The Econometrics Journal

Econometrics Journal (2021), volume **24**, pp. 264–289.
https://doi.org/10.1093/ectj/utaa027

Debiased machine learning of conditional average treatment effects and other causal functions

VIRA SEMENOVA[†] AND VICTOR CHERNOZHUKOV[‡]

[†]*Department of Economics, University of California, Berkeley 530 Evans Hall Berkeley, California 94720.*
Email: semenovavira@gmail.com

[‡]*MIT Department of Economics, 50 Memorial Drive, Cambridge, MA 02142.*
Email: vchern@mit.edu

First version received: 28 April 2020; final version accepted: 18 August 2020.

Summary: This paper provides estimation and inference methods for the best linear predictor (approximation) of a structural function, such as conditional average structural and treatment effects, and structural derivatives, based on modern machine learning tools. We represent this structural function as a conditional expectation of an unbiased signal that depends on a nuisance parameter, which we estimate by modern machine learning techniques. We first adjust the signal to make it insensitive (Neyman-orthogonal) with respect to the first-stage regularisation bias. We then project the signal onto a set of basis functions, which grow with sample size, to get the best linear predictor of the structural function. We derive a complete set of results for estimation and simultaneous inference on all parameters of the best linear predictor, conducting inference by Gaussian bootstrap. When the structural function is smooth and the basis is sufficiently rich, our estimation and inference results automatically target this function. When basis functions are group indicators, the best linear predictor reduces to the group average treatment/structural effect, and our inference automatically targets these parameters. We demonstrate our method by estimating uniform confidence bands for the average price elasticity of gasoline demand conditional on income.

Keywords: *High-dimensional statistics*, *heterogeneous treatment effect*, *conditional average treatment effect*, *group average effects*, *debiased/orthogonal estimation*, *machine learning*, *double robustness*, *continuous treatment effects*, *dose–response functions*.

JEL codes: *C14*, *C55*.

1. INTRODUCTION AND MOTIVATION

This paper gives a method for estimating and conducting inference on a nonparametric function $g(x)$ that summarises heterogeneous treatment/causal/structural effects conditional on a small set of covariates X. We assume that this function can be represented as a conditional expectation function

$$g(x) = \mathbb{E}[Y(\eta_0)|X = x], \tag{1.1}$$

where the random variable $Y(\eta_0)$, which we refer to as a signal, depends on a nuisance function $\eta_0(z)$ of a (potentially very) high-dimensional control vector Z. Examples of the nonparametric

target function include the conditional average treatment effect (CATE), continuous treatment effects (CTEs), as well as many others discussed below. Examples of the nuisance functions

$$\eta_0 = \eta_0(z)$$

include the propensity score, the conditional density, and the regression function, among others. In summary,

$$\dim(Z) \text{ is high;} \quad \dim(X) \text{ is low.}$$

Although there are many possible choices of signals $Y(\eta)$, we focus on signals that have the orthogonality property (Neyman, 1959). Formally, we require the pathwise derivative of the conditional expectation to be zero conditional on X:

$$\partial_r \mathbb{E}[Y(\eta_0 + r(\eta - \eta_0))|X = x]|_{r=0} = 0, \quad \text{for all } x \text{ and } \eta. \tag{1.2}$$

If the signal $Y(\eta)$ is orthogonal, its plug-in estimate $Y(\widehat{\eta})$ is insensitive to bias in the estimation of $\widehat{\eta}$, which results from applying modern adaptive learning methods in high dimensions. Under mild conditions, $Y(\widehat{\eta})$ delivers a high-quality estimator of the target function $g(x)$.

We demonstrate the importance of the orthogonality property for CTEs, studied in Imbens (2000), Gill and Robins (2001), and Kennedy et al. (2017). Let $X \in \mathbb{R}$ be a one-dimensional continuous treatment, and let Y^x be the potential outcome corresponding to the subject's response after receiving x units of treatment. The observed data vector $V = (X, Z, Y)$ consists of the treatment X, the control vector Z, and the observed outcome $Y = Y^X$. If potential outcomes $\{Y^x, x \in \mathbb{R}\}$ are independent of treatment X conditional on controls Z, the average potential outcome is identified as

$$\mathbb{E}[Y^x] = \mathbb{E}\mu_0(x, Z) = \int \mu_0(x, z) d P_Z(z), \tag{1.3}$$

where $\mu_0(x, z) = \mathbb{E}[Y|X = x, Z = z]$ is the regression function of the observed outcome. Because the control vector Z is high dimensional, it is necessary to estimate the regression function $\mu_0(x, z)$ with some regularised technique to achieve convergence.

A naive approach to estimate $\mathbb{E}[Y^x]$ is to consider a sample average

$$\widetilde{g}(x) = \int \widehat{\mu}(x, z) d \widehat{P}_Z(z),$$

where $\widehat{\mu}(x, z)$ is a regularised estimate of $\mu_0(x, z)$, and $\widehat{P}_Z$ is the empirical analog of P_Z. This approach results in a biased estimate, and the bias of estimation error, $\widehat{\mu}(x, Z) - \mu_0(x, Z)$, does not vanish faster than $N^{-1/2}$. The plug-in estimator inherits this first-order bias, because the moment equation (1.3) is not orthogonal to perturbations of μ:

$$\partial_r \mathbb{E}[(\mu_0 + r(\mu - \mu_0)) \circ (x, Z)]|_{r=0} = \mathbb{E}[\mu(x, Z) - \mu_0(x, Z)] \neq 0.$$

This bias implies that the plug-in estimator $\widetilde{g}$ will not converge at the optimal rate.

To deliver a high-quality estimate of $\mathbb{E}[Y^x]$, we represent

$$g(x) = \mathbb{E}[Y^x]$$

as a special case of the signal framework (1.1). We choose the signal $Y(\eta)$ to be the doubly robust signal from Kennedy et al. (2017):

$$Y(\eta) := \frac{Y - \mu(X, Z)}{s(X|Z)} w(X) + \int \mu(X, z) d P_Z(z), \tag{1.4}$$

where the nuisance parameter

$$\eta_0(x, z) = \{s_0(x|z), \mu_0(x, z), w_0(x)\}$$

consists of the regression function $\mu_0(x, z)$, the conditional density $s_0(x|z)$ of X given Z, and the marginal treatment density $w_0(x)$. This procedure is more costly because the nuisance parameter now includes two more functions, $s_0(x|z)$ and $w_0(x)$, in addition to $\mu_0(x, z)$. However, the signal in (1.4) has the benefit of being conditionally orthogonal with respect to each nuisance function in $\eta_0(x, z)$:

$$\mathbb{E}\left[\begin{array}{l} -\int_{z\in\mathcal{Z}}(\mu(X, z) - \mu_0(X, z))dP_Z(z) + \int_{z\in\mathcal{Z}}(\mu(x, z) - \mu_0(x, z))dP_Z(z) \\ \dfrac{\mu_0(X, Z) - Y}{s_0^2(X|Z)}(s(X|Z) - s_0(X|Z)) \\ \dfrac{Y - \mu_0(X, Z)}{s_0(X|Z)}(w(X) - w_0(X)) \end{array}\middle| X = x\right] = 0.$$

Because this signal is conditionally orthogonal to the nuisance function, the bias of the estimation error, $\widehat{\eta}(X, Z) - \eta_0(X, Z)$, does not create first-order bias in the estimated signal $Y(\widehat{\eta})$ and affects only its higher-order bias. As a result, the estimate of the target function based on $Y(\widehat{\eta})$ is high quality under plausible conditions.

In the second stage, we consider a linear projection of an orthogonal signal $Y(\eta)$ onto a vector of basis functions $p(X)$,

$$\beta := \arg\min_{b\in\mathbb{R}^d} \mathbb{E}(Y(\eta) - p(X)'b)^2.$$

The choice of basis functions depends on the desired interpretation of the linear approximation. For example, consider partitioning the support of X into d mutually exclusive groups $\{G_k\}_{k=1}^d$. Setting

$$p_k(x) = \mathbf{1}\{x \in G_k\}, \quad k \in \{1, 2, \ldots, d\}$$

implies that $p(x)'\beta_0$ is a group average treatment effect for group k such that $x \in G_k$. Our inference will target this parameter, allowing the number of groups to increase at some rate.

For another example, let $p(x) \in \mathbb{R}^d$ be a d-dimensional dictionary of series/sieve basis functions, e.g., polynomials, splines, or wavelets. Then, $p(x)'\beta_0$ corresponds to the best linear approximation to the target function $g(x)$ in the given dictionary. Under some smoothness conditions, as the dimension of the dictionary becomes large, $p(x)'\beta_0$ will approximate $g(x)$, and our inference will target this function. We derive a complete set of results for estimation and simultaneous inference on all parameters of the best linear predictor, conducting inference by Gaussian bootstrap. When the structural function is smooth and the basis is sufficiently rich, our estimation and inference results automatically target this function. When basis functions are group indicators, the best linear predictor reduces to the group average treatment/structural effect, and our inference automatically targets these parameters.

1.1. Literature review

This paper builds on three bodies of research within the semiparametric literature: orthogonal (debiased) machine learning, least-squares series estimation, and treatment effects / missing data problems. Orthogonal machine learning (Chernozhukov et al., 2016; Chernozhukov et al.,

2018) proposes inference on a fixed-dimensional target parameter β_0 in the presence of a high-dimensional nuisance function η in a semiparametric moment problem. If the moment condition is orthogonal to perturbations of η, estimating η by machine learning methods has no first-order effect on the asymptotic distribution of the target parameter β_0. In particular, plugging in an estimate of η obtained on a separate sample results in a $\sqrt{N}$-consistent asymptotically normal estimate whose asymptotic variance is the same as if the econometrician knew $\eta = \eta_0$. This result makes it possible to use highly complex machine learning methods to estimate the nuisance function η, such as ℓ_1 penalised methods in sparse models (Bühlmann and van der Geer, 2011; Belloni et al., 2016; ℓ_2 boosting in sparse linear models Luo and Spindler, 2016), and other methods for classes of neural nets, regression trees, and random forests. The present paper extends the orthogonal machine learning literature by allowing the target parameter to be a function—that is, an infinite-dimensional parameter. Next, our paper contributes to a large body of work on "debiased" inference for parameters after regularisation or model selection (Belloni et al., 2017; Belloni et al., 2014; Belloni et al., 2016; van der Geer et al., 2014; Javanmard and Montanari, 2014; Zhang and Zhang, 2014), with the crucial difference being that our (ultimate) target parameter $g(x)$ is infinite dimensional, whereas in those papers the target parameter is finite dimensional.

The second building block of our paper is the literature on least-squares series estimation (Newey, 2007; Newey, 2009; Belloni et al., 2015; Chen and Christensen, 2015), which establishes pointwise and uniform limit theory for least-squares series estimation. We extend this theory by allowing the dependent variable of the series projection to depend on an unknown nuisance parameter η. We show that series properties continue to hold without any additional strong assumptions on the problem design.

Finally, we also contribute to the literature on estimating the CATEs and group average treatment effects with missing data (Robins and Rotnitzky, 1995; Hahn, 1998; Graham, 2011; Graham et al., 2012; Hirano et al., 2003; Abrevaya et al., 2015; Athey and Imbens, 2016; Grimmer et al., 2017; Oprescu et al., 2018, among others). After we released the working paper version of the present article (Semenova and Chernozhukov, 2018), many methods (Jacob, 2019; Fan et al., 2019; Zimmert and Lechner, 2019; Colangelo and Lee, 2020) have been proposed for estimating group, incremental, or heterogeneous treatment effects in the presence of high-dimensional controls. Our framework covers many more examples than just CATEs or CTEs and uses series estimators, as opposed to kernels, to localise the structural function.

In a related paper, Chernozhukov et al. (2017) studied CATEs in randomised control trials with a known propensity score. Recognising a widespread interest in estimating CATE by modern machine learning techniques, Chernozhukov et al. (2017) studies the best linear projection of the true CATE function onto an arbitrary machine learning estimator of CATE under consideration, constructed on an auxiliary sample. The analysis of Chernozhukov et al. (2017) does not require any assumptions on the machine learning estimator; however, that paper targets a specific feature of CATE—the best linear projection of CATE—rather than CATE itself. In contrast, our work operates in a classic observational setting, with many potential controls, and targets the true CATE function. In our setting, modern regularised methods are used to estimate the propensity score, but they are required to approximate this parameter sufficiently well. To sum up, the present paper delivers a sharper characterisation of CATE in a more challenging setting, in exchange for stronger assumptions about the first-stage machine learning estimate.

2. SETUP

2.1. Examples

In this section, we describe our main examples. For each example, we provide an orthogonal signal $Y(\eta)$ obeying (1.1)–(1.2).

EXAMPLE 2.1 (CONTINUOUS TREATMENT EFFECTS). Let $X \in \mathbb{R}$ be a continuous treatment variable, Z be a vector of the controls, Y^x stand for the potential outcomes corresponding to the subject's response after receiving x units of treatment, and $Y = Y^X$ be the observed outcome. The observed data V is $V = (X, Z, Y)$. For a given value x, the target function is the average potential outcome

$$g(x) = \mathbb{E}[Y^x].$$

A standard way to identify the function $g(x)$ is to assume unconfoundedness. Suppose all of the potential outcomes $\{Y^x, x \in \mathbb{R}\}$ are independent of X conditional on Z,

$$\{Y^x, x \in \mathbb{R}\} \perp X|Z.$$

Then, $g(x)$ is identified as

$$g(x) = \mathbb{E}\mu_0(x, Z),$$

where $\mu_0(x, z) = \mathbb{E}[Y|X = x, Z = z]$ is the regression function. Lemma 4.3 shows that the doubly robust signal from Kennedy et al. (2017),

$$Y(\eta) := \frac{Y - \mu(X, Z)}{s(X|Z)} w(X) + \int \mu(X, z) dP_Z(z), \tag{2.1}$$

is conditionally orthogonal with respect to the nuisance parameter

$$\eta_0(x, z) := \{s_0(x|z), \mu_0(x, z), w_0(x)\},$$

consisting of the conditional treatment density (a.k.a. generalised propensity score)

$$s_0(x|z) = \left. \frac{dP(X \leq t|Z = z)}{dt} \right|_{t=x},$$

regression function $\mu_0(x, z) = \mathbb{E}[Y|X = x, Z = z]$, and the marginal treatment density

$$w_0(x) = \left. \frac{d\mathrm{P}(X \leq t)}{dt} \right|_{t=x} = \mathbb{E}_Z s_0(x|Z).$$

Theorems 4.6 and 4.7 establish pointwise and uniform asymptotic normality for the orthogonal estimator of CTEs.

In the examples below, the vector X represents a low-dimensional subset of potential controls Z.

EXAMPLE 2.2 (CONDITIONAL AVERAGE TREATMENT EFFECT). Let Y_1 and Y_0 be the potential outcomes corresponding to a subject's response with and without receiving a binary treatment, respectively. Let $D = 1$ be a dummy for whether a subject is treated. The object of interest is the CATE

$$g(x) := \mathbb{E}[Y_1 - Y_0|X = x].$$

Because an individual cannot be treated and untreated at the same time, the econometrician observes only the actual outcome $Y = DY_1 + (1 - D)Y_0$ but not the treatment effect $Y_1 - Y_0$.

A standard way to make progress in this problem is to assume unconfoundedness (Rosenbaum and Rubin, 1983). Suppose there exists an observable control vector Z such that treatment status D is independent of the potential outcomes Y_1, Y_0 conditional on Z,

$$Y_1, Y_0 \perp D|Z.$$

Define the conditional probability of treatment receipt as $s_0(z) = \mathrm{P}(D = 1|Z = z)$. Consider a Robins and Rotnitzky (1995)–type orthogonal signal $Y(\eta)$,

$$Y(\eta) := \mu(1, Z) - \mu(0, Z) + \frac{D[Y - \mu(1, Z)]}{s(Z)} - \frac{(1 - D)[Y - \mu(0, Z)]}{1 - s(Z)}, \tag{2.2}$$

where $\mu_0(d, z) = \mathbb{E}[Y|D = d, Z = z]$ is the conditional expectation function of Y. Corollary 4.1 shows that (2.2) is orthogonal with respect to the nuisance parameter $\eta_0(z) := \{s_0(z), \mu_0(1, z), \mu_0(0, z)\}$ and establishes pointwise and uniform asymptotic theory for the orthogonal estimator of CATE.

EXAMPLE 2.3 (REGRESSION FUNCTION WITH PARTIALLY MISSING OUTCOME). Suppose a researcher is interested in the conditional expectation of a variable Y^* given X:

$$g(x) := \mathbb{E}[Y^*|X = x],$$

where Y^* is partially missing. Let $D = 1$ be a dummy for whether the outcome Y^* is observed, Z be a control vector, $Y = DY^*$ be the observed outcome, and $V = (D, Z, Y)$ be the data vector. Because the researcher does not control D, a standard way to make progress is to assume there exists an observable control vector Z such that Y^* is independent of D given Z,

$$Y^* \perp D|Z.$$

Corollary 4.2 shows that the signal $Y(\eta)$, defined as

$$Y(\eta) := \mu(Z) + \frac{D[Y - \mu(Z)]}{s(Z)}, \tag{2.3}$$

is orthogonal with respect to the nuisance parameter

$$\eta_0(z) := \{s_0(z), \mu_0(z)\},$$

where $\mu_0(z) = \mathbb{E}[Y|Z = z, D = 1]$ is the conditional expectation function of the observed outcome Y.

EXAMPLE 2.4 (CONDITIONAL AVERAGE PARTIAL DERIVATIVE). Let $D \in \mathbb{R}$ be a continuous treatment variable, Z be a vector of the controls, Y^d stand for the potential outcomes corresponding to the subject's response after receiving d units of treatment, $Y = Y^D$ be the observed outcome, and $V = (D, Z, Y)$ be the data vector. Let X be a subvector of controls Z. The target function is the average partial derivative conditional on a covariate vector X,

$$g(x) = \partial_d \mathbb{E}[Y^D|X = x].$$

A standard way to identify the function $g(x)$ is to assume unconfoundedness. Suppose the potential outcome Y^d is independent of D conditional on Z,

$$\{Y^d, d \in \mathbb{R}\} \perp D|Z.$$

Then, $g(x)$ is identified as

$$g(x) = \mathbb{E}[\partial_d \mu_0(D, Z)|X = x],$$

where $\mu_0(d, z) = \mathbb{E}[Y|D = d, Z = z]$ is the regression function. Corollary 4.3 shows that the signal

$$Y(\eta) := -\partial_d \log s(D|Z)[Y - \mu(D, Z)] + \partial_d \mu(D, Z) \tag{2.4}$$

is orthogonal with respect to the nuisance parameter

$$\eta_0(d, z) = \{\mu_0(d, z), s_0(d|z)\},$$

where $s_0(d|z)$ is the conditional density of D given Z.

2.2. *Overview of main results*

The first main contribution of this paper is to provide sufficient conditions for pointwise and uniform asymptotic Gaussian approximation of the target function. We approximate the target function $g(x)$ by a linear form $p(x)'\beta_0$:

$$g(x) = p(x)'\beta_0 + r_g(x),$$

where $p(x)$ is a d-vector of basis functions of x, $r_g(x)$ is the linear approximation error, and β_0 is the best linear predictor / approximation parameter, defined by the normal equation

$$\mathbb{E}p(X)[g(X) - p(X)'\beta_0] = \mathbb{E}p(X)r_g(X) = 0.$$

We construct the *orthogonal estimator* $\widehat{\beta}$, the two-stage estimator of β_0, as follows. In the first stage, we construct an estimate $\widehat{\eta}$ of the nuisance parameter η_0, using a high-quality machine learning estimator capable of dealing with the high-dimensional covariate vector Z. In the second stage we construct an estimate $\widehat{Y}_i$ of the signal Y_i as $\widehat{Y}_i := Y_i(\widehat{\eta})$ and run ordinary least squares of $\widehat{Y}_i$ on $p(X_i)$. We use different samples to estimate η in the first stage and β_0 in the second stage in a form of cross-fitting.

DEFINITION 2.1 (CROSS-FITTING).

(1) *For a random sample of size N, denote a K-fold random partition of the sample indices* $[N] = \{1, 2, \ldots, N\}$ *by* $(J_k)_{k=1}^K$, *where K is the number of partitions, and the sample size of each fold is* $n = N/K$. *For each* $k \in [K] = \{1, 2, \ldots, K\}$ *define* $J_k^c = \{1, 2, \ldots, N\}\backslash J_k$.
(2) *For each* $k \in [K]$, *construct an estimator* $\widehat{\eta}_k = \widehat{\eta}(V_{i \in J_k^c})$ *of the nuisance parameter* η_0 *by using only the data* $\{V_j : j \in J_k^c\}$. *For any observation* $i \in J_k$, *define* $\widehat{Y}_i := Y_i(\widehat{\eta}_k)$.

DEFINITION 2.2 (ORTHOGONAL ESTIMATOR). *Given* $(\widehat{Y}_i)_{i=1}^N$, *define*

$$\widehat{\beta} := \left(\frac{1}{N}\sum_{i=1}^N p(X_i)p(X_i)'\right)^{-1} \frac{1}{N}\sum_{i=1}^N p(X_i)\widehat{Y}_i.$$

Figure 1 gives a schematic description of the orthogonal estimator. Under mild conditions on η, the orthogonal estimator delivers a high-quality estimate $p(x)'\widehat{\beta}$ of the pseudo-target function $p(x)'\beta_0$ with the following properties:

(1) With probability (w.p.) $\to 1$, the mean squared error of $p(x)'\widehat{\beta}$ is bounded by

$$\left(\frac{1}{N}\sum_{i=1}^{N}(p(X_i)'(\widehat{\beta}-\beta_0))^2\right)^{1/2} = O_P\left(\sqrt{\frac{d}{N}}\right).$$

(2) The estimator $p(x)'\widehat{\beta}$ of the pseudo-target function $p(x)'\beta_0$ is asymptotically linear:

$$\sqrt{N}\frac{p(x)'(\widehat{\beta}-\beta_0)}{\sqrt{p(x)'\Omega p(x)}} = G_N(x) + o_P(1/\sqrt{\log N}),$$

where the empirical process $G_N(x)$ is approximated by a Gaussian process

$$G(x) = \frac{p(x)'}{\sqrt{p(x)'\Omega p(x)}}N(0,\Omega)$$

uniformly over $x \in \mathcal{X}$, and the covariance matrix Ω can be consistently estimated by a sample analog $\widehat{\Omega}$.

(3) If the misspecification error $r_g(x)$ is small, the pseudo-target function $p(x)'\beta_0$ can be replaced by the target function $g(x)$:

$$\sqrt{N}\frac{p(x)'\widehat{\beta}-g(x)}{\sqrt{p(x)'\Omega p(x)}} = G_N(x) + o_P(1/\sqrt{\log N}).$$

(4) Simultaneous inference is facilitated by Gaussian bootstrap, which relies on simulating the empirical Gaussian process:

$$G^\star(x) = \frac{p(x)'}{\sqrt{p(x)'\Omega p(x)}}N(0,\widehat{\Omega}).$$

The quantiles of the suprema of $x \mapsto G(x)$ can be consistently approximated by simulation, which makes it possible to construct uniform confidence bands for $x \mapsto p(x)'\beta$ and $x \mapsto g(x)$.

Our results accommodate high-dimensional / highly complex modern machine learning methods to estimate η, such as random forests, neural networks, and ℓ_1-shrinkage estimators, as well as procedures that estimate $\widehat{\eta}$ by classic nonparametric methods. The only requirement we impose on the estimation of $\widehat{\eta}$ is that it converges to the true nuisance parameter η_0 at a fast enough rate $o_P(N^{-1/4-\delta})$ for some $\delta \geq 0$. This requirement is satisfied under structural assumptions on η_0, such as approximate sparsity of η_0 with respect to some dictionary, or if η_0 is well approximated by trees or by sparse neural and deep neural nets. For the CATE, it is straightforward to apply our method by using the best_linear_predictor command in the *R* grf package, available from Tibshirani et al. (2017).

Define the covariance matrix of the basis functions as

$$Q = \mathbb{E}p(X)p(X)',$$

and define its empirical analog as

$$\widehat{Q} = \frac{1}{N}\sum_{i=1}^{N}p(X_i)p(X_i)'.$$

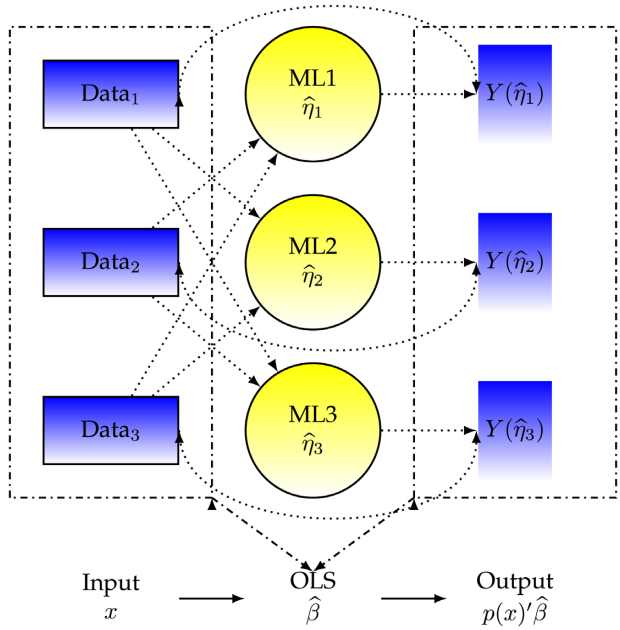


Figure 1. Graphical representation of the orthogonal estimator (OE) with cross-fitting. *First stage*. The rectangles represent the partition of the data into $K = 3$ subsets. The circles represent $K = 3$ instances of the machine learning algorithm (ML), whose training sets are indicated by straight arrows. For each partition $k \in \{1, 2, 3\}$, the signal $Y(\widehat{\eta}_k)$ is estimated using the Data$_k$ and ML instance $\widehat{\eta}_k$. *Second stage*. The ordinary least squares (OLS) estimator $\widehat{\beta} = (\sum_{i=1}^N p(X_i)p(X_i)')^{-1} \sum_{i=1}^N p(X_i)Y_i(\widehat{\eta})$ is estimated on the full data set with covariate vector $p(X)$ and outcome variable $Y(\widehat{\eta})$, where $p(X)$ is a vector of series terms. For a given point of interest x (input), the OE estimator of function $g(x)$ is $p(x)'\widehat{\beta}$ (output).

In Examples 2.2 through 2.4, the asymptotic covariance matrix of the orthogonal estimator is

$$\Omega = Q^{-1}\mathbb{E}p(X)p(X)'(U + r_g(X))^2 Q^{-1},$$

and its empirical analog is

$$\widehat{\Omega} := \widehat{Q}^{-1}\mathbb{E}_N p(X_i)p(X_i)'(Y_i(\widehat{\eta}) - p(X_i)'\widehat{\beta})^2 \widehat{Q}^{-1}. \tag{2.5}$$

In the case of CTEs (Example 2.1), the asymptotic variance contains an additional component that we describe in Section 4.

DEFINITION 2.3 (POINTWISE AND UNIFORM CONFIDENCE BANDS). *Let* $\widehat{g}(x) = p(x)'\widehat{\beta}$. *Denote the t-statistic as*

$$t_N(x) := \frac{\widehat{g}(x) - g(x)}{\widehat{\sigma}_N(x)}, \tag{2.6}$$

where $\widehat{\sigma}_N(x) = \sqrt{p(x)'\widehat{\Omega}p(x)/N}$, *and denote the bootstrapped t-statistic as*

$$\widehat{t}_N^b(x) := \frac{p(x)'\widehat{\Omega}^{1/2}/\sqrt{N}}{\widehat{\sigma}_N(x)}\mathcal{N}_d^b,$$

where $\mathcal{N}_d^b$ is a bootstrap draw from $N(0, I_d)$. Define the confidence bands for $g(x)$ as

$$[\underline{i}(x), \bar{i}(x)] := [\widehat{g}(x) - c_N(1-\alpha)\widehat{\sigma}_N(x), \quad \widehat{g}(x) + c_N(1-\alpha)\widehat{\sigma}_N(x)], \quad x \in \mathcal{X}, \tag{2.7}$$

where the critical value $c_N(1-\alpha)$ is the $(1-\alpha)$-quantile of $N(0,1)$ for the pointwise bands, and $c_N(1-\alpha)$ is the $(1-\alpha)$-quantile of $\sup_{x\in\mathcal{X}} |\widehat{t}_N^b(x)|$ for the uniform bands.

In the case of CTEs (Example 2.1), the orthogonal signal (2.1) involves an auxiliary nuisance parameter—the expectation of $\mu(x, Z)$ with respect to Z for each value of x. We estimate this parameter by a leave-one-out sample average,

$$Y_i^\dagger(\widehat{\eta}) = \frac{Y_i^o - \widehat{\mu}(X_i, Z_i)}{\widehat{s}(X_i|Z_i)}\widehat{w}(X_i) + \frac{1}{n-1}\sum_{j\in J_k, j\neq i} \widehat{\mu}(X_i, Z_j), \quad i \in J_k, \tag{2.8}$$

where $\widehat{\eta}(x,z) = \widehat{\eta}_k(x,z)$ is estimated on J_k^c for each k and the sample average in the second summand is taken over data $(V_j)_{j\in J_k}$ excluding observation i. Having replaced $Y_i(\widehat{\eta})$ by $Y_i^\dagger(\widehat{\eta})$ in Definition 2.2, we obtain an asymptotically linear estimator $p(x)'\widehat{\beta}^\dagger$ of the pseudo-target function:

$$\frac{\sqrt{N}\,p(x)'(\widehat{\beta}^\dagger - \beta_0)}{\sqrt{p(x)'\Omega^\dagger p(x)}} = G_N(x) + o_P(1/\sqrt{\log N}),$$

where $\Omega^\dagger$ is the asymptotic variance of $\widehat{\beta}^\dagger$, and the empirical process $G_N(x)$ is approximated by a Gaussian process

$$G(x) = \frac{p(x)'}{\sqrt{p(x)'\Omega^\dagger p(x)}} N(0, \Omega^\dagger)$$

uniformly over $x \in \mathcal{X}$.

3. MAIN THEORETICAL RESULTS

We use the empirical process notation. For a generic function f and a generic sample $(V_i)_{i=1}^N$, denote the empirical sample average by $\mathbb{E}_N f(V_i) := \frac{1}{N}\sum_{i=1}^N f(V_i)$ and the scaled, demeaned sample average by

$$\mathbb{G}_N f(V_i) := 1/\sqrt{N}\sum_{i=1}^N [f(V_i) - \int f(v)dP(v)].$$

The following assumptions impose regularity conditions on the covariate distribution, error terms, and the estimator of the nuisance parameter.

ASSUMPTION 3.1 (IDENTIFICATION). *Let $Q := \mathbb{E}p(X)p(X)' = Q_d$ denote the population covariance matrix of $p(X)$. Assume that there exist $0 < C_{\min} < C_{\max} < \infty$ that do not depend on d so that $C_{\min} \leq \min \text{eig}(Q) \leq \max \text{eig}(Q) \leq C_{\max}$ for all d.*

ASSUMPTION 3.2 (GROWTH CONDITION). *We assume that the* sup*-norm of the basis functions* $\xi_d := \sup_{x\in\mathcal{X}} \|p(x)\| = \sup_{x\in\mathcal{X}}(\sum_{j=1}^d p_j(x)^2)^{1/2}$ *grows sufficiently slowly:*

$$\sqrt{\frac{\xi_d^2 \log N}{N}} = o(1).$$

ASSUMPTION 3.3 (MISSPECIFICATION ERROR). *There exists a sequence of finite constants* l_d, r_d *such that the norms of the misspecification error are controlled as follows:*

$$\|r_g\|_{P,2} := \sqrt{\int r_g(x)^2 dP(x)} \lesssim r_d \text{ and } \|r_g\|_{P,\infty} := \sup_{x\in\mathcal{X}} |r_g(x)| \lesssim l_d r_d.$$

Assumption 3.3 introduces the rate of decay of the misspecification error. Specifically, the sequence of constants r_d bounds the mean squared misspecification error. In addition, the sequence $l_d r_d$ bounds the worst-case misspecification error uniformly over the domain $\mathcal{X}$, where l_d is the modulus of continuity of the worst-case error with respect to mean squared error.

Define the stochastic error U as

$$U := Y - g(X)$$

and the lower and upper bounds on its second moment conditional on X as

$$\underline{\sigma}^2 := \inf_{x\in\mathcal{X}} \mathbb{E}[U^2|X = x], \quad \overline{\sigma}^2 := \sup_{x\in\mathcal{X}} \mathbb{E}[U^2|X = x].$$

ASSUMPTION 3.4 (ERROR ASSUMPTION). *The second moment of the sampling error* U *conditional on* X *is bounded from above:* $\overline{\sigma}^2 \lesssim 1$.

To describe the first-stage rate requirement, Assumption 3.5 introduces a sequence of nuisance realisation sets $\mathcal{T}_N$ for the nuisance parameter η_0. As sample size N increases, the sets $\mathcal{T}_N$ shrink around the true value η_0. The shrinkage speed is described in terms of the statistical rates B_N and Λ_N.

ASSUMPTION 3.5 (SMALL BIAS CONDITION). *There exists a sequence* $\epsilon_N = o(1)$*, such that with probability at least* $1 - \epsilon_N$*, for all* $k \in [K]$*, the first-stage estimate* $\widehat{\eta}_k$*, obtained by cross-fitting (Definition 2.1), belongs to a shrinking neighborhood of* η_0*, denoted by* $\mathcal{T}_N$*. Uniformly over* $\mathcal{T}_N$*, the following mean square convergence holds:*

$$B_N := \sqrt{N} \sup_{\eta\in\mathcal{T}_N} \|\mathbb{E}p(X)[Y(\eta) - Y(\eta_0)]\| = o(1),$$

$$\Lambda_N := \sup_{\eta\in\mathcal{T}_N} (\mathbb{E}\|p(X)[Y(\eta) - Y(\eta_0)]\|^2)^{1/2} = o(1).$$

In particular, Λ_N *can be bounded as* $\Lambda_N \lesssim \xi_d \sup_{\eta\in\mathcal{T}_N} (\mathbb{E}(Y(\eta) - Y(\eta_0))^2)^{1/2}$.

REMARK 3.1 (SUFFICIENT CONDITIONS FOR ASSUMPTION 3.5). Assumption 3.5 is stated in a high-level form in order to accommodate various machine learning estimators. We demonstrate the plausibility of Assumption 3.5 for a high-dimensional sparse model for Example 2.3 in Appendix B, adapting the work of Belloni et al. (2017). Furthermore, one can also use deep neural networks (Schmidt-Hieber, 2017; Farrell et al., 2018) and random forest in small (Wager and Walther, 2015) dimensions and high (Syrganis and Zampetakis, 2020) dimensions with sparsity structure.

3.1. Pointwise limit theory

In this section, we establish pointwise asymptotic properties for the orthogonal estimator. Our first result is concerned with mean square convergence rate and pointwise linearisation.

LEMMA 3.1 (CONVERGENCE RATE AND POINTWISE LINEARIZATION). *Let Assumptions 3.1 through 3.5 hold. Then, the following statements hold:*

(a) The ℓ_2-norm of the estimation error is bounded as:

$$\|\widehat{\beta} - \beta_0\|_2 \lesssim_P \sqrt{\frac{d}{N}} + \Big[\sqrt{\frac{d}{N}} l_d r_d \wedge \xi_d r_d/\sqrt{N}\Big],$$

which implies a bound on the mean squared error of the estimate $p(x)'\widehat{\beta}$ of the pseudo-target function $p(x)'\beta_0$:

$$(\mathbb{E}_N(p(X_i)'(\widehat{\beta} - \beta_0))^2)^{1/2} \lesssim_P \sqrt{\frac{d}{N}} + \Big[\sqrt{\frac{d}{N}} l_d r_d \wedge \xi_d r_d/\sqrt{N}\Big].$$

(b) For any $\alpha \in \mathcal{S}^{d-1} := \{\alpha \in \mathbb{R}^d : \|\alpha\| = 1\}$, the estimator $\widehat{\beta}$ is approximately linear:

$$\sqrt{N}\alpha'(\widehat{\beta} - \beta) = \alpha' Q^{-1} \mathbb{G}_N p(X_i)(U_i + r_g(X_i)) + R_{1,N}(\alpha),$$

where the remainder term $R_{1,N}(\alpha)$ is bounded as

$$R_{1,N}(\alpha) \lesssim_P B_N + \Lambda_N + \sqrt{\frac{\xi_d^2 \log N}{N}} \left(1 + \min\left\{l_d r_d \sqrt{d}, \xi_d r_d\right\}\right).$$

Under a small bias condition, Lemma 3.1 states that the orthogonal estimator converges at the oracle rate and achieves oracle asymptotic linearity representation, where the oracle knows the true value of the nuisance parameter η_0.

THEOREM 3.1 (POINTWISE NORMALITY OF THE ORTHOGONAL ESTIMATOR). *Suppose Assumptions 3.1 through 3.5 hold. In addition, suppose $(\xi_d^2 \log N/N)^{1/2} \cdot (1 + l_d r_d \sqrt{d}) = o(1)$, $1 \lesssim \underline{\sigma}^2$, and the Lindeberg condition holds: $\sup_{x \in \mathcal{X}} \mathbb{E}[U^2 1_{|U|>M} | X = x] \to 0, \quad M \to \infty$. Then, for any $\alpha \in \mathcal{S}^{d-1}$, the orthogonal estimator is asymptotically normal:*

$$\lim_{N\to\infty} \sup_{t \in \mathbb{R}} \left| \mathrm{P}\left(\frac{\sqrt{N}\alpha'(\widehat{\beta} - \beta_0)}{\sqrt{\alpha' \Omega \alpha}} < t \right) - \Phi(t) \right| = 0.$$

Moreover, for any $x_0 = x_{0,N} \in \mathcal{X}$, the estimator $p(x_0)'\widehat{\beta}$ of the pseudo-target value $p(x_0)'\beta_0$ is asymptotically normal:

$$\lim_{N\to\infty} \sup_{t \in \mathbb{R}} \left| \mathrm{P}\left(\frac{\sqrt{N}\, p(x_0)'(\widehat{\beta} - \beta_0)}{\sqrt{p(x_0)' \Omega p(x_0)}} < t \right) - \Phi(t) \right| = 0,$$

and if the approximation error is negligible relative to the estimation error, namely $\sqrt{N} r_g(x_0) = o(\|\Omega^{1/2} p(x_0)\|)$, then $\widehat{g}(x) = p(x_0)'\widehat{\beta}$ is asymptotically normal:

$$\lim_{N\to\infty} \sup_{t \in \mathbb{R}} \left| \mathrm{P}\left(\frac{\sqrt{N}(\widehat{g}(x_0) - g(x_0))}{\sqrt{p(x_0)' \Omega p(x_0)}} < t \right) - \Phi(t) \right| = 0.$$

Theorem 3.1 delivers the pointwise convergence in distribution of the orthogonal estimator for any point x_0 that can depend on N.

3.2. Uniform limit theory

In this section, we establish uniform asymptotic properties for the orthogonal estimator. Not surprisingly, stronger conditions are required for our results to hold when compared with the pointwise case. Let $m > 2$. The following assumption controls the tails of the regression errors.

ASSUMPTION 3.6 (TAIL BOUNDS). *There exists a constant $m > 2$ such that the upper bound of the m^{th} moment of $|U|$ is bounded conditional on X:*

$$\sup_{x \in \mathcal{X}} \mathbb{E}[|U|^m | X = x] \lesssim 1.$$

Denote by $\alpha(x) := p(x)/\|p(x)\|$ the normalised value of basis functions vector $p(x)$. Define the Lipschitz constant as:

$$\xi_d^L = \sup_{x, x' \in \mathcal{X}, x \neq x'} \frac{\|\alpha(x) - \alpha(x')\|}{\|x - x'\|}.$$

ASSUMPTION 3.7 (BASIS). *Basis functions are well behaved, namely (i)$(\xi_d^L)^{2m/(m-2)} \log N/N \lesssim 1$ and (ii) $\log \xi_d^L \lesssim \log d$ for the same m as in Assumption 3.6.*

ASSUMPTION 3.8 (CONDITION FOR MATRIX ESTIMATION). *Let $\mathcal{T}_N$ be as in Assumption 3.5. Uniformly over $\mathcal{T}_N$, the following convergence holds:*

$$\kappa_N^1 := \sup_{\eta \in \mathcal{T}_N} \mathbb{E}[\max_{1 \leq i \leq N} |Y_i(\eta) - Y_i(\eta_0)|] = o(1),$$
$$\kappa_N := \sup_{\eta \in \mathcal{T}_N} (\mathbb{E} \max_{1 \leq i \leq N} (Y_i(\eta) - Y_i(\eta_0))^2)^{1/2} = o(1).$$

Lemma 3.2 establishes asymptotic linearity representation uniformly over the domain $\mathcal{X}$ and uniform convergence rate.

LEMMA 3.2 (UNIFORM RATE AND UNIFORM LINEARIZATION). *Suppose Assumptions 3.1 through 3.7 hold.*

(a) The orthogonal estimator is approximately linear uniformly over $\mathcal{X}$:

$$|\sqrt{N}\alpha(x)'(\widehat{\beta} - \beta_0) - \alpha'(x)Q^{-1}\mathbb{G}_N p(X_i)[U_i + r_g(X_i)]| \leq R_{1,N}(\alpha(x)),$$

where $R_{1,N}(\alpha(x))$, summarising the impact of unknown design and the first-stage error, obeys

$$\sup_{x \in \mathcal{X}} R_{1,N}(\alpha(x)) \lesssim_P B_N + \Lambda_N + \sqrt{\frac{\xi_d^2 \log N}{N}}(N^{1/m}\sqrt{\log N} + \sqrt{d} l_d r_d) =: \bar{R}_{1N}$$

uniformly over $x \in \mathcal{X}$. Moreover,

$$|\sqrt{N}\alpha(x)'(\widehat{\beta} - \beta_0) - \alpha'(x)Q^{-1}\mathbb{G}_N p(X_i)U_i| \leq R_{1,N}(\alpha(x)) + R_{2,N}(\alpha(x)),$$

where $R_{2,N}(\alpha(x))$, summarising the impact of misspecification error, obeys

$$R_{2,N}(\alpha(x)) \lesssim_P \sqrt{\log N} l_d r_d =: \bar{R}_{2N}$$

uniformly over $x \in \mathcal{X}$.

(b) *The estimator* $p(x)'\widehat{\beta}$ *of the pseudo-target* $p(x)'\beta_0$ *converges uniformly over* $\mathcal{X}$*:*

$$\sup_{x\in\mathcal{X}} |p(x)'(\widehat{\beta} - \beta_0)| \lesssim_P \frac{\xi_d}{\sqrt{N}}[\sqrt{\log N} + \bar{R}_{1N} + \bar{R}_{2N}].$$

REMARK 3.2 (OPTIMAL UNIFORM RATE IN HOLDER CLASS). Suppose the true function $g(x)$ belongs to the Holder smoothness class of order k, denoted by $\Sigma_k(\mathcal{X})$. Suppose $l_d r_d \lesssim d^{-k/\dim(X)}$, $\xi_d \lesssim \sqrt{d}$, $\bar{R}_{1N} + \bar{R}_{2N} \lesssim (\log N)^{1/2}$. Then, the optimal number d of technical regressors that comprise a vector $p(x)$ obeys

$$d \asymp (\log N/N)^{-\dim(X)/(2k+\dim(X))}.$$

This choice of d yields the optimal uniform rate:

$$\sup_{x\in\mathcal{X}} |\widehat{g}(x) - g(x)| \lesssim_P \left(\frac{\log N}{N}\right)^{k/(2k+\dim(X))}.$$

Theorem 3.2 establishes a strong approximation of the orthogonal estimator's series process by a sequence of zero-mean Gaussian processes.

THEOREM 3.2 (STRONG APPROXIMATION BY A GAUSSIAN PROCESS). *Suppose Assumptions 3.1 through 3.7 hold with* $m \geq 3$. *Let* $\bar{a}_N$ *be a sequence of positive numbers s.t.* $\bar{a}_N^{-1} = o(1)$. *Suppose (i)* $\bar{R}_{1N} = o(\bar{a}_N^{-1})$, *(ii)* $1 \lesssim \underline{\sigma}^2$, *and (iii)* $d^4 \bar{a}_N^6 \xi_d^2 (1 + l_d^3 r_d^3)^2 \log^2 N/N = o(1)$. *Then, for some* $\mathcal{N}_d \sim N(0, I_d)$, *the following statement holds for* $e(x) = \Omega^{1/2} p(x)$

$$\sqrt{N}\frac{p(x)'(\widehat{\beta} - \beta_0)}{\|e(x)\|} =_d \frac{e(x)}{\|e(x)\|}\mathcal{N}_d + o_P(\bar{a}_N^{-1}) \text{ in } \ell^\infty(\mathcal{X}).$$

In addition, if $\sup_{x\in\mathcal{X}} \sqrt{N}|r(x)|/\|e(x)\| = o(\bar{a}_N^{-1})$, *then, for* $\widehat{g}(x) = p(x)'\widehat{\beta}$,

$$\sqrt{N}\frac{\widehat{g}(x) - g(x)}{\|e(x)\|} =_d \frac{e(x)}{\|e(x)\|}\mathcal{N}_d + o_P(\bar{a}_N^{-1}) \text{ in } \ell^\infty(\mathcal{X}).$$

Theorem 3.3 establishes the convergence rate for the covariance matrix estimator $\widehat{\Omega}$. It extends Theorem 4.6 of Belloni et al. (2015), allowing the signal $Y(\eta_0)$ to depend on an unknown nuisance parameter η_0.

THEOREM 3.3 (MATRICES ESTIMATION). *Suppose Assumptions 3.1 through 3.8 hold. In addition, suppose (i)* $\bar{R}_{1N} + \bar{R}_{2N} \lesssim \sqrt{\log N}$ *and (ii)* $(N^{1/m} + l_d r_d)(\sqrt{\frac{\xi_d^2 \log N}{N}} + \kappa_N^1) = o(1)$. *Then, the estimator* $\widehat{\Omega}$, *defined in (2.5), converges in the matrix operator norm with the following rate:*

$$\|\widehat{\Omega} - \Omega\| \lesssim_P \left(N^{1/m} + l_d r_d\right) \cdot \left(\sqrt{\frac{\xi_d^2 \log N}{N}} + \kappa_N^1\right) + \kappa_N^2 =: a_N.$$

Moreover, for $\sigma_N(x) = \sqrt{p(x)'\Omega p(x)/N}$ *and* $\widehat{\sigma}_N(x) = \sqrt{p(x)'\widehat{\Omega} p(x)/N}$, *the following bound holds:*

$$\sup_{x\in\mathcal{X}} \left|\frac{\widehat{\sigma}_N(x)}{\sigma_N(x)} - 1\right| \lesssim_P \|\widehat{\Omega} - \Omega\| \lesssim_P a_N. \tag{3.1}$$

Theorem 3.4 establishes validity of empirical (Gaussian) bootstrap.

THEOREM 3.4 (VALIDITY OF GAUSSIAN BOOTSTRAP). *Suppose the assumptions of Theorem 3.2 hold with* $\bar{a}_N = \log N$ *and the assumptions of Theorem 3.3 hold with* $a_N = O(N^{-b})$ *for some* $b > 0$. *In addition, suppose (i)* $1 \lesssim \underline{\sigma}^2$ *and (ii) there exists a sequence* ξ'_N *obeying* $1 \lesssim \xi'_N \lesssim \|p(x)\|$ *uniformly for all* $x \in \mathcal{X}$ *so that* $\|p(x) - p(x')\|/\xi'_N \leq L_N \|x - x'\|$, *where* $\log L_N \lesssim \log N$. *Let* $\mathcal{N}_d^b$ *be a bootstrap draw from* $N(0, I_d)$ *and* P^* *be a probability conditional on data* $(V_i)_{i=1}^N$. *Then, the following approximation holds uniformly in* $\ell^\infty(\mathcal{X})$:

$$\frac{p(x)'\widehat{\Omega}^{1/2}}{\|\widehat{\Omega}^{1/2}p(x)\|}\mathcal{N}_d^b =^d \frac{p(x)'\Omega^{1/2}}{\|\Omega^{1/2}p(x)\|}\mathcal{N}_d^b + o_{P^*}(\log^{-1} N).$$

Theorem 3.5 establishes the asymptotic validity of uniform confidence bands. It also shows that the uniform width of the bands is of the same order as the uniform rate of convergence.

THEOREM 3.5 (VALIDITY OF UNIFORM CONFIDENCE BANDS). *Let Assumptions 3.1 through 3.8 hold with* $m \geq 4$. *In addition, suppose (i)* $\bar{R}_{1N} + \bar{R}_{2N} \lesssim \log^{-1/2} N$, *(ii)* $\xi_d \log^2 N / N^{1/2-1/m} = o(1)$, *(iii)* $1 \lesssim \underline{\sigma}^2$, *(iv)* $\sup_{x\in\mathcal{X}} \sqrt{N}|r_g(x)|/\|p(x)\| = o(\log^{-1/2} N)$, *and (v)* $d^4\xi_d^2(1 + l_d^3 r_d^3)^2 \log^5 N/N = o(1)$. *Then,*

$$\mathrm{P}\Big(\sup_{x\in\mathcal{X}} |t_N(x)| \leq c_N(1-\alpha)\Big) = 1 - \alpha + o(1)$$

for t_N *defined in (2.6). As a consequence, the confidence bands defined in (2.7) satisfy*

$$\mathrm{P}\big(g(x) \in [\underline{i}(x), \bar{i}(x)] \quad \forall x \in \mathcal{X}\big) = 1 - \alpha + o(1).$$

The width of the confidence bands $2c_N(1-\alpha)\widehat{\sigma}_N(x)$ *obeys*

$$2c_N(1-\alpha)\widehat{\sigma}_N(x) \lesssim_P \sigma_N(x)\sqrt{\log N} \lesssim \sqrt{\frac{\xi_d^2 \log N}{N}}$$

uniformly over $x \in \mathcal{X}$.

4. APPLICATIONS

In this section, we apply the results of Section 3 for empirically relevant settings, described in Examples 2.1 through 2.4.

4.1. Continuous Treatment Effects

Consider the setup of Example 2.1. We provide sufficient low-level conditions on the first-stage nuisance parameter $\eta_0(x, z)$ such that the pointwise and uniform Gaussian approximations established in Section 3 hold.

Assume that there exists a sequence of numbers $\epsilon_N = o(1)$ and sequences of neighborhoods S_N of $s_0(\cdot|\cdot)$, M_N of $\mu_0(\cdot, \cdot)$, and W_N of $w_0(\cdot)$ such that the first-stage estimate

$$\{\widehat{s}(\cdot|\cdot), \widehat{\mu}(\cdot, \cdot), \widehat{w}(\cdot)\}$$

belongs to the set $\{S_N \times M_N \times W_N\}$ with probability at least $1 - \epsilon_N$. The shrinkage speed of this set is measured by the following statistical rates:

$$\mathbf{s}_{N,q} := \sup_{s \in S_N} (\mathbb{E}(s(X|Z) - s_0(X|Z))^q)^{1/q},$$

$$\mathbf{m}_{N,q} := \sup_{\mu \in M_N} (\mathbb{E}(\mu(X, Z) - \mu_0(X, Z))^q)^{1/q},$$

$$\mathbf{w}_{N,q} := \sup_{w \in W_N} (\mathbb{E}(w(X) - w_0(X))^q)^{1/q},$$

where q is either a positive number $q \geq 2$ or $q = \infty$, which corresponds to ℓ_∞-norm (sup-norm). For $q = 2$, we will refer to $\mathbf{s}_N := \mathbf{s}_{N,2}$ as the conditional density mean square rate, $\mathbf{m}_N := \mathbf{m}_{N,2}$ as the regression function mean square rate, and $\mathbf{w}_N = \mathbf{w}_{N,2}$ as the marginal density mean square rate.

ASSUMPTION 4.9 (FIRST-STAGE RATE OF CTE). *Assume that mean square rates* $\mathbf{s}_N, \mathbf{m}_N,$ *and* $\mathbf{w}_N$ *decay sufficiently fast,*

$$\xi_d(\mathbf{s}_N \vee \mathbf{m}_N \vee \mathbf{w}_N) = o(1),$$

and one of two alternative conditions holds. (a) Bounded basis. There exists $\bar{B} < \infty$ *so that* $\sup_{x \in \mathcal{X}} \|p(x)\|_\infty \leq \bar{B}$ *and* $\sqrt{N}\sqrt{d}(\mathbf{m}_N \mathbf{s}_N \vee \mathbf{m}_N \mathbf{w}_N) = o(1)$. *(b) Unbounded basis. There exist* $\kappa, \gamma \in [1, \infty]$, $1/\kappa + 1/\gamma = 1$ *so that* $\sqrt{N}\sqrt{d}(\mathbf{m}_{N,2\gamma} \mathbf{s}_{N,2\kappa} \vee \mathbf{m}_{N,2\gamma} \mathbf{w}_{N,2\kappa}) = o(1)$. *Furthermore, there exists a constant* $\bar{C} < \infty$ *such that* $\sup_{\mu \in M_N} \sup_{(x,z) \in \mathcal{X} \times \mathcal{Z}} |\mu(x, z)| < \bar{C}$, $\sup_{s \in S_N} \sup_{(x,z) \in \mathcal{X} \times \mathcal{Z}} s^{-1}(x|z) < \bar{C}$, *and* $\sup_{s \in S_N} \sup_{(x,z) \in \mathcal{X} \times \mathcal{Z}} s(x|z) < \bar{C}$,

$$\sup_{w \in W_N} \sup_{x \in \mathcal{X}} w^{-1}(x) < \bar{C}.$$

LEMMA 4.3 (ORTHOGONAL SIGNAL FOR CTE). *Suppose Assumption 4.9 holds. Then, the orthogonal signal* $Y(\eta)$, *defined in (2.1), satisfies Assumption 3.5.*

As discussed in the Introduction and Motivation, the estimator $\widehat{\beta}^\dagger$ takes the form

$$\widehat{\beta}^\dagger = \widehat{Q}^{-1} \frac{1}{N} \sum_{i=1}^{N} p(X_i) Y_i^\dagger(\widehat{\eta}),$$

where $Y_i^\dagger(\widehat{\eta})$ is as in (2.8). Because μ enters linearly in (2.1), the error term $Y_i^\dagger(\widehat{\eta}) - Y_i(\widehat{\eta})$ does not introduce bias in $\widehat{\beta}^\dagger$ but introduces an extra term in asymptotic variance, which we characterise below. For a function $\mu(x, z)$, define its demeaned analog $\mu^0(x, z)$ as

$$\mu^0(x, z) := \mu(x, z) - \mathbb{E}\mu(x, Z)$$

and the kernel function as

$$\tau(v_1, v_2; \mu) = \frac{1}{2}(p(x_1)\mu^0(x_1, z_2) + p(x_2)\mu^0(x_2, z_1)),$$

where $v_1 = (x_1, z_1)$ and $v_2 = (x_2, z_2)$. Finally, define the Hajek projection of $\tau(v_1, v_2; \mu)$ as

$$\tau_1(v; \mu) := \mathbb{E}\tau(v, V; \mu) = \mathbb{E}p(X)\mu^0(X, z) = \tau_1(z; \mu). \tag{4.1}$$

Decompose the orthogonal estimator $\widehat{\beta}^{\dagger}$ into the sum of its infeasible analog $\widehat{\beta}$ and a mean-zero U-statistic with the kernel function $\tau(v_1, v_2; \mu)$

$$\widehat{\beta}^{\dagger} =: \widehat{\beta} + \widehat{Q}^{-1} \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n(n-1)} \sum_{i,j \in J_k, i \neq j} \tau(V_i, V_j; \widehat{\mu}).$$

Theorem 4.6 establishes pointwise asymptotic normality of the orthogonal estimator. Its asymptotic variance is

$$\Omega^{\dagger} = Q^{-1} \Sigma^{\dagger} Q^{-1},$$

where $\Sigma^{\dagger}$ is

$$\Sigma^{\dagger} = \mathbb{E}[p(X)(U + r_g(X)) + \tau_1(Z; \mu_0)][p(X)(U + r_g(X)) + \tau_1(Z; \mu_0)]'.$$

THEOREM 4.6 (POINTWISE ASYMPTOTIC THEORY FOR CTES) *Suppose Assumptions 3.1 through 3.4 and 4.9 hold. Let $C^{\dagger}_{\min} > 0$ be an absolute constant. In addition, suppose (i) $(\xi_d^2 \log N/N)^{1/2} \cdot (1 + l_d r_d \sqrt{d}) = o(1)$, (ii) $\sqrt{d\xi_d^4 \log^3 N/N^2} + \sqrt{d}\mathbf{m}_N = o(1)$, $\xi_d l_d r_d = o(N^{1/2})$, (iii) $\min \operatorname{eig} \Omega^{\dagger} \geq C^{\dagger}_{\min}$, and (iv) the Lindeberg condition holds: $\sup_{x \in \mathcal{X}} \mathbb{E}[U^2 1_{|U|>M} | X = x] \to 0$, $M \to \infty$. Then, for any $x_0 = x_{0,N} \in \mathcal{X}$ the estimator $p(x_0)'\widehat{\beta}^{\dagger}$ of the pseudo-target $p(x_0)'\beta_0$ is asymptotically normal:*

$$\lim_{N \to \infty} \sup_{t \in \mathbb{R}} \left| \mathrm{P}\left(\frac{\sqrt{N} p(x_0)'(\widehat{\beta}^{\dagger} - \beta_0)}{\sqrt{p(x_0)' \Omega^{\dagger} p(x_0)}} < t \right) - \Phi(t) \right| = 0,$$

and if the approximation error is negligible relative to the estimation error, namely $\sqrt{N} r_g(x_0) = o(\|(\Omega^{\dagger})^{1/2} p(x_0)\|)$, then $\widehat{g}^{\dagger}(x_0) = p(x_0)'\widehat{\beta}^{\dagger}$ is asymptotically normal:

$$\lim_{N \to \infty} \sup_{t \in \mathbb{R}} \left| \mathrm{P}\left(\frac{\sqrt{N}(\widehat{g}^{\dagger}(x_0) - g(x_0))}{\sqrt{p(x_0)' \Omega^{\dagger} p(x_0)}} < t \right) - \Phi(t) \right| = 0.$$

THEOREM 4.7 (UNIFORM ASYMPTOTIC THEORY FOR CTES). *Suppose Assumptions 3.1 through 3.7 hold with $\xi_d^L / C_{min} \geq e^2/16 \vee e$. In addition, suppose $\sqrt{d\xi_d^2 \log N/N} = o(1)$ and $\xi_d \log^2 N/N = o(1)$. Then, the following statements hold.*

(a) The orthogonal estimator converges uniformly over $\mathcal{X}$:

$$\sup_{x \in \mathcal{X}} |p(x)'(\widehat{\beta}^{\dagger} - \beta_0)| \lesssim_P \frac{\xi_d}{\sqrt{N}} \left(\sqrt{\log N} + \bar{R}_{1N} + \bar{R}_{2N} + \sqrt{d}\mathbf{m}_N \right),$$

where $\bar{R}_{1N}$ and $\bar{R}_{2N}$ are as in Lemma 3.2.

(b) Suppose Assumptions 3.6 and 3.7 hold with $m \geq 3$. Then, the statement of Theorem 3.2 holds for $\widehat{\beta}^{\dagger}$ in place of $\widehat{\beta}$, $\Omega^{\dagger}$ in place of Ω, and $e(x) := (\Omega^{\dagger})^{1/2} p(x)$.

Theorem 4.7 establishes a uniform convergence rate and strong Gaussian approximation for CTEs. We compare our Theorems 4.6 and 4.7 with Kennedy et al. (2017), who introduced the doubly robust score for the average potential outcome. First, by virtue of sample splitting, we do not impose any complexity requirements on the estimator of the first-stage nuisance parameters. In particular, unlike Theorem 2 of Kennedy et al. (2017), we do not require the function class containing $\widehat{\mu}(\cdot, \cdot), \widehat{s}(\cdot|\cdot), \widehat{w}(\cdot)$ to have bounded uniform entropy integrals. As a result, our method accommodates a wide class of modern regularised methods to be used to

estimate first-stage parameters, as discussed in Remark 3.1. Second, Theorem 4.7 offers uniform asymptotic statements, whereas the work by Kennedy et al. (2017) offers only pointwise results. Finally, Kennedy et al. (2017) uses local linear regression in the second stage and delivers convergence at a $\sqrt{Nh^{\dim(X)}}$ rate, where $h = h(N)$ is the kernel bandwidth. In contrast, our method delivers $\sqrt{N}$-approximation for the normalised projection $p(x)'\beta_0/\|p(x)\|$.

4.2. Conditional Average Treatment Effect

Consider the setup of Example 2.2. We provide sufficient low-level conditions on the regression functions $\mu_0(1,\cdot)$, $\mu_0(0,\cdot)$ and the propensity score $s_0(\cdot)$ such that the pointwise and uniform Gaussian approximations of Section 3 hold.

ASSUMPTION 4.10 (STRONG OVERLAP).

(a) The propensity score is bounded above and below. Specifically, there exists $\bar{\pi}_0 > 0$ such that $0 < \bar{\pi}_0 < s_0(z) < 1 - \bar{\pi}_0 < 1$ for any $z \in \mathcal{Z}$.

(b) The propensity score is bounded below. Specifically, there exists $\bar{\pi}_0 > 0$ such that $0 < \bar{\pi}_0 < s_0(z) < 1$ for any $z \in \mathcal{Z}$.

In the context of Example 2.2, Assumption 4.10(a) ensures that the probability of assignment to the treatment and control group is bounded away from zero. In the context of Example 2.3, Assumption 4.10(b) ensures that the probability of observing the response Y^* is bounded away from zero.

Given the true functions $s_0(\cdot)$, $\mu_0(1,\cdot)$, $\mu_0(0,\cdot)$ and sequences of shrinking neighborhoods S_N of $s_0(\cdot)$ and M_N of $\mu_0(1,\cdot)$ and of $\mu_0(0,\cdot)$, define the following rates:

$$\mathbf{s}_{N,q} := \sup_{s \in S_N} (\mathbb{E}(s(Z) - s_0(Z))^q)^{1/q},$$

$$\mathbf{m}_{N,q} := \sup_{\mu \in M_N} (\mathbb{E}(\mu(1,Z) - \mu_0(1,Z))^q)^{1/q} \vee \sup_{\mu \in M_N} (\mathbb{E}(\mu(0,Z) - \mu_0(0,Z))^q)^{1/q},$$

where $q \geq 2$ is either a positive number or $q = \infty$. We will refer to $\mathbf{s}_N := \mathbf{s}_{N,2}$ as the propensity score mean square rate and $\mathbf{m}_N := \mathbf{m}_{N,2}$ as the regression function mean square rate.

ASSUMPTION 4.11 (FIRST-STAGE RATE FOR CATE). *Assume that there exists a sequence of numbers $\epsilon_N = o(1)$ and sequences of neighborhoods S_N of $s_0(\cdot)$ and M_N of $\mu_0(1,\cdot)$ and $\mu_0(0,\cdot)$ such that the first-stage estimate $\{\widehat{s}(\cdot), \widehat{\mu}(1,\cdot), \widehat{\mu}(0,\cdot)\}$ belongs to the set $\{S_N \times M_N \times M_N\}$ w.p. at least $1 - \epsilon_N$. Assume that mean square rates $\mathbf{s}_N, \mathbf{m}_N$ decay sufficiently fast:*

$$\xi_d(\mathbf{s}_N \vee \mathbf{m}_N) = o(1),$$

and one of two alternative conditions holds. (a) Bounded basis. There exists $\bar{B} < \infty$ so that $\sup_{x \in \mathcal{X}} \|p(x)\|_\infty \leq \bar{B}$ and $\sqrt{N}\sqrt{d}\mathbf{m}_N\mathbf{s}_N = o(1)$. (b) Unbounded basis. There exist $\kappa, \gamma \in [1,\infty]$, $1/\kappa + 1/\gamma = 1$ so that $\sqrt{N}\sqrt{d}\mathbf{m}_{N,2\gamma}\mathbf{s}_{N,2\kappa} = o(1)$. Finally, the functions in S_N and M_N are bounded uniformly over their domain:

$$\sup_{\mu \in M_N} \sup_{z \in \mathcal{Z}} \sup_{d \in \{1,0\}} |\mu(d,z)| \vee \sup_{s \in S_N} \sup_{z \in \mathcal{Z}} s^{-1}(z) < \bar{C} < \infty.$$

COROLLARY 4.1 (ASYMPTOTIC THEORY FOR CATE). *Under Assumptions 4.10(a) and 4.11, the orthogonal signal $Y(\eta)$, given by Equation (2.2), satisfies Assumption 3.5. As a result, the statements of Theorems 3.1 through 3.5 hold for CATE.*

4.3. Regression function with partially missing outcome

Consider the setup of Example 2.3. Define the regression function rate $\mathbf{m}_{N,q}$ as

$$\mathbf{m}_{N,q} := \sup_{\mu \in M_N} (\mathbb{E}(\mu(Z) - \mu_0(Z))^q)^{1/q},$$

where $q \geq 2$ is either a positive number or $q = \infty$, and let $\mathbf{s}_{N,q}$ be as defined in Section 4.2. We show that pointwise and uniform Gaussian approximations of Section 3 hold for regression function with partially missing outcome.

COROLLARY 4.2 (ASYMPTOTIC THEORY FOR REGRESSION FUNCTION WITH PARTIALLY MISSING OUTCOME). *Suppose Assumptions 4.10(b) and 4.11 hold for $\mathbf{s}_{N,q}$, defined in Example 4.2, and $\mathbf{m}_{N,q}$, redefined above. Then, the orthogonal signal $Y(\eta)$, given by Equation (2.3), satisfies Assumption 3.5. Then, the statements of Theorems 3.1 through 3.5 hold for the regression function with partially missing outcome.*

We give an example of low-level sparse conditions for the propensity score $s(\cdot)$ and regression $\mu(\cdot)$ in Appendix B.

4.4. Conditional average partial derivative

Consider the setup of Example 2.4. We provide sufficient low-level conditions on the regression functions $s(d|z)$, $\mu(d, z)$ such that the pointwise and uniform Gaussian approximations of Section 3 hold.

Given a true function $s_0(\cdot|\cdot)$, $\mu_0(\cdot, \cdot)$, let S_N, M_N be a sequence of shrinking neighborhoods of $s_0(\cdot|\cdot)$ and $\mu_0(\cdot, \cdot)$, constrained as follows:

$$\mathbf{s}_{N,q} := \sup_{s \in \mathcal{S}_N} (\mathbb{E}(s(D|Z)) - s_0(D|Z))^q)^{1/q} \vee \sup_{s \in \mathcal{S}_N} (\mathbb{E}(\partial_d s(D|Z)) - \partial_d s_0(D|Z))^q)^{1/q}$$

$$\mathbf{m}_N := \sup_{\mu \in M_N} (\mathbb{E}(\mu(D, Z) - \mu_0(D, Z))^q)^{1/q} \vee \sup_{\mu \in M_N} (\mathbb{E}(\partial_d \mu(D, Z) - \partial_d \mu_0(D, Z))^q)^{1/q},$$

where $q \geq 2$ is either a positive number or $q = \infty$. We will refer to $\mathbf{s}_N := \mathbf{s}_{N,2}$ as the mean square conditional density rate and $\mathbf{m}_N := \mathbf{m}_{N,2}$ as the mean square regression function rate.

ASSUMPTION 4.12 (FIRST-STAGE RATE FOR CONDITIONAL AVERAGE PARTIAL DERIVATIVE). *Assume that there exists a sequence of numbers $\epsilon_N = o(1)$ and a sequence of neighborhoods S_N, M_N of functions $s_0(\cdot|\cdot)$, $\mu_0(\cdot, \cdot)$ such that the first-stage estimate $\{\widehat{s}(\cdot|\cdot), \widehat{\mu}(\cdot)\}$ belongs to the set $\{S_N \times M_N\}$ w.p. at least $1 - \epsilon_N$. Assume that mean square rates $\mathbf{s}_N, \mathbf{m}_N$ decay sufficiently fast $\xi_d(\mathbf{s}_N \vee \mathbf{m}_N) = o(1)$, and one of two alternative conditions hold. (a) Bounded basis. There exists $\bar{B} < \infty$ so that $\sup_{x \in \mathcal{X}} \|p(x)\|_\infty \leq \bar{B}$ and $\sqrt{N}\sqrt{d}\mathbf{m}_N\mathbf{s}_N = o(1)$. (b) Unbounded basis. There exist $\kappa, \gamma \in [1, \infty]$, $1/\kappa + 1/\gamma = 1$ so that $\sqrt{N}\sqrt{d}\mathbf{m}_{N,2\gamma}\mathbf{s}_{N,2\kappa} = o(1)$. The functions in M_N and S_N are bounded uniformly over their domain:*

$$\sup_{\mu \in M_N} \sup_{(d,z) \in \mathrm{R} \times \mathcal{Z}} |\mu(d, z)| \vee \sup_{s \in \mathcal{S}_N} \sup_{(d,z) \in \mathrm{R} \times \mathcal{Z}} \max\{s(d|z), s^{-1}(d|z)\} < \bar{C}.$$

COROLLARY 4.3 (ASYMPTOTIC THEORY FOR CONDITIONAL AVERAGE PARTIAL DERIVATIVE). *Suppose Assumption 4.12 holds. Then, the orthogonal signal $Y(\eta)$, given by Equation (2.3), satisfies Assumption 3.5. Then, the statements of Theorems 3.1 through 3.5 hold for conditional average partial derivative.*

5. EMPIRICAL APPLICATION: INFERENCE ON CONDITIONAL AVERAGE ELASTICITY

To show the immediate usefulness of the method, we consider an important problem of inference on structural derivatives. We apply our methods to study the household demand for gasoline, a question studied in Hausman and Newey (1995), Schmalensee and Stoker (1999), Yatchew and No (2001), and Blundell et al. (2012). These papers estimated the demand function and the average price elasticity for various demographic groups. The dependence of the price elasticity on the household income was highlighted in Blundell et al. (2012), who have estimated the elasticity by low, middle, and high-income groups and found its relationship with income to be non-monotonic. To gain more insight into this question, we estimate the average price elasticity as a function of income and provide simultaneous confidence bands for it.

The data for our analysis are the same as in Yatchew and No (2001), coming from the National Private Vehicle Use Survey, conducted by Statistics Canada between October 1994 and September 1996. The data set is based on fuel purchase diaries and contains detailed information about fuel prices, fuel consumption patterns, vehicles, and demographic characteristics. We use the same selection procedure as in Yatchew and No (2001) and Belloni et al. (2019), focusing on a sample of the households with non-zero licensed drivers, vehicles, and distances driven, which leaves us with 5,001 observations.

The object of interest is the average predicted percentage change in the demand due to a unit percentage change in the price, holding the observed demographic characteristics fixed, conditional on income. In the context of Example 2.4, this corresponds to the conditional average derivative

$$g(x) = \mathbb{E}[\partial_d \mu(D, Z)|X = x],$$
$$\mu(d, z) = \mathbb{E}[Y|D = d, Z = z],$$

where Y is the logarithm of gas consumption, D is the logarithm of price per liter, X is log income, and Z are the observed subject characteristics, such as household size and composition, distance driven, the type of fuel used, and income. We use the orthogonal signal $Y(\eta)$ of (2.4).

The choice of the estimators in the first and the second stages is as follows. To estimate the conditional expectation function $\mu(d, z)$ and its partial derivative $\partial_d \mu(d, z)$, we consider a linear model

$$\mu(d, z) = b(d, z)'\omega,$$

where the basis function $b(d, z)$ includes price, price squared, income, income squared, and their interactions with 28 time, geographical, and household composition dummies. All in all, we have 91 explanatory variables. We estimate the coefficient vector ω using Lasso with the penalty level chosen as in Belloni et al. (2014) and plug the estimate $\widehat{\omega}$ into the expression for the derivative,

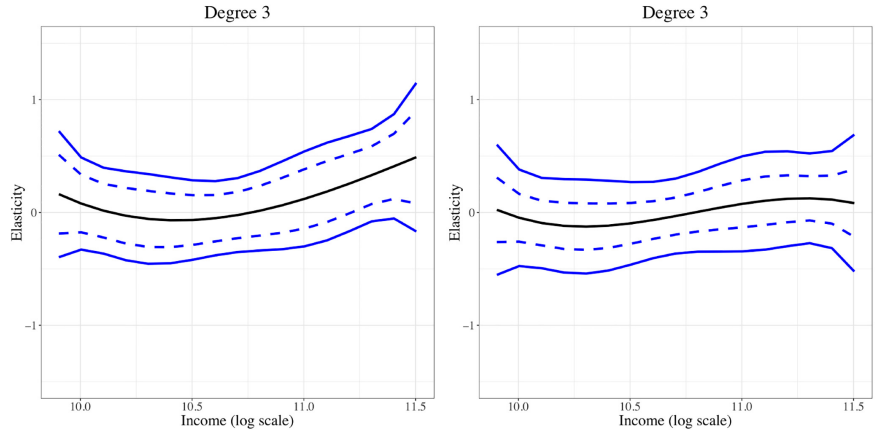
$$\partial_d \mu(d, z) = \partial_d b(d, z)'\omega,$$

to estimate $\partial_d \mu(d, z)$.

To estimate the conditional density $s(d|z)$, we consider a model:

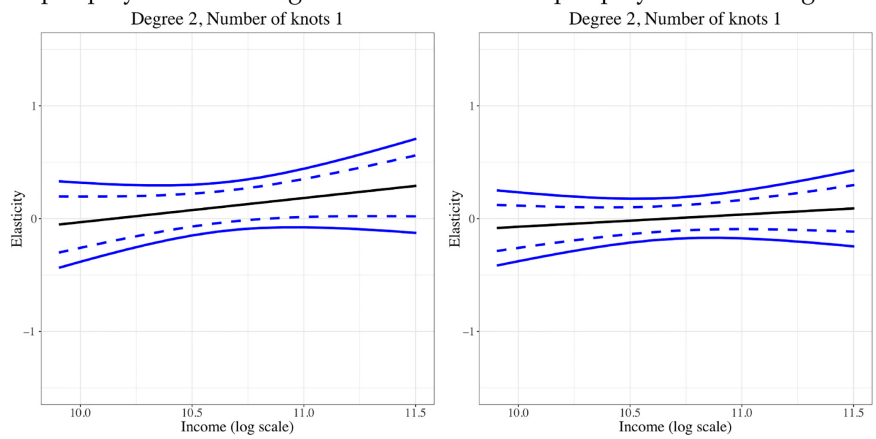
$$D = l(Z) + U, \quad U \perp Z,$$

where $l(z) = \mathbb{E}[D|Z = z]$ is the conditional expectation of price variable D given covariates Z, and U is an independent continuously distributed shock with univariate density $\phi(\cdot)$. Under this



(a) Step 2: $l(z)$ is estimated by Lasso. Step 3: polynomials of degree 3.

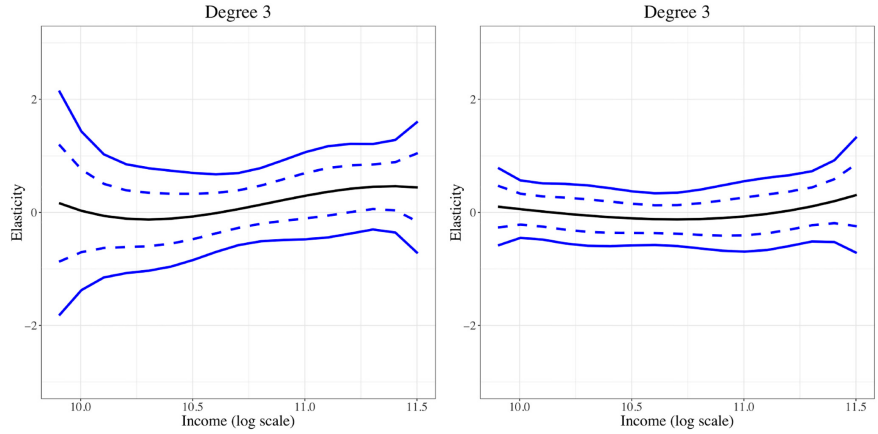
(b) Step 2: $l(z)$ is estimated by random forest. Step 3: polynomials of degree 3.



(c) Step 2: $l(z)$ is estimated by Lasso. Step 3: B-splines of order 2 with 1 knot.

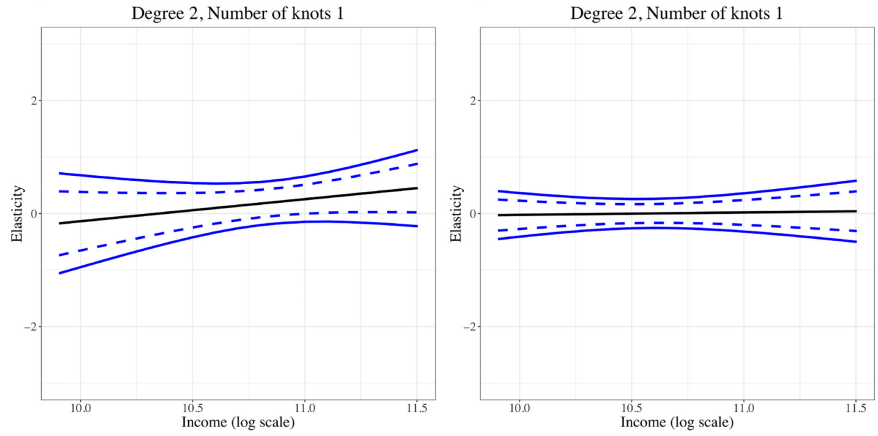
(d) Step 2: $l(z)$ is estimated by random forest. B-splines of order 2 with 1 knot.

Figure 2. 95% confidence bands for the best linear approximation of the average price elasticity conditional on income with accounting for the demographic controls in the first stage. The black line is the estimated function, and the dashed (solid) blue lines are the pointwise (uniform) confidence bands. The estimation algorithm has three steps: (1) first-stage estimation of the conditional expectation function $\mu(d, z)$, (2) second-stage estimation of the conditional density $s(d|z)$, and (3) third-stage estimation of the target function $g(x)$ by least-squares series. Step 1 is performed by using Lasso with standardised covariates and the penalty choice $\lambda = 2.2\sqrt{n}\widehat{\sigma}\Phi^{-1}(1 - \gamma/2p)$, where $\gamma = 0.1/\log n$ and $\widehat{\sigma}$ is the estimate of the residual variance. Step 2 is performed by estimating the regression function of $l(z) = \mathbb{E}[D|Z = z]$ and estimating the density $\phi(d - l(z))$ of the residual $d - l(z)$ by the adaptive kernel density estimator of Portnoy and Koenker (1989) with the Silverman choice of bandwidth. The regression function $l(z)$ is estimated Lasso (a, c) and random forest (b, d). Step 3 is performed using B-splines of order 2 with the number of knots equal to one (c, d) and polynomial functions of order 3 (2a, b). Uniform confidence bands are based on $B = 200$ repetitions of weighted (Bayes) bootstrap algorithm, described in Belloni et al. (2015).



(a) Large Households, Polynomials of degree 3.

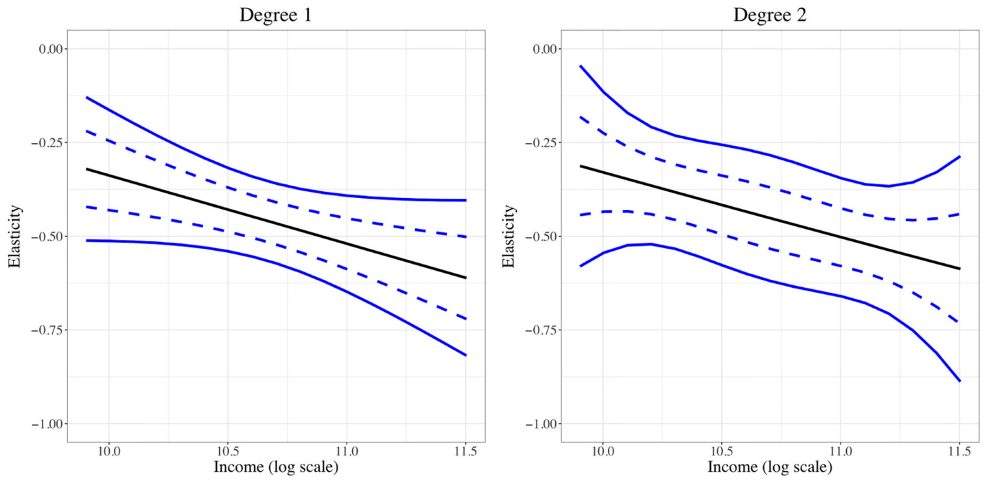
(b) Small Households, Polynomials of degree 3.



(c) Large Households, B-splines of degree 2 with 1 knot.

(d) Small Households, B-splines of degree 2 with 1 knot.

Figure 3. 95% confidence bands for the best linear approximation of the average price elasticity conditional on income with accounting for the demographic controls in the first stage by household size. The black line is the estimated function, and the dashed (solid) blue lines are the pointwise (uniform) confidence bands. The estimation algorithm has three steps: (1) first-stage estimation of the conditional expectation function $\mu(d, z)$, (2) second-stage estimation of the conditional density $s(d|z)$, and (3) third-stage estimation of the target function $g(x)$ by least-squares series. Step 1 is performed by using Lasso with standardised covariates and the penalty choice $\lambda = 2.2\sqrt{n}\widehat{\sigma}\Phi^{-1}(1 - \gamma/2p)$, where $\gamma = 0.1/\log n$ and $\widehat{\sigma}$ is the estimate of the residual variance. Step 2 is performed by estimating the regression function of $l(z) = \mathbb{E}[D|Z = z]$ and estimating the density $\phi(d - l(z))$ of the residual $d - l(z)$ by adaptive kernel density estimator of Portnoy and Koenker (1989) with the Silverman choice of bandwidth. The regression function $l(z)$ is estimated Lasso. Step 3 is performed by using B-splines of order 2 with the number of knots equal to one (c, d) and using non-orthogonal polynomial functions of degree 3 (3a, b). Uniform confidence bands are based on $B = 200$ repetitions of weighted (Bayes) bootstrap algorithm, described in Belloni et al. (2015).



(a) Polynomial degree $d = 1$ (b) Polynomial degree $d = 2$

Figure 4. 95% confidence bands for the best linear approximation of the average price elasticity conditional on income without accounting for the demographic controls in the first stage. The black line is the estimated function, and the dashed blue lines and the solid blue lines are the pointwise and the uniform confidence bands. The estimation algorithm has three steps: (1) first-stage estimation of the conditional expectation function $\mu(d, x) = \mathbb{E}[Y|D = d, X = x]$, (2) second-stage estimation of the conditional density $s(d|x)$, and (3) third-stage estimation of the target function $g(x)$ by least-squares series. Step 1 is performed by using least-squares series regression using polynomial functions $\{1, x, \ldots, x^q\}$, $q = 3$ whose power q is chosen by cross-validation out of $\{1, 2, 3\}$. Step 2 is performed by kernel density estimator with the Silverman choice of bandwidth. Step 3 is performed by using polynomial functions $\{1, x, \ldots, x^d\}$ and is shown for $d = 1$ and $d = 2$. Uniform confidence bands are based on $B = 200$ repetitions of weighted (Bayes) bootstrap algorithm, described in Belloni et al. (2015).

assumption, the log density $\partial_d \log s(d|z)$ equals to

$$\partial_d \log s(d|z) = \frac{\phi'(d - l(z))}{\phi(d - l(z))}.$$

We estimate $\phi(u) : \mathbb{R} \to \mathbb{R}^+$ by an adaptive kernel density estimator of Portnoy and Koenker (1989) with Silverman choice of bandwidth. Finally, we plug in the estimates of $\mu(d, z)$, $\partial_d \mu(d, z)$, and $s(d|z)$ into Equation (1.4) to get an estimate of $\widehat{Y}$ and estimate $g(x)$ by least-squares series regression of $\widehat{Y}$ on X. We try both polynomial basis function and B-splines to construct technical regressors.

Figures 2 and 3 report the estimate of the target function (the black line) and the pointwise (the dashed blue lines) and the uniform confidence (the solid blue lines) bands for the average price elasticity conditional on income, where the significance level $\alpha = 0.05$. The panels of Figure 2 correspond to different choices of the first-stage estimates of the nuisance functions $\mu(d, z)$ and $s(d|z)$ and dictionaries of technical regressors. The panels of Figure 3 correspond to the subsamples of large and small households and to different choices of the dictionaries.

The summary of our empirical findings based on Figures 2 and 3 is as follows. We find the elasticity to be in the range $(-1, 0)$ and significant for the majority of income levels. The estimates based on B-splines (Figure 2c and d) are monotonically increasing in income, which is intuitive. The estimates based on polynomial functions are non-monotonic in income. For every algorithm in Figure 2 we cannot reject the null hypothesis of constant price elasticity for all income levels; for each estimation procedure, the uniform confidence bands contain the constant function. Figure 3 shows the average price elasticity conditional on income for small and large households. For the majority of income levels, we find large households to be more price elastic than the small ones, but the difference is not significant at any income level.

To demonstrate the relevance of demographic data Z in the first-stage estimation, we also show the average predicted effect of the price change on the gasoline consumption (in logs), without accounting for the covariates in the first stage. In particular, this effect equals to $\mathbb{E}[\partial_d \mu(D, X)|X = x]$, where $\mu(d, x) = \mathbb{E}[Y|D = d, X = x]$ is the conditional expectation of gas consumption given income and price. Figure 4 shows this predictive effect, approximated by the polynomials of degree $d \in \{1, 2\}$, conditional on income. By contrast to the results in Figure 2, the slope of the polynomial of degree $d = 1$ has a negative relationship between income and price elasticity, which presents evidence that the demographics strongly confound the relationship between income and price elasticity.

REFERENCES

Abrevaya, J., Y.-C. Hsu and R. Lieli (2015). Estimating conditional average treatment effects. *Journal of Business and Economic Statistics 33*(4), 485–505.

Athey, S. and G. Imbens (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences 113*(27), 7353–460.

Belloni, A., V. Chernozhukov, D. Chetverikov and I. Fernandez-Val (2019). Conditional quantile processes based on series or many regressors. *Journal of Econometrics 213*(260), 4–29.

Belloni, A., V. Chernozhukov, D. Chetverikov and K. Kato (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics 186*(2), 345–66.

Belloni, A., V. Chernozhukov, I. Fernandez-Val and C. Hansen (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica 85*, 233–98.

Belloni, A., V. Chernozhukov and C. Hansen (2014). Inference on treatment effects after selection amongst high-dimensional controls. *Journal of Economic Perspectives 28*(2), 608–50.

Belloni, A., V. Chernozhukov and Y. Wei (2016). Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics 34*(4), 606–19.

Blundell, R., J. Horowitz and M. Parey (2012). Measuring the price responsiveness of gasoline demand: Economic shape restrictions and nonparametric demand estimation. *Quantitative Economics 3*(1), 29–51.

Bühlmann, P. and S. van der Geer (2011). Statistics for high-dimensional data: methods, theory and applications. *Springer Series in Statistics*.

Chen, X. and T. Christensen (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics 188*, 447–65.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal 21*, C1–68.

Chernozhukov, V., M. Demirer, E. Duflo and I. Fernández-Val (2017). Generic machine learning inference on heterogenous treatment effects in randomized experiments. *arXiv e-prints*, arXiv:1712.04802.

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey and J. M. Robins (2016). Locally robust semiparametric estimation. *arXiv e-prints*, arXiv:1608.00033.

Colangelo, K. and Y.-Y. Lee (2020). Double debiased machine learning nonparametric inference with continuous treatments. *arXiv e-prints*, arXiv:2004.03036.

Fan, Q., Y.-C. Hsu, R. P. Lieli and Y. Zhang (2020). Estimation of conditional average treatment effects with high-dimensional data. *Forthcoming in Journal of Business and Economic Statistics.*

Farrell, M. H., T. Liang and S. Misra (2020). Deep neural networks for estimation and inference. *Forthcoming in Econometrica*.

Gill, R. and J. Robins (2001). Causal inference for complex longitudinal data: the continuous case. *Annals of Statistics 29*(6), 1785–811.

Graham, B. (2011). Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica 79*(2), 437–52.

Graham, B., C. Pinto and D. Egel (2012). Inverse probability tilting for moment condition models with missing data. *Review of Economic Studies 79*(3), 1053–79.

Grimmer, J., S. Messing and S. Westwood (2017). Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis 25*(4), 413–34.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica 66*(2), 315–31.

Hausman, J. and W. Newey (1995). Nonparametric estimation of exact consumers surplus and deadweight loss. *Econometrica 63*(6), 1445–76.

Hirano, K., G. Imbens and G. Reeder (2003). Efficient estimation of average treatment effects under the estimated propensity score. *Econometrica 71*(4), 1161–89.

Imbens, G. (2000). The role of propensity score in estimating dose-response functions. *Biometrika 87*(3), 706–10.

Jacob, D. (2019). Group average treatment effects for observational studies. *arXiv e-prints*, arXiv:1911.02688.

Javanmard, A. and A. Montanari (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research 2*(4), 2869–909.

Kennedy, E., Z. Ma, M. McHugh and D. Small (2017). Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B 79*(4), 1229–45.

Luo, Y. and M. Spindler (2016). High-dimensional L_2 boosting: rate of convergence. *arXiv e-prints*, arXiv:1602.08927.

Newey, W. (2007). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics 79*(1), 147–68.

Newey, W. (2009). Two-step series estimation of sample selection models. *Econometrics Journal 12*, 217–29.

Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. *Probability and Statistics 213*(57), 416–44.

Oprescu, M., V. Syrgkanis and Z. S. Wu (2018). Orthogonal random forest for causal inference. *arXiv e-prints*, arXiv:1806.03467.

Portnoy, S. and R. Koenker (1989). Adaptive *l*-estimation for linear models. *Annals of Statistics 17*(1), 362–81.

Robins, J. and A. Rotnitzky (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of American Statistical Association 90*(429), 122–9.

Rosenbaum, P. and D. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Schmalensee, R. and T. Stoker (1999). Household gasoline demand in the united states. *Econometrica 67*(3), 645–62.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Annals of Statistics 48*(4), 1875–97.

Semenova, V. and V. Chernozhukov (2018). Estimation and inference about conditional average treatment effect and other structural functions. *arXiv e-prints*, arXiv:1702.06240.

Syrganis, V. and M. Zampetakis (2020). Estimation and inference with trees and forests in high dimensions. *arXiv e-prints*, arXiv:2007.03210.

Tibshirani, J., S. Wager and S. Athey (2017). Generalized random forests. https://grf-labs.github.io/grf/reference/best_linear_projection.html.

van der Geer, S., P. Bühlmann, Y. Ritov and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics 42*(3), 1166–202.

Wager, S. and G. Walther (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv e-prints*, arXiv:1503.06388.

Yatchew, A. and J. A. No (2001). Household gasoline demand in Canada. *Econometrica 69*, 1697–709.

Zhang, C.-H. and S. Zhang (2014). Confidence intervals for low-dimensional parameters in high-dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 217–42.

Zimmert, M. and M. Lechner (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. *arXiv e-prints*, arXiv:1908.08779.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Online Appendix
Replication Package

Managing editor Jaap Abbring handled this manuscript.