W)) := \mathbb{G}_{n,I}(\psi(W)) := \frac{1}{\sqrt{n}} \sum_{i \in I} f(W_i) - \int f(w) dP(w).$$

Analogously, we defined the empirical expectation as:

$$\mathbb{E}_n(\psi(W)) := \mathbb{E}_{n,I}(\psi(W)) := \frac{1}{\sqrt{n}} \sum_{i \in I} f(W_i).$$

Proof of Theorem 2.1. We split the proof into five steps.

Step 1. (Preliminary Rate Result). We claim that with probability $1 - o(1)$,

$$\|\check{\theta}_0 - \theta_0\| \lesssim B_{1n}\tau_{\pi n}.$$

By definition of $\check{\theta}_0$, we have

$$\Big\|\mathbb{E}_n[\psi(W, \check{\theta}_0, \widehat{\eta}_0)]\Big\| \leqslant \inf_{\theta \in \Theta} \Big\|\mathbb{E}_n[\psi(W, \theta, \widehat{\eta}_0)]\Big\| + \epsilon_n,$$

which implies via the triangle inequality that, with probability $1 - o(1)$,

$$\Big\| \ \mathrm{E}_P[\psi(W, \theta, \eta_0)]|_{\theta = \check{\theta}_0} \ \Big\| \leqslant \epsilon_n + 2I_1 + 2I_2 \lesssim B_{1n}\tau_{\pi n}, \quad \text{where} \tag{A.1}$$

$$I_1 := \sup_{\theta \in \Theta} \Big\|\mathbb{E}_n[\psi(W, \theta, \widehat{\eta}_0)] - \mathbb{E}_n[\psi(W, \theta, \eta_0)]\Big\| \lesssim B_{1n}\tau_{\pi n},$$

$$I_2 := \sup_{\theta \in \Theta} \Big\|\mathbb{E}_n[\psi(W, \theta, \eta_0)] - \mathrm{E}_P[\psi(W, \theta, \eta_0)]\Big\| \lesssim \tau_{\pi n}.$$

and the bounds on I_1 and I_2 are derived in Step 2 (note also that $\epsilon_n = o(\tau_{\pi n})$ by construction of the estimator and Assumption 2.2(vi)). Since by Assumption 2.1(iv), $2^{-1}\|J_0(\check{\theta}_{uj} - \theta_{uj})\| \wedge c_0$ does not exceed the left-hand side of (A.1), minimal singular values J_0 are bounded away from zero, and by Assumption 2.2(vi), $B_{1n}\tau_{\pi n} = o(1)$, we conclude that

$$\|\check{\theta}_0 - \theta_0\| \lesssim B_{1n}\tau_{\pi n}, \tag{A.2}$$

with probability $1 - o(1)$ yielding the claim of this step.

Step 2. (Bounds on I_1 and I_2) We claim that with probability $1 - o(1)$,

$$I_1 \lesssim B_{1n}\tau_{\pi n} \quad \text{and} \quad I_2 \lesssim \tau_{\pi n}.$$

To show these relations, observe that with probability $1-o(1)$, we have $I_1 \leqslant 2I_{1a} + I_{1b}$ and $I_2 \leqslant I_{1a}$, where

$$\begin{aligned} I_{1a} &:= \max_{\eta \in \{\eta_0, \widehat{\eta}_0\}} \sup_{\theta \in \Theta} \Big\| \mathbb{E}_n[\psi(W,\theta,\eta)] - \mathrm{E}_P[\psi(W,\theta,\eta)] \Big\|, \\ I_{1b} &:= \sup_{\theta \in \Theta, \eta \in \mathcal{T}} \Big\| \mathrm{E}_P[\psi(W,\theta,\eta)] - \mathrm{E}_P[\psi(W,\theta,\eta_0)] \Big\|. \end{aligned}$$

To bound I_{1b}, we employ Taylor's expansion:

$$I_{1b} \lesssim \max_{j \leqslant d_\theta} \sup_{\theta \in \Theta, \eta \in \mathcal{T}, r \in [0,1)} \Big| \partial_r \mathrm{E}_P \Big[\psi_j(W, \theta, \eta_0 + r(\eta - \eta_0)) \Big] \Big| \lesssim B_{1n} \sup_{\eta \in \mathcal{T}} \|\eta - \eta_0\|_e \leqslant B_{1n}\tau_{\pi n},$$

by Assumptions 2.1(v) and 2.2(ii).

To bound I_{1a}, we can apply the maximal inequality of Lemma A.1 to the function class $\mathcal{F}_{1,\eta}$ for $\eta = \eta_0$ and $\eta = \widehat{\eta}_0$ defined in Assumption 2.2, conditional on $(W_i)_{i \in I^c}$ so that $\widehat{\eta}_0$ is fixed after conditioning. Note that $(W_i)_{i \in I}$ are i.i.d. conditional on I^c. We conclude that with probability $1 - o(1)$,

$$I_{1a} \lesssim n^{-1/2}\Big(1 + n^{-1/2+1/q}\Big) \lesssim \tau_n. \tag{A.3}$$

Combining presented bounds gives the claim of this step.

Step 3. (Linearization) Here we prove the claim of the theorem. By definition of $\check{\theta}_0$, we have

$$\sqrt{n}\Big\| \mathbb{E}_n[\psi(W, \check{\theta}_0, \widehat{\eta}_0)] \Big\| \leqslant \inf_{\theta \in \Theta} \sqrt{n} \Big\| \mathbb{E}_n[\psi(W,\theta,\widehat{\eta}_0)] \Big\| + \epsilon_n \sqrt{n}. \tag{A.4}$$

Also, for any $\theta \in \Theta$ and $\eta \in \mathcal{T}$, we have

$$\begin{aligned} \sqrt{n}\mathbb{E}_n[\psi(W,\theta,\eta)] &= \sqrt{n}\mathbb{E}_n[\psi(W,\theta_0,\eta_0)] - \mathbb{G}_n\psi(W,\theta_0,\eta_0) \\ &\quad - \sqrt{n}\Big(\mathrm{E}_P[\psi(W,\theta_0,\eta_0)] - \mathrm{E}_P[\psi(W,\theta,\eta)]\Big) + \mathbb{G}_n\psi(W,\theta,\eta). \end{aligned} \tag{A.5}$$

Moreover, by Taylor's expansion of the function $r \mapsto \mathrm{E}_P[\psi(W, \theta_0 + r(\theta - \theta_0), \eta_0 + r(\eta - \eta_0))]$,

$$\begin{aligned} &\mathrm{E}_P[\psi(W,\theta,\eta)] - \mathrm{E}_P[\psi(W,\theta_0,\eta_0)] \\ &\quad = J_0(\theta - \theta_0) + \mathrm{D}_0[\eta - \eta_0] + \partial_r^2 \mathrm{E}_P[W, \theta_0 + r(\theta - \theta_0), \eta_0 + r(\eta - \eta_0)]\big|_{r = \bar{r}} \end{aligned} \tag{A.6}$$

for some $\bar r \in (0,1)$, which may differ for each row of the vector in the display. Substituting this equality into (A.5), taking $\theta = \check\theta_0$ and $\eta = \eta_0$, and using (A.4) gives

$$\sqrt{n}\Big\|\mathbb{E}_n[\psi(W,\theta_0,\eta_0)] + J_0(\check\theta_0 - \theta_0) + \mathrm{D}_0[\widehat\eta_0 - \eta_0]\Big\| \\ \leqslant \epsilon_n\sqrt{n} + \inf_{\theta\in\Theta}\sqrt{n}\|\mathbb{E}_n[\psi(W,\theta,\widehat\eta_0)]\| + II_1 + II_2, \tag{A.7}$$

where

$$II_1 := \sqrt{n}\sup_{r\in[0,1)}\Big\|\ \partial_r^2\mathrm{E}_P\Big[\psi(W,\theta_0 + r(\theta-\theta_0),\eta_0 + r\{\eta-\eta_0\})\Big|_{\theta=\check\theta_0,\eta=\widehat\eta_0}\ \Big\|,$$

$$II_2 := \Big\|\ \mathbb{G}_n\Big(\psi(W,\theta,\eta) - \psi(W,\theta_0,\eta_0)\Big)\Big|_{\theta=\check\theta_0,\eta=\widehat\eta_0}\ \Big\|.$$

It will be shown in Step 4 that

$$II_1 + II_2 = O_P(\delta_n). \tag{A.8}$$

In addition, it will be shown in Step 5 that

$$\inf_{\theta\in\Theta}\sqrt{n}\|\mathbb{E}_n[\psi(W,\theta,\widehat\eta_0)]\| = O_P(\delta_n). \tag{A.9}$$

Moreover, $\epsilon_n\sqrt{n} = o(\delta_n)$ by construction of the estimator. Therefore, the expression in (A.7) is $O_P(\delta_n)$. Further,

$$\|\mathrm{D}_0[\widehat\eta_0 - \eta_0]\| = 0$$

by the orthogonality condition since $\widehat\eta_0 \in \mathcal{T}_0$ with probability $1 - o(1)$ by Assumption 2.2(i). Therefore, Assumption 2.1(iv) gives

$$\|J_0^{-1}\sqrt{n}\mathbb{E}_n[\psi(W,\theta_0,\eta_0)] + \sqrt{n}(\check\theta_0 - \theta_0)\| = O_P(\delta_n).$$

The asserted claim now follows by multiplying both parts of the display by $\Sigma_0^{-1/2}$ (under the supremum on the left-hand side) and noting that singular values of Σ_0 are bounded from below and from above by Assumptions 2.1(iv) and 2.2(v).

Step 4. (Bounds on II_1 and II_2). Here we prove (A.8). First, with probability $1 - o(1)$,

$$II_1 \lesssim \sqrt{n}B_{2n}\|\check\theta_0 - \theta_0\|^2 \vee \|\widehat\eta_0 - \eta_0\|_e^2 \lesssim \sqrt{n}B_{1n}^2B_{2n}\tau_{\pi n}^2 \lesssim \delta_n,$$

where the first inequality follows from Assumptions 2.1(v) and 2.2(i), the second from Step 1 and Assumptions 2.2(ii) and 2.2(vi), and the third from Assumption 2.2(vi).

Second, with probability $1 - o(1)$,

$$II_2 \lesssim \sup_{f\in\mathcal{F}_2}|\mathbb{G}_n(f)|$$

where

$$\mathcal{F}_2 = \Big\{\psi_j(\cdot,\theta,\widehat\eta_0) - \psi_j(\cdot,\theta_0,\eta_0)\colon j = 1,...,d_\theta, \|\theta - \theta_0\| \leqslant CB_{1n}\tau_{\pi n}\Big\}$$

for sufficiently large constant C. To bound $\sup_{f \in \mathcal{F}_2} |\mathbb{G}_n(f)|$, we apply Lemma A.1. Observe that

$$\begin{aligned} \sup_{f \in \mathcal{F}_2} \|f\|_{P,2}^2 &\leqslant \sup_{j \leqslant d_\theta, \|\theta - \theta_0\| \leqslant C B_{1n}\tau_{\pi n}, \eta \in \mathcal{T}} \mathrm{E}_P \left[|\psi_j(W, \theta, \eta) - \psi_j(W, \theta_0, \eta_0)|^2 \right] \\ &\leqslant \sup_{j \leqslant d_\theta, \|\theta - \theta_0\| \leqslant C B_{1n}\tau_{\pi n}, \eta \in \mathcal{T}} C_0 (\|\theta - \theta_0\| \vee \|\eta - \eta_0\|_e)^\omega \lesssim (B_{1n}\tau_{\pi n})^\omega, \end{aligned}$$

where we used Assumption 2.1(v) and Assumption 2.2(ii). An application of Lemma A.1 to the empirical process $\{\mathbb{G}_n(f), f \in \mathcal{F}_2\}$ with an envelope $F_2 = 2F_{1,\widehat{\eta}_0}$ and $\sigma = (CB_{1n}\tau_{\pi n})^{\omega/2}$, conditionally on $(W_i)_{i \in I^c}$, so that $\widehat{\eta}_0$ can be treated as fixed, yields that with probability $1 - o(1)$,

$$\sup_{f \in \mathcal{F}_2} |\mathbb{G}_n(f)| \lesssim (B_{1n}\tau_{\pi n})^{\omega/2} + n^{-1/2+1/q}, \tag{A.10}$$

since $\sup_{f \in \mathcal{F}_2} |f| \leqslant 2 \sup_{f \in \mathcal{F}_{1,\widehat{\eta}_0}} |f| \leqslant 2F_{1,\widehat{\eta}_0}$ and $\|F_{1,\widehat{\eta}_0}\|_{P,q} \leqslant K$ by Assumption 2.2(iv) and

$$\log \sup_Q N(\epsilon \|F_2\|_{Q,2}, \mathcal{F}_2, \|\cdot\|_{Q,2}) \leqslant 2v \log(2a/\epsilon), \quad \text{for all } 0 < \epsilon \leqslant 1$$

because $\mathcal{F}_2 \subset \mathcal{F}_{1,\widehat{\eta}_0} - \mathcal{F}_{1,\widehat{\eta}_0}$ for $\mathcal{F}_{1,\eta}$ defined in Assumption 2.2(iv), so that

$$\log \sup_Q N(\epsilon \|F_2\|_{Q,2}, \mathcal{F}_2, \|\cdot\|_{Q,2}) \leqslant 2 \log \sup_Q N((\epsilon/2)\|F_{1,\widehat{\eta}_0}\|_{Q,2}, \mathcal{F}_{1,\widehat{\eta}_0}, \|\cdot\|_{Q,2})$$

by a standard argument. The claim of this step now follows from an application of Assumption 2.2(vi) to bound the right-hand side of (A.10).

Step 5. Here we prove (A.9). Let $\bar{\theta}_0 = \theta_0 - J_0^{-1}\mathbb{E}_n[\psi(W, \theta_0, \eta_0)]$. Then $\|\bar{\theta}_0 - \theta_0\| = O_P(1/\sqrt{n})$ since $\mathrm{E}_P[\|\sqrt{n}\mathbb{E}_n[\psi(W, \theta_0, \eta_0)]\|]$ is bounded and J_0 is bounded in absolute value below by Assumption 2.1(iv). Therefore, $\check{\theta}_0 \in \Theta$ with probability $1 - o(1)$ by Assumption 2.1(i). Hence, with the same probability,

$$\inf_{\theta \in \Theta} \sqrt{n} \Big\| \mathbb{E}_n[\psi(W, \theta, \widehat{\eta}_0)] \Big\| \leqslant \sqrt{n} \Big\| \mathbb{E}_n[\psi(W, \bar{\theta}_0, \widehat{\eta}_0)] \Big\|,$$

and so it suffices to show that

$$\sqrt{n} \Big\| \mathbb{E}_n[\psi(W, \bar{\theta}_0, \widehat{\eta}_0)] \Big\| = O_P(\delta_n). \tag{A.11}$$

To prove (A.11), substitute $\theta = \bar{\theta}_0$ and $\eta = \widehat{\eta}$ into (A.5) and use Taylor's expansion in (A.6). This gives

$$\sqrt{n} \Big\| \mathbb{E}_n[\psi(W, \bar{\theta}_0, \widehat{\eta}_0)] \Big\| \leqslant \sqrt{n} \Big\| \mathbb{E}_n[\psi(W, \theta_0, \eta_0)] + J_0(\bar{\theta}_0 - \theta_0) + \mathrm{D}_0[\widehat{\eta}_0 - \eta_0] \Big\| + \widetilde{II}_1 + \widetilde{II}_2$$

where $\widetilde{II}_1$ and $\widetilde{II}_2$ are defined as II_1 and II_2 in Step 3 but with $\check{\theta}_0$ replaced by $\bar{\theta}_0$. Then, given that $\|\bar{\theta}_0 - \theta_0\| \lesssim \log n / \sqrt{n}$ with probability $1 - o(1)$, the argument in Step 4 shows that

$$\widetilde{II}_1 + \widetilde{II}_2 = O_P(\delta_n).$$

In addition,

$$\mathbb{E}_n[\psi(W, \theta_0, \eta_0)] + J_0(\bar{\theta}_0 - \theta_0) = 0$$

by the definition of $\bar{\theta}_0$, and

$$\|\mathrm{D}_0[\widehat{\eta}_0 - \eta_0]\| = 0$$

by the orthogonality condition. Combining these bounds gives (A.11), so that the claim of this step follows, and completes the proof of the theorem. ■

A.1. **Useful Lemmas.** Let $(W_i)_{i=1}^n$ be a sequence of independent copies of a random element W taking values in a measurable space $(\mathcal{W}, \mathcal{A}_{\mathcal{W}})$ according to a probability law P. Let $\mathcal{F}$ be a set of suitably measurable functions $f\colon \mathcal{W} \to \mathbb{R}$, equipped with a measurable envelope $F\colon \mathcal{W} \to \mathbb{R}$.

Lemma A.1 (Maximal Inequality, [23])**.** *Work with the setup above. Suppose that $F \geqslant \sup_{f \in \mathcal{F}} |f|$ is a measurable envelope for $\mathcal{F}$ with $\|F\|_{P,q} < \infty$ for some $q \geqslant 2$. Let $M = \max_{i \leqslant n} F(W_i)$ and $\sigma^2 > 0$ be any positive constant such that $\sup_{f \in \mathcal{F}} \|f\|_{P,2}^2 \leqslant \sigma^2 \leqslant \|F\|_{P,2}^2$. Suppose that there exist constants $a \geqslant e$ and $v \geqslant 1$ such that*

$$\log \sup_Q N(\epsilon \|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}) \leqslant v \log(a/\epsilon), \ \ 0 < \epsilon \leqslant 1.$$

Then

$$\mathrm{E}_P[\|\mathbb{G}_n\|_{\mathcal{F}}] \leqslant K \left(\sqrt{v\sigma^2 \log\left(\frac{a\|F\|_{P,2}}{\sigma}\right)} + \frac{v\|M\|_{P,2}}{\sqrt{n}} \log\left(\frac{a\|F\|_{P,2}}{\sigma}\right) \right),$$

where K is an absolute constant. Moreover, for every $t \geqslant 1$, with probability $> 1 - t^{-q/2}$,

$$\|\mathbb{G}_n\|_{\mathcal{F}} \ \leqslant \ (1+\alpha)\mathrm{E}_P[\|\mathbb{G}_n\|_{\mathcal{F}}] + K(q)\Big[(\sigma + n^{-1/2}\|M\|_{P,q})\sqrt{t} + \alpha^{-1} n^{-1/2} \|M\|_{P,2} t\Big], \ \ \forall \alpha \ > \ 0,$$

where $K(q) > 0$ is a constant depending only on q. In particular, setting $a \geqslant n$ and $t = \log n$, with probability $> 1 - c(\log n)^{-1}$,

$$\|\mathbb{G}_n\|_{\mathcal{F}} \leqslant K(q,c) \left(\sigma \sqrt{v \log\left(\frac{a\|F\|_{P,2}}{\sigma}\right)} + \frac{v\|M\|_{P,q}}{\sqrt{n}} \log\left(\frac{a\|F\|_{P,2}}{\sigma}\right) \right), \tag{A.12}$$

where $\|M\|_{P,q} \leqslant n^{1/q} \|F\|_{P,q}$ and $K(q,c) > 0$ is a constant depending only on q and c.

References

[1] Ai, C. and Chen, X. (2012). The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics*, 170(2):442–457.

[2] Andrews, D.W.K. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica*, 62(1):43–72.

[3] Belloni, A., Chen, D, Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80:2369–2429. ArXiv, 2010.

[4] Belloni, A. and Chernozhukov, V. (2011). ℓ_1-penalized quantile regression for high dimensional sparse models. *Annals of Statistics*, 39(1):82–130. ArXiv, 2009.

[5] Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547. ArXiv, 2009.

[6] Belloni, A., Chernozhukov, V., and Hansen, C. (2010). Lasso methods for gaussian instrumental variables models. ArXiv, 2010.

[7] Belloni, A., Chernozhukov, V., and Hansen, C. (2013). Inference for high-dimensional sparse econometric models. *Advances in Economics and Econometrics. 10th World Congress of Econometric Society. August 2010*, III:245–295. ArXiv, 2011.

[8] Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection amongst high-dimensional controls. *Review of Economic Studies*, 81:608–650. ArXiv, 2011.

[9] Belloni, A., Chernozhukov, V., and Kato, K. (2013). Robust inference in high-dimensional approximately sparse quantile regression models. ArXiv, 2013.

[10] Belloni, A., Chernozhukov, V., and Kato, K. (2015). Uniform post selection inference for LAD regression models and other Z-estimators. *Biometrika*, (102):77–94. ArXiv, 2013.

[11] Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root-lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806. Arxiv, 2010.

[12] Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2013). Program evaluation with high-dimensional data. ArXiv, 2013; to appear in *Econometrica*.

[13] Belloni, A., Chernozhukov, V., and Wang, L. (2014). Pivotal estimation via square-root lasso in nonparametric regression. *The Annals of Statistics*, 42(2):757–788. ArXiv, 2013.

[14] Bickel, P. J. (1982). On Adaptive Estimation. *Annals of Statistics*, 10(3):647–671.

[15] Bickel, P. J., Klaassen, C. A. J., Ritov, Y., Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models.* Springer.

[16] Bickel, P. J., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732. ArXiv, 2008.

[17] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

[18] Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica*, 60:567–596.

[19] Chen, X., Linton, O. and van Keilegom, I. (2003). Estimation of Semiparametric Models when the Criterion Function Is Not Smooth. *Econometrica*, 71(5):1591–1608.

[20] Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819. ArXiv, 2012.

[21] Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 42(5):1787–1818. ArXiv, 2013.

[22] Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Central limit theorems and bootstrap in high dimensions. ArXiv, 2014. *The Annals of Probability* (to appear).

[23] Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564–1597. ArXiv, 2012.

[24] Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Comparison and anti-concentration bounds for maxima of gaussian random vectors. *Probability Theory and Related Fields*, 162:47–70.

[25] Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related gaussian couplings. ArXiv, 2015; to appear, *Stochastic Processes and Applications.*

[26] Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013). Inference on counterfactual distributions. *Econometrica*, 81:2205–2268.

[27] Chernozhukov, V. and Hansen, C. The impact of 401(k) participation on the wealth distribution: An instrumental quantile regression analysis *Review of Economics and Statistics*, 86:735–751.

[28] Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Post-selection and post-regularization inference in linear models with very many controls and instruments. *Americal Economic Review: Papers and Proceedings*, 105:486–490.

[29] Dudley, R. (1999). *Uniform central limit theorems*, volume 63 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge.

[30] Hahn, J. (1998) "On the role of the propensity score in efficient semiparametric estimation of average treatment effects." *Econometrica* (1998): 315-331.

[31] Javanmard, A. and Montanari, A. (2014). Hypothesis testing in high-dimensional regression under the gaussian random design model: asymptotic theory. *IEEE Transactions on Information Theory*, 60:6522–6554. ArXiv, 2013.

[32] Jing, B.-Y., Shao, Q.-M., and Wang, Q. (2003). Self-normalized Cramer-type large deviations for independent random variables. *Ann. Probab.*, 31(4):2167–2215.

[33] Kosorok, M. (2008). *Introduction to Empirical Processes and Semiparametric Inference.* Series in Statistics. Springer, Berlin.

[34] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces (Isoperimetry and processes).* Ergebnisse der Mathematik undihrer Grenzgebiete, Springer-Verlag.

[35] Leeb, H. and Pötscher, B. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(2):338–376.

[36] Leeb, H. and Pötscher, B. (2008). Recent developments in model selection and related areas. *Econometric Theory*, 24(2):319–322.

[37] Leeb, H. and Pötscher, B. (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *J. Econometrics*, 142(1):201–211.

[38] Linton, O. (1996). Edgeworth approximation for MINPIN estimators in semiparametric regression models. *Econometric Theory*, 12(1):30–60.

[39] Negahban, S., Ravikumar, P., Wainwright, P., and Yu, B. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557. ArXiv, 2010.

[40] Newey, W. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2):99–135.

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

[42] Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. *Probability and Statistics, the Harold Cramer Volume.*

[43] Neyman, J. (1979). $c(\alpha)$ tests and their use. *Sankhya*, 41:1–21.

[44] Pisier, G. (1999). *The volume of convex bodies and Banach space geometry*, volume 94. Cambridge University Press.

[45] Poterba, J. M., Venti, S. F., and Wise, D. A. (1994). 401(k) plans and tax-deferred savings. In Wise, D., ed., *Studies in the Economics of Aging.* Chicago: University of Chicago Press, 105–142.

[46] Poterba, J. M., Venti, S. F., and Wise, D. A. (1994). Do 401(k) contributions crowd out other personal saving?. *Journal of Public Economics*, 58:1–32.

[47] Pötscher, B. and Leeb, H. (2009). On the distribution of penalized maximum likelihood estimators: the LASSO, SCAD, and thresholding. *J. Multivariate Anal.*, 100(9):2065–2082.

[48] Ridgeway, G. (2006). Generalized boosted regression models. *Documentation on the R Package gbm, version 1.57.*

[49] Robins, J. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.*, 90(429):122–129.

[50] Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica*, 56(4):931–954.

[51] Rudelson, M. and Vershynin, R. (2008). On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61:1025–1045.

[52] van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42:1166–1202. ArXiv, 2013.

[53] van der Laan, M. J. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data.* Springer.

[54] van der Vaart, A. W. (1991). On Differentiable Functionals. *Annals of Statistics*, 19(1):178–204.

[55] van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press.

[56] van der Vaart, A. W. and Wellner, J. (1996). *Weak Convergence and Empirical Processes*. Springer Series in Statistics.

[57] Zhang, C.-H, and Zhang, S. (2014). Confidence intervals for low-dimensional parameters with high-dimensional data. *J. R. Statist. Soc. B*, 76:217–242. ArXiv, 2012.