JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2022, VOL. 40, NO. 1, 313–327
https://doi.org/10.1080/07350015.2020.1811102

Taylor & Francis
Taylor & Francis Group

Estimation of Conditional Average Treatment Effects With High-Dimensional Data

Qingliang Fan[a], Yu-Chin Hsu[b,c,d], Robert P. Lieli[e], and Yichong Zhang[f]

[a]Department of Economics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong; [b]Institute of Economics, Academia Sinica, Taipei, Taiwan; [c]Department of Finance, National Central University, Taoyuan, Taiwan; [d]Department of Economics, National Chengchi University, Taipei, Taiwan; [e]Department of Economics and Business, Central European University, Budapest, Hungary; [f]School of Economics, Singapore Management University, Singapore

ABSTRACT
Given the unconfoundedness assumption, we propose new nonparametric estimators for the reduced dimensional conditional average treatment effect (CATE) function. In the first stage, the nuisance functions necessary for identifying CATE are estimated by machine learning methods, allowing the number of covariates to be comparable to or larger than the sample size. The second stage consists of a low-dimensional local linear regression, reducing CATE to a function of the covariate(s) of interest. We consider two variants of the estimator depending on whether the nuisance functions are estimated over the full sample or over a hold-out sample. Building on Belloni at al. and Chernozhukov et al., we derive functional limit theory for the estimators and provide an easy-to-implement procedure for uniform inference based on the multiplier bootstrap. The empirical application revisits the effect of maternal smoking on a baby's birth weight as a function of the mother's age.

ARTICLE HISTORY
Received March 2020
Accepted August 2020

KEYWORDS
Heterogeneous treatment effects; High-dimensional data; Uniform confidence band

1. Introduction

In settings with individual-level treatment effect heterogeneity, the unconfoundedness assumption theoretically permits identification and consistent estimation of the conditional average treatment effect (CATE) for all possible values of the set of covariates X used in adjusting for selection bias. One way to think about these covariates is that they are ex-ante predictors of an individual's potential outcomes with and without treatment, and hence are highly correlated with the treatment participation decision as well. Unconfoundedness states that the econometrician observes all relevant predictors so that conditional on X, the treatment takeup decision is no longer statistically related to the potential outcomes.[1] Nevertheless, in many situations the individual deciding on treatment participation is likely to have access to private signals about their potential outcomes. Relying on the unconfoundedness assumption amounts to hoping that a set of publicly observed characteristics can still proxy for the information content of these signals. Therefore, the unconfoundedness assumption is more plausible in applications in which X is a rich, detailed set of covariates, that is, the dimension of X is high.

While CATE as a function of X provides a detailed characterization of treatment effect heterogeneity across observable subpopulations, this information is very hard to analyze and convey if X is high dimensional. Of course, one could examine slices of this function along some component(s) X_1 of X while holding the other components X_{-1} of X constant. Nevertheless, how CATE varies as a function of X_1 will generally depend on the level at which X_{-1} is held constant, requiring the examination of (infinitely) many different slices. For this reason, instead of holding the variables in X_{-1} constant, Abrevaya, Hsu, and Lieli (2015) (henceforth AHL) suggested integrating them out with respect to the conditional distribution of X_{-1} given X_1 or, in practice, a smoothed estimate of this distribution. This gives rise to a reduced dimensional CATE function that is easier to present and interpret.[2]

In this article, we propose two-step estimators of the reduced dimensional CATE function where in the first step the required high-dimensional nuisance regressions are conducted by machine learning methods designed specifically to handle such problems, while the second integration step is implemented by a traditional local linear nonparametric regression.[3] We derive the statistical properties of two variants of the estimator. In the first case, the first step (nuisance function estimation) and the second step (local linear regression) are both implemented over the full sample of available observations. In the second case, the available sample is split into parts, and the first step is implemented in one subsample while the second step is done in

CONTACT Qingliang Fan michaelqfan@gmail.com Department of Economics, The Chinese University of Hong Kong, 903, Esther Lee Building, Shatin, N.T., Hong Kong.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/UBES.

[1] This condition was formalized by Rosenbaum and Rubin (1983); since then, unconfoundedness (or "selection on observables" or "conditional independence") has become one of the standard paradigms for modeling selection effects. See, for example, Imbens and Wooldridge (2009) for further discussion.

[2] If all covariates X are integrated out, one obtains an estimator of ATE as in Hahn (1998) or Hirano, Imbens, and Ridder (2003).

[3] This step assumes that X_1 is a continuous variable, which is the technically challenging and interesting case.

the complement sample. The roles of the subsamples are then rotated and the results are averaged. This is the "cross-fitting" approach to machine-learning-aided causal inference advocated by Chernozhukov et al. (2018). The first approach is used by Belloni et al. (2017) in estimating unconditional treatment effects.

In proposing and studying these estimators, we contribute to two recent strands of the econometrics literature. First, we advance the currently available flexible methods for the estimation of reduced dimensional CATE functions due to AHL and Lee, Okui, and Whang (2017) (henceforth LOW). Second, we make technical contributions to the recent literature that employs machine learning methods in tackling the prediction component of causal inference problems (see, e.g., Belloni, Chernozhukov, and Hansen 2014a, 2014b; Belloni et al. 2017; Chernozhukov et al. 2017, 2018). Taking a broader perspective, our article is also related to a large statistics literature on regular estimation and the use of orthogonal (doubly robust) moment conditions.

Regarding the first set of articles, AHL use an inverse probability weighted conditional moment of the data to identify CATE. They consider both kernel-based and parametric estimation of the propensity score in the first step, and derive the asymptotic distribution of the estimated CATE function evaluated at a fixed point x_1 in the support of X_1. LOW advance these results in two respects: their estimator is based on a Neyman-orthogonal moment condition and they also provide a method for uniform inference about the CATE function as a whole (rather than point by point). While LOW only use parametric models to estimate the nuisance functions involved in the moment condition, orthogonality lends their estimator a "double robustness" property: either the model for the propensity score or the models for the conditional means of the potential outcomes are allowed to be misspecified (but not both).

The CATE estimator proposed here is based on the same orthogonal moment condition as in LOW, but the required nuisance functions are estimated by machine learning methods, which allow for data-driven flexible functional forms as well as a (very) high-dimensional set of covariates. Neyman orthogonality is crucial in ensuring that the proposed CATE estimators are robust to the regularization bias inherent in the first stage, making post-selection inference possible. As the asymptotic theory is derived from high-level assumptions, there are a number of applicable first-stage estimation methods in practice, such as a random forest or ℓ_1-penalized lasso or post-lasso. In this article, we use lasso estimation as the leading example.

In light of the discussion of the unconfoundedness assumption above, replacing the parametric estimators in LOW with machine learning methods greatly enhances the applicability and empirical relevance of flexible CATE estimation. At the same time, the asymptotic theory remains tractable: we provide methods for pointwise as well as uniform inference about the CATE function under both the full sample and sample-splitting implementation schemes. The uniform methods use the multiplier bootstrap, while pointwise inference can be based either on the bootstrap or the analytic results.

Turning to the literature on machine learning in treatment effect estimation, we build primarily on Belloni et al. (2017) for the full-sample method and Chernozhukov et al. (2017, 2018) for the split-sample method, while providing the necessary extension of the theory to account for the use of local linear regression in second step. In these articles the parameter of interest is identified by the restriction that the unconditional expectation of a "score function" evaluated at the true parameter value (and the true nuisance functions) is zero. By contrast, the identifying restriction in our case is that the *conditional* expectation of the same score function is zero. Hence, our estimation procedure does not simply consist of substituting in the estimated nuisance functions and setting the sample average score to zero; instead, the score function will enter a local linear regression with kernel weight $\mathcal{K}((X_{1i} - x_1)/h)/h$ on each observation, where h denotes a smoothing parameter (bandwidth).

The key high-level assumptions we employ in deriving our asymptotic results involve bounding the L_∞ norm of the difference between the true and estimated nuisance functions, and the L_2 norm of the same difference multiplied by the kernel. The rates at which these error bounds are required to converge to zero are closely linked to the rate at which the bandwidth sequence converges to zero. From a purely technical standpoint, incorporating the bandwidth conditions into the high-level norm bounds in the full-sample as well as the cross-fitting case is a central contribution of the article. Similarly to AHL and LOW, the resulting convergence rate of the CATE estimators is $\sqrt{Nh^d}$, where N is the sample size and $d = \dim(X_1)$.

In addition to the error bounds, the full-sample estimator also requires controlling the complexity (entropy) of the function space in which the nuisance functions take values. In the case of lasso estimation, this can be accomplished by restricting how fast the number of covariates and the sparsity indices associated with the nuisance functions are allowed to increase with the sample size. These conditions are more stringent than in the case of estimating ATE.

There are several articles in the broader statistics literature that have considered estimation problems related to ours (Robins 2004; van der Laan 2013; Luedtke and van der Laan 2016a, 2016b; Nie and Wager 2017; Lechner 2019).[4] Nie and Wager (2017), in particular, estimated the full-dimensional CATE function in a data-rich environment using penalized regression, and established the quasi-oracle error bounds for their estimator. While we also use a high-dimensional set of covariates and machine learning methods to deal with selection into treatment, the ultimate parameter of interest, being a function of a low-dimensional subset of the covariates, is then targeted by a traditional nonparametric estimator. We also complement Nie and Wager (2017) by establishing both pointwise and uniform inference procedures. In a related article, Zimmert and Lechner (2019) considered the local constant estimation of CATE in the high-dimensional setting but only provide pointwise asymptotic results.

Another closely related article, Chernozhukov and Semenova (2019), proposes an approach to CATE estimation that also includes a dimension-reduction step. There are, however,

[4] We thank Edward Kennedy and an anonymous referee for these references.

substantial technical differences between their article and ours. First, the traditional nonparametric estimator used by Chernozhukov and Semenova (2019) in the second stage is series regression rather than local linear regression. Second, they only consider the cross-fitting approach and do not address the problem of estimating both the nuisance functions and the target function on the full sample. Third, we also provide a reasonably detailed discussion of the primitive conditions under which lasso estimation fulfills the high-level conditions posited in the article, while Chernozhukov and Semenova (2019) restricted attention to high-level analysis.

Finally, our estimation method based on doubly robust moments is tied to the classic literature on regular estimation and semiparametric efficiency (Begun et al. 1983; Pfanzagl 1990; Bickel et al. 1993; Newey 1994a; van der Vaart 2000). As mentioned above, in a parametric setting, estimation of CATE based on doubly robust moments is consistent as long as either the treatment assignment process or the outcome processes is correctly specified. If both processes are nonparametrically estimated, the method can achieve a faster convergence rate than the nuisance estimators employed. The use of doubly robust methods for causal inference has also been considered by Robins and Rotnitzky (1995), Hahn (1998), van der Laan and Robins (2003), Hirano, Imbens, and Ridder (2003), van der Laan and Rubin (2006), Firpo (2007), Tsiatis (2007), van der Laan and Rose (2011), Belloni et al. (2017), Farrell (2015), Kennedy et al. (2017), Robins et al. (2017), Wager and Athey (2018), and Su, Ura, and Zhang (2019), among others.

In addition to providing theoretical results, we study and illustrate our methods through Monte Carlo simulations. The proposed estimators perform well in terms of bias, MSE, and coverage rates. In general, we find that the cross-fitting estimator has somewhat better finite sample properties than the full sample estimator, and thus we suggest using the cross-fitting estimator in empirical studies with reasonably large sample sizes.

Our application uses vital statistics data from North Carolina to estimate the effect of a (first-time) mother's smoking during pregnancy on the baby's birth weight as a function of the mother's age. Despite a number of previous analyses, the application is well worth revisiting with the help of machine learning methods, as there are a large number of covariates describing the mother's characteristics and events during pregnancy, and the specification of the propensity score is known to have a substantial impact on the results (see AHL, sec. 4.2). Our results provide some corroborating evidence that the negative effect of smoking on birth weight becomes more detrimental with age. This pattern is less prevalent than some of the results reported in AHL but stronger than that found by LOW.

The rest of the article proceeds as follows. In Section 2, we describe the formal setup and the estimators. Section 3 states and discusses the assumptions underlying the first-order asymptotic theory and provides the main results. Section 4 describes how to conduct uniform inference using the multiplier bootstrap. The application is presented in Section 5, while Section 6 concludes. An online supplement contains additional empirical studies, the Monte Carlo exercise as well as detailed proofs of the theoretical results.

2. The Formal Framework, Identification, and the Estimators

Population units are characterized by a random vector $(D, Y(1), Y(0), X)$, where $D \in \{0, 1\}$ indicates the receipt of a binary treatment, $Y(1)$ and $Y(0)$ are the potential outcomes with and without the treatment, respectively, and X is a vector of pretreatment covariates. The observed variables are given by the vector $W = (D, Y, X)$, where $Y = DY(1) + (1 - D)Y(0)$. The distribution of $(D, Y(1), Y(0), X)$, and hence W, is induced by an underlying probability measure $\mathbb{P}$; parameter values computed under $\mathbb{P}$ will be denoted by the subscript "0" and represent the true values of these parameters. The expectation operator corresponding to $\mathbb{P}$ is denoted by $\mathbb{E}$, but we also use the linear functional notation $\mathbb{P}f := \int f(w)d\mathbb{P} = \mathbb{E}[f(W)]$.

To accommodate high-dimensional data, we follow the conceptual considerations in Farrell (2015) and treat the DGP (the measure $\mathbb{P}$) as dependent on the sample size N, allowing, in particular, the dimension of X to grow with N.[5] This has two practical interpretations. First, the number of raw controls can already be comparable to the sample size or, second, X may be composed of a large dictionary of sieve bases derived from a fixed dimensional vector X^* through suitable transformations (e.g., powers and interactions). Thus, the high dimensionality of X can also stem from the desire to provide a flexible approximation to the required nuisance functions. We explicitly allow for the use of lasso-type methods in the first stage that select a smaller subset of terms from the dictionary to approximate these functions.

To ease the already heavy notational burden in the article, the dependence of the DGP on the sample size is left implicit throughout, but is of course accounted for in the theoretical analysis. Most arguments in the article are based on concentration inequalities, which are nonasymptotic in nature. In our Assumption 3.2, we also take into account the fact that, as the dimension of X grows, the complexity of the first-stage estimator will generally diverge, which can affect the rate of convergence of our second-stage estimator. Furthermore, we establish the uniform inference results using the multiplier bootstrap based on the strong approximation theory developed by Chernozhukov, Chetverikov, and Kato (2014), which does not require the existence of an asymptotic distribution.

Given a d-dimensional subvector $X_1 \subset X$ composed of continuous variables, the reduced dimensional CATE function is defined as

$$\tau_0(x_1) = \text{CATE}(x_1) = \mathbb{E}[Y(1) - Y(0)|X_1 = x_1].^{6}$$

The identification of $\tau_0(x_1)$ from the joint distribution of W is facilitated by the unconfoundedness assumption along with some technical conditions:

Assumption 2.1. The distribution $\mathbb{P}$ satisfies:

[5] This implies that the nuisance functions $\mu_0(j, X)$, $j = 0, 1$ and $\pi_0(X)$, to be defined below, may generally depend on N as well.

[6] The most relevant case in practice is $d = 1$ or perhaps $d = 2$, otherwise the motivating properties of the reduced dimensional CATE function (interpretability and presentability) are lost. As we will see below, under a fourth-moment condition on Y, the general theory requires $d \leq 3$. There are no restrictions on d for bounded outcomes.

(i) (Unconfoundedness) $(Y(1), Y(0)) \perp D \big| X$.
(ii) (Moments) $\mathbb{E}\big[|Y(j)|^q\big] < \infty, j = 0, 1$ and $q \geq 4$.
(iii) (Propensity score) Let $\pi_0(x) = \mathbb{P}(D = 1|X = x)$. There exists some constant $\underline{C} > 0$ such that $\mathbb{P}(\underline{C} \leq \pi_0(X) \leq 1 - \underline{C}) = 1$.

Assumption 2.1(i) is the standard unconfoundedness condition. Although we are interested in CATE for a low-dimensional subset X_1 of the covariates, we still use the full vector of X to address selection into treatment. Allowing for X to be high-dimensional makes it more plausible to have conditional independence between the potential outcomes and the treatment indicator. Assumption 2.1(ii) is a usual sufficient condition for the estimation of standard errors. Assumption 2.1(iii) is the overlapping support condition commonly assumed in the literature. We also need it to establish that our CATE(x_1) estimator converges at the usual nonparametric rate.

Let $\mu_0(j, x) = \mathbb{E}[Y|X = x, D = j], j = 0, 1$. It follows immediately from Assumption 2.1 that $E[Y(j)|X_1 = x_1] = E[\mu_0(j, X)|X_1 = x_1]$, and hence $\tau_0(x_1)$ is identified as $\tau_0(x_1) = \mathbb{E}[\mu_0(1, X) - \mu_0(0, X)|X_1 = x_1]$.

We now state a less obvious but more robust result based on a Neyman-orthogonal moment condition. Given any probability measure satisfying Assumption 2.1, let $\tau(\cdot), \mu(1, \cdot), \mu(0, \cdot), \pi(\cdot)$ denote the functions corresponding to $\tau_0(\cdot)$, $\mu_0(1, \cdot)$, $\mu_0(0, \cdot)$, $\pi_0(\cdot)$, respectively. Let $\eta = (\pi(\cdot), \mu(1, \cdot), \mu(0, \cdot))$ represent the infinite dimensional nuisance parameters needed to identify CATE, and define

$$\psi(W; \eta) = \frac{D(Y - \mu(1, X))}{\pi(X)} + \mu(1, X) - \frac{(1 - D)(Y - \mu(0, X))}{1 - \pi(X)} - \mu(0, X).$$

The following theorem gives a moment condition that is (at least approximately) satisfied at (τ_0, η) even when η deviates from η_0.

Theorem 2.1.

(i) Under Assumption 2.1,

$$\mathbb{E}\left[\frac{D(Y - \mu_0(1, X))}{\pi_0(X)} + \mu_0(1, X)\bigg|X_1 = x_1\right] = \mathbb{E}[Y(1) \mid X_1 = x_1]$$

$$\mathbb{E}\left[\frac{(1 - D)(Y - \mu_0(0, X))}{1 - \pi_0(X)} + \mu_0(0, X)\bigg|X_1 = x_1\right] = \mathbb{E}[Y(0) \mid X_1 = x_1]$$

for all x_1 in the support of X_1.

(ii) $\mathbb{E}[\psi(W; \eta_0) - \tau_0(X_1)|X_1 = x_1] = 0$ by part (i), and this moment equation satisfies the Neyman-orthogonality condition

$$\partial_r \mathbb{E}\big[\psi\big(W; \eta_0 + r(\eta - \eta_0)\big) - \tau_0(X_1)\big|X_1 = x_1\big]\big|_{r=0} = 0. \quad (1)$$

Remarks.

1. Assumption 2.1(iii) is not necessary for Theorem 2.1; a weaker moment condition such as $\mathbb{E}[1/\pi_0^2(X)] < \infty$ would suffice. Nevertheless, the overlap condition stated under Assumption 2.1(iii) is indispensable for subsequent results concerned with the asymptotic distribution of our CATE estimators. Similarly, for identification only, the fourth moment condition in Assumption 2.1(ii) could be replaced by a second moment condition.
2. If $\eta = (\pi_0, \mu(0, \cdot), \mu(1, \cdot))$ or $\eta = (\pi, \mu_0(0, \cdot), \mu_0(1, \cdot))$, that is, η deviates from η_0 along one set of coordinates at a time, then $\mathbb{E}\big[\psi\big(W; \eta_0 + r(\eta - \eta_0)\big) - \tau_0(X_1)\big|X_1 = x_1\big] = 0$ for any value of r, which of course implies (1). This is the "double robustness property" emphasized by LOW; it implies that if $\pi(\cdot)$ and $(\mu(0, \cdot), \mu(1, \cdot))$ are parametric models for $\pi_0(\cdot)$ and $(\mu_0(0, \cdot), \mu_0(1, \cdot))$, respectively, and one of these models is misspecified, then one can still consistently estimate $\tau_0(x_1)$ based on the moment condition $\mathbb{E}\big[\psi(W; \eta) - \tau_0(X_1)\big|X_1 = x_1\big] = 0$.

The following assumption describes the properties and use of the sample data:

Assumption 2.2.

(i) The observed data consist of N independent and identically distributed (iid) random vectors $\{W_i\}_{i=1}^N = \{(D_i, Y_i, X_i)\}_{i=1}^N$ with the same distribution as the population distribution of W.
(ii) Let K be a (small) positive integer, and (for simplicity) suppose that $n = N/K$ is also an integer. Let $I_1, \ldots, I_K$ be a random partition of the index set $I = \{1, \ldots, N\}$ so that $\#I_k = n$ for $k = 1, \ldots K$.

We now propose two versions of the CATE estimator, depending on whether the first-stage approximation to η_0 and the second-stage local linear regression targeting τ_0 take place over the same sample or not.

- The full-sample estimator:
  Let $\hat{\eta}(I) = (\hat{\mu}(0, \cdot; I), \hat{\mu}(1, \cdot; I), \hat{\pi}(\cdot; I))$, where $\hat{\mu}(0, \cdot; I)$, $\hat{\mu}(1, \cdot; I)$, and $\hat{\pi}(\cdot; I)$ are estimators of $\mu_0(0, \cdot)$, $\mu_0(1, \cdot)$, and $\pi_0(\cdot)$, respectively, over the full sample I. Furthermore, let $\mathcal{K}$ be a d-dimensional product kernel, h be a smoothing parameter (bandwidth), and $\mathcal{K}_h(u) = \mathcal{K}\left(\frac{u}{h}\right)$.[7] The second stage of the full-sample estimator $\hat{\tau}(x_1)$ is obtained as the intercept of the local linear regression

$$(\hat{\tau}(x_1), \hat{\beta}(x_1)) = \arg\min_{a,b} \sum_{i \in I} \big[\psi(W_i, \hat{\eta}(I)) - a - (X_{1i} - x_1)'b\big]^2 \mathcal{K}_h(X_{1i} - x_1). \quad (2)$$

- The K-fold cross-fitting estimator:
  For each $k = 1, \ldots, K$, let $\hat{\eta}(I_k^c) = (\hat{\mu}(0, \cdot; I_k^c), \hat{\mu}(1, \cdot; I_k^c), \hat{\pi}(\cdot; I_k^c))$, where $\hat{\mu}(0, \cdot; I_k^c)$, $\hat{\mu}(1, \cdot; I_k^c)$, and $\hat{\pi}(\cdot; I_k^c)$ are estimators for $\mu_0(0, \cdot)$, $\mu_0(1, \cdot)$, and $\pi_0(\cdot)$, respectively, constructed over the subsample $I_k^c = I \setminus I_k$. The second stage of the K-fold cross-fitting estimator consists of

[7] To be specific, let $k_j(\cdot)$ be a one-dimensional kernel, then a d-dimensional product kernel $\mathcal{K}$ where bandwidth h is defined as $\mathcal{K}_h(u) = \Pi_{j=1}^d k_j(u_j/h)$. More generally, we can allow h to be different for each j such that $\mathcal{K}_h(u) = \Pi_{j=1}^d k_j(u_j/h_j)$ given that h_j's are of the same order. For notational simplicity, we focus on the first case in the theory.

K nonparametric regressions over the samples $I_1, \dots, I_K$:

$$(\hat{\tau}_k(x_1), \hat{\beta}_k(x_1)) = \arg\min_{a,b} \sum_{i \in I_k} \left[\psi(W_i, \hat{\eta}(I_k^c)) - a - (X_{1i} - x_1)'b\right]^2 \mathcal{K}_h\,(X_{1i} - x_1)\,. \quad (3)$$

Finally, in the third stage we take the average of the K preliminary estimates to obtain an efficient estimator: $\check{\tau}(x_1) = \frac{1}{K}\sum_{k=1}^{K} \hat{\tau}_k(x_1)$.

We use the local linear smoother studied by Fan (1992, 1993) and Fan and Gijbels (1992) to estimate $\tau_0(x_1)$ in the second stage. While it is possible to extend our results to local polynomial estimators with an extra degree of smoothness, we focus on the linear case for simplicity.[8]

Our second-stage estimator is related to the partial mean estimator studied by Newey (1994b) and Lee (2018). However, Lee (2018) and our article are distinct in two important ways. First, the parameters of interest, and thus the estimators, are different. We are interested in CATE(x_1) when the treatment variable is binary, while Lee (2018) considered a model with a continuous treatment. Second, as explained above, our analysis is compatible with the use of high-dimensional data. For the full-sample first-stage estimation, we allow the complexity of our first-stage estimator to increase with the dimensionality of the data and investigate its impact on the rate of convergence. For the split-sample first-stage estimation, we show that the impact of the increasing complexity is eliminated due to the independence between the observations used in the first- and second-stage estimations.

When X_1 is discrete and takes the values $x_{1,1}, \dots, x_{1,M}$, the function CATE($x_{1,m}$), $m = 1, \dots, M$ can be interpreted as the average treatment effect for the subpopulation $X_1 = x_{1,m}$. In this case one can restrict the sample to observations with $X_1 = x_{1,m}$, and directly apply the full-sample or cross-fitting estimation methods developed in Belloni et al. (2017) and Chernozhukov et al. (2017).

3. Asymptotic Properties of CATE Estimators

3.1. CATE Estimators Based on General First-Step ML Estimators

In this section, we first provide the fundamental asymptotic results for our CATE estimators which form the basis of the uniform inference procedures to be given in Section 4. To this end, we state and discuss several assumptions. Let $\mathcal{X}_1 \subset \mathbb{R}^d$ denote the support of X_1 and let $\overline{\mathcal{X}}_1$ be the subset of $\mathcal{X}_1$ over which $\tau_0(x_1)$ is to be estimated. In addition, let $f(x_1)$ denote the p.d.f. of X_1.

Assumption 3.1. Assume that

(i) The set $\overline{\mathcal{X}}_1$ is contained in the interior of $\mathcal{X}_1$ and is the Cartesian product of closed intervals, that is, $\overline{\mathcal{X}}_1 = \Pi_{j=1}^{d}[x_{1\ell}^{(j)}, x_{1u}^{(j)}]$ with $x_{1\ell}^{(j)} < x_{1u}^{(j)}$. Furthermore, there exist positive constants $\underline{C}$ and $\overline{C}$ such that:

$$\underline{C} \le \inf_{x_1 \in \overline{\mathcal{X}}_1} f(x_1) \le \sup_{x_1 \in \overline{\mathcal{X}}_1} f(x_1) \le \overline{C} \quad \text{and}$$

$$\sup_{x_1 \in \overline{\mathcal{X}}_1} (|\mathbb{E}[Y(1)|X_1 = x_1]| + |\mathbb{E}[Y(0)|X_1 = x_1]|) \le \overline{C}.$$

(ii) The functions $f(x_1)$, $\mathbb{E}[Y(0)|X_1 = x_1]$, and $\mathbb{E}[Y(1)|X_1 = x_1]$ are twice differentiable with bounded derivatives over $\overline{\mathcal{X}}_1$; more formally,

$$\sup_{x_1 \in \overline{\mathcal{X}}_1, 1 \le j,s \le d} \Big(|\partial_j f(x_1)| + |\partial_{j,s} f(x_1)| + |\partial_j \tau_0(x_1)| + |\partial_{j,s} \tau_0(x_1)| \Big) \le \overline{C},$$

where $\partial_{j,s} f(x_1)$ is the derivative of $f(x_1)$ w.r.t. x_{1j} and x_{1s}.

(iii) For $u \in \mathbb{R}^d$, $\mathcal{K}(u) = \kappa(u_1) \times \dots \times \kappa(u_d)$, where κ is a bounded, symmetric p.d.f. with $\int t\kappa(t)dt = 0$ and $\int t^2\kappa(t)dt = \nu < \infty$. Furthermore, there exists a positive constant $\overline{C}$ such that $|t|\kappa\,(t) \le \overline{C}$ for all $t \in \mathbb{R}$.

(iv) The bandwidth $h = h_N$ satisfies $h = CN^{-H}$ for some $H > 1/(4+d)$ and $H < (1-2/q)/d$, where $C > 0$ and q satisfies Assumption 2.1(ii).

(v) Let $\beta_0(x_1) = \partial_{x_1}\tau_0(x_1)$ and $\tau_0^{(2)}(x_1) = \partial_{x_1 x_1^T}\tau_0(x_1)$. Then $\sup_{x_1 \in \overline{\mathcal{X}}_1} \lambda_{\max}(\tau_0^{(2)}(x_1)) < \overline{C}$, where $\lambda_{\max}(G)$ denotes the maximum singular value of matrix G. In addition, we have

$$\sup_{x_1, x_1' \in \overline{\mathcal{X}}_1} \frac{\left|\tau_0(x_1') - \tau_0(x_1) - (x_1' - x_1)^T\beta_0(x_1) - \frac{1}{2}(x_1' - x_1)^T\tau_0^{(2)}(x_1)(x_1' - x_1)\right|}{||x_1' - x_1||_2^3} \le \overline{C},$$

where $||\cdot||_2$ denotes the Euclidean norm of a vector.

For the most part, Assumption 3.1 is a collection of standard regularity conditions used in the nonparametric treatment effect estimation literature. The functions $f(x_1)$, $\mu_0(0, x_1)$, and $\mu_0(1, x_1)$ are required to be sufficiently smooth over $\overline{\mathcal{X}}_1$, the density of X_1 must be bounded away from zero over the same set, and the kernel function $\mathcal{K}$ must obey some mild restrictions, satisfied by usual choices of κ such as the Gaussian or the Epanechnikov kernel (in the simulations and the empirical study we use the former). Of course, Assumption 3.1(ii) also implies that we restrict attention to the technically more interesting case in which the distribution of X_1 is continuous, which means that one cannot simply use sample splitting to estimate CATE at various points in the support of X_1.

The conditions imposed on the bandwidth in Assumption 3.1(iv) are motivated as follows. The restriction $H > 1/(4+d)$ means that h converges to zero faster than the MSE-optimal bandwidth choice; this undersmoothing condition is needed to ensure that the bias from the second-stage kernel regression is asymptotically negligible. In addition, we require $H < (1-2/q)/d$ to be able to use a Gaussian approximation as in Chernozhukov, Chetverikov, and Kato (2014, Proposition 3.2). If the outcome variable is bounded, one can set $q = \infty$ in Assumption 2.1(ii) so that $H < 1/d$ as in Chernozhukov, Chetverikov, and Kato (2014, Proposition 3.1). If one only

[8] An early version of the article considered kernel-based (local constant) nonparametric regression in the second stage. The results are available upon request.

assumes $q = 4$, then the convergence rate must satisfy $H \in (1/(4+d), 1/2d)$. For this interval to be nonempty, d can be at most 3, which is consistent with Assumption 1 in LOW. In principal, it would also be possible to use the optimal bandwidth, that is, $H = 1/(4+d)$, and conduct bias correction as in Calonico, Cattaneo, and Farrell (2018), while accounting for the impact of the estimated bias correction term on the standard error. This approach is, however, beyond the scope of the present article. A key conceptual difference between our setup and Calonico, Cattaneo, and Farrell (2018) is that in our case the dependent variable is not directly observed, but is rather constructed based on first-stage nuisance estimators. Finally, Assumption 3.1(v) is a standard bound for the Taylor remainder, commonly assumed in local linear regression theory. See, for example, Li and Racine (2007, chap. 2).

We now state high-level conditions that specify the convergence rates required of the first-stage nuisance function estimators. The stated rates are linked to the bandwidth sequence h used in the second-stage regressions. More specifically, we make the following assumption about the full-sample first-stage estimator $\hat{\eta}(I)$.

Assumption 3.2 (Full sample, first stage). Let δ_{1N}, δ_{2N}, δ_{4N}, and A_N be sequences of positive numbers, and $\mathcal{G}_N^{(j)}, j \in \{0, 1, \pi\}$ be classes of real-valued functions defined on the support of X with corresponding envelope functions $G_N^{(j)}, j \in \{0, 1, \pi\}$. For $\epsilon > 0$, let $\mathcal{N}(\mathcal{G}_N^{(j)}, \|\cdot\|, \epsilon)$ be the covering number associated with $\mathcal{G}_N^{(j)}$ under some norm $\|\cdot\|$ defined on $\mathcal{G}_N^{(j)}$.[9] The following conditions are satisfied.

(i) The estimator $\hat{\eta}(I)$ obeys the error bounds

$$\sup_{x_1 \in \overline{\mathcal{X}}_1, j=0,1} \left\| (\hat{\mu}(j, X; I) - \mu_0(j, X)) \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P},2} \times \left\| (\hat{\pi}(X; I) - \pi_0(X)) \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P},2} = O_p(\delta_{1N}^2) \quad (4)$$

$$\sum_{j=0,1} \left\| \hat{\mu}(j, \cdot; I) - \mu_0(j, \cdot) \right\|_{\mathbb{P},\infty} + \left\| \hat{\pi}(X; I) - \pi_0(X) \right\|_{\mathbb{P},\infty} = O(\delta_{2N}). \quad (5)$$

$$\sup_{x_1 \in \overline{\mathcal{X}}_1, j=0,1} \left\| (\hat{\mu}(j, X; I) - \mu_0(j, X)) ||X_1 - x_1||_2^{1/2} \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P},2} \times \left\| (\hat{\pi}(X; I) - \pi_0(X)) ||X_1 - x_1||_2^{1/2} \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P},2} = O_p(\delta_{3N}^2). \quad (6)$$

(ii) With probability approaching one,

$$\hat{\mu}(j, \cdot; I) \in \mathcal{G}_N^{(j)}, \quad j = 0, 1, \quad \text{and} \quad \hat{\pi}(\cdot; I) \in \mathcal{G}_N^{(\pi)},$$

where the classes of functions $\mathcal{G}_N^{(j)}, j \in \{0, 1, \pi\}$ are such that

$$\sup_Q \log \mathcal{N}\big(\mathcal{G}_N^{(j)}, ||\cdot||_{Q,2}, \varepsilon ||G_N^{(j)}||_{Q,2}\big) \le \delta_{4N}(\log(A_N) + \log(1/\varepsilon) \vee 0), \quad j = 0, 1, \pi \quad (7)$$

with the supremum taken over all finitely supported discrete probability measures Q.

(iii) The sequences δ_{1N}, δ_{2N}, δ_{3N}, δ_{4N} and A_N satisfy:

$$\min(\delta_{1N}/h^{d/2}, \delta_{2N}) = o\big((\log(N) N h^d)^{-1/4}\big), \quad \delta_{2N} = o(1), \quad (8)$$

$$\min(\delta_{3N}/h^{d/2+1}, \delta_{2N}) = o\big((\log(N) N h^{d+2})^{-1/4}\big), \quad (9)$$

$$\delta_{4N} \log(A_N \vee N) \log(N) \delta_{2N}^2 = o(1), \quad (10)$$

$$\text{and} \quad \delta_{2N} \delta_{4N} \log^{1/2}(N) N^{1/q} \log(A_N \vee N) = o((N h^d)^{1/2}). \quad (11)$$

Remarks.

1. Part (i) of Assumption 3.2 controls the difference between η_0 and $\hat{\eta}(I)$ (i.e., the estimation error) in various norms.
2. Part (ii) controls the complexity of the nuisance functions and the estimators through restrictions on the entropy of the classes $\mathcal{G}_N^{(j)}$.
3. It is of course part (iii) that fills parts (i) and (ii) with content through specifying the behavior of the sequences δ_{1N}, δ_{2N}, δ_{4N}, and A_N. In particular, conditions (8) and (9) extend the fairly standard requirement in semiparametric settings that the first-stage nuisance function estimators converge faster than $N^{-1/4}$; see Ai and Chen (2003) and Belloni et al. (2017). However, in estimating $\tau_0(x_1)$ and $\beta_0(x_1)$, the second-stage kernel regression relies only on observations local to x_1, and hence the relevant effective sample size is Nh^d and Nh^{d+2} rather than N. The extra $\log(N)$ factor that appears in the required convergence rate is the price to pay for uniform results in x_1.
4. If the first-stage estimators are based on (correctly specified) parametric models, then, under standard regularity conditions, $\hat{\eta}(I)$ converges to η_0 at the rate of $N^{1/2}$ both in L_2 and L_∞ norm. Thus, in this case (4) and (5) both hold with $\delta_{1N} = \delta_{2N} = N^{-1/2}$ (recall that $\mathcal{K}_h$ is bounded). In addition, conditions (8), (10), and (11) are also easily satisfied with $\delta_{4N} = O(1)$, and $A_N = O(1)$. This is essentially the setting in LOW (with allowance for partial misspecification).
5. Assumption 3.2 imposes rate restrictions on the complexity of the first-stage estimators. The lasso-type regularization method achieves variable selection along with estimation, which greatly reduces the complexity of the estimator. Thus, it is especially suitable for first-stage estimation when using the full sample.
6. It is possible to establish sufficient conditions on the convergence rate of $\hat{\pi}$, $\hat{\mu}$ and the kernel individually. For example, following Kennedy et al. (2017), we can assume

$$\sup_{x_1 \in \overline{\mathcal{X}}_1} \sqrt{\frac{\mathbb{E}\left[\left(\hat{\mu}(j, X; I) - \mu_0(j, X)\right)^2 | X_1 = x_1\right]}{\mathbb{E}\left[\left(\hat{\pi}(X; I) - \pi_0(X)\right)^2 | X_1 = x_1\right]}} = O_p(\delta_{5N}^2).$$

Note here we require the bound to hold uniformly over $\overline{\mathcal{X}}_1$ rather than a neighborhood of x_1, because we aim to conduct uniform inference over $\overline{\mathcal{X}}_1$. Then, it is easy to see that our $\delta_{1N} = \delta_{5N} h^{d/2}$ and $\delta_{3N} = \delta_{5N} h^{(d+1)/2}$. Consequently, (8) and (9) reduce to $\delta_{5N} = o\big((\log(N) N h^d)^{-1/4}\big)$ as $\delta_{2N} \le \delta_{5N}$. However, this condition is sufficient but necessary. It is possible to bound directly the estimation error of the nuisance parameters weighted by the kernel, as shown in Su, Ura, and Zhang (2019).

[9] The covering number is the minimal number of balls with radius ϵ needed to cover $\mathcal{G}_N^{(j)}$. A ball with radius ϵ centered on g is the collection of functions $g' \in \mathcal{G}_N^{(j)}$ with $\|g' - g\| < \epsilon$.

7. Alternatively, because the kernel function is bounded, we can write
$$\sup_{x_1 \in \overline{\mathcal{X}}_1, j=0,1} \left\| (\hat{\mu}(j, X; I) - \mu_0(j, X)) \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P},2} \left\| (\hat{\pi}(X; I) - \pi_0(X)) \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P},2}$$
$$\leq M \max_{j=0,1} \left\| (\hat{\mu}(j, X; I) - \mu_0(j, X)) \right\|_{\mathbb{P},2} \left\| (\hat{\pi}(X; I) - \pi_0(X)) \right\|_{\mathbb{P},2}.$$
for some constant $M > 0$. Thus, one could also state sufficient conditions for (4) solely in terms of the L_2-norm of the error bounds associated with $\hat{\mu}(j, X; I)$ and $\hat{\pi}(X; I)$.
8. Note that we only require bounds on the product of the L_2-norms of two estimation errors as in (4) and (6). The product structure of these conditions allows for tradeoffs between how fast $\hat{\pi}(\cdot; I)$ versus $\hat{\mu}(j, \cdot; I)$ converges.

The corresponding assumption about the cross-fitting (split-sample) estimator is as follows.

Assumption 3.3 (Split sample, first stage). The split-sample first-stage estimators $\hat{\eta}(I_k^c)$, $k = 1, \ldots, K$ are assumed to satisfy:
$$\sup_{x_1 \in \overline{\mathcal{X}}_1} \left\{ \left\| (\hat{\pi}(X; I_k^c) - \pi_0(X)) \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P}_{I_k},2} \times \left\| (\hat{\mu}(j, X; I_k^c) - \mu_0(j, X)) \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P}_{I_k},2} \right\} = O_p\big(\delta_{1n}^2\big), \tag{12}$$
$$||\pi_0(X) - \hat{\pi}_0(X; I_k^c)||_{\mathbb{P},\infty} + \sum_{j=0,1} ||\mu_0(j, X) - \hat{\mu}(0, X; I_k^c)||_{\mathbb{P},\infty} = O(\delta_{2n}), \tag{13}$$
$$\left\| (\hat{\mu}(j, X; I_k^c) - \mu_0(j, X)) ||X_1 - x_1||_2^{1/2} \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P}_{I_k},2} \times \left\| (\hat{\pi}(X; I_k^c) - \pi_0(X)) ||X_1 - x_1||_2^{1/2} \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P}_{I_k},2} = O_p(\delta_{3n}^2), \tag{14}$$
where $\mathbb{P}_{I_k} f = \mathbb{E}(f(W_1, \ldots, W_N) | W_i, i \in I_k^c)$ for a generic function f, $h^{-d}\delta_{1n}^2 = o((\log(n)nh^d)^{-1/2})$, $\delta_{2n} = o((\log(n))^{-1})$, and $h^{-d-2}\delta_{3n}^2 = o((\log(n)nh^{d+2})^{-1/2})$.

Remarks.

1. Because K is fixed and $n = N/K$, $\log(N)Nh^d$ and $\log(n)nh^d$ have the same order of magnitude.
2. For the split-sample estimation, there is no requirement on the entropy of the space where the estimated nuisance functions take values. This weakening of the theoretical conditions is due to the fact that, because of the cross-fitting technique, we can treat the estimators of the nuisance parameters as fixed by conditioning on the subsample of the data used for the estimation.
3. Assumption 3.3 does not impose restrictions on the complexity of the first-stage estimator and thus accommodates various machine learning methods. One can verify Assumption 3.3 given the error bounds of machine learning first-stage estimators in both L_∞ and L_2 norms by the same argument as described in Section 3.2. Deriving these error bounds for various machine learning methods is beyond the scope of our article. Partial results are available in the literature. For example, the L_2 bounds for the random forest method and deep neural networks have already been established in Wager and Athey (2018) and Farrell, Liang, and Misra (2018), respectively.
4. Remarks 6–8 after Assumption 3.2 apply here as well. In essence, Assumption 3.3 is the local analog of Assumption 5.1(f) used by Chernozhukov et al. (2018) to estimate the *unconditional* average treatment effect via the cross-fitting (split-sample) approach.
5. Similarly to Remark 7 after Assumption 3.2, one sufficient condition for the requirement on δ_{1N} is that
$$\sqrt{n/h^d} \max_{j=0,1} \left\| (\hat{\mu}(j, X; I_k^c) - \mu_0(j, X)) \right\|_{\mathbb{P},2} \left\| (\hat{\pi}(X; I_k^c) - \pi_0(X)) \right\|_{\mathbb{P},2} = o((\log(n)^{-1/2})).$$

Chernozhukov and Semenova (2019) consider sieved estimation of CATE with high-dimensional control variables and require
$$\sqrt{nr} \max_{j=0,1} \left\| (\hat{\mu}(j, X; I_k^c) - \mu_0(j, X)) \right\|_{\mathbb{P},2} \left\| (\hat{\pi}(X; I_k^c) - \pi_0(X)) \right\|_{\mathbb{P},2} = o(1),$$
where r is the dimension of the sieve bases. In nonparametric estimation, we know the variances of sieve- and kernel-based estimators are of order r/n and $1/(nh^d)$, respectively. This implies our rate requirement is equivalent to Chernozhukov and Semenova (2019, Assumption 4.4) up to some logarithmic factor. The requirement on δ_{3N} is not essential and can be avoided if one uses the local constant regression instead.

Theorem 3.1.

(a) If Assumptions 2.1, 2.2, 3.1, and 3.2 are satisfied, then
$$\hat{\tau}(x_1) - \tau_0(x_1) = (\mathbb{P}_N - \mathbb{P}) \left[\frac{1}{h^d f(x_1)} (\psi(W, \eta_0) - \tau_0(x_1)) \mathcal{K}_h(X_1 - x_1) \right] + R_\tau(x_1),$$
where $\mathbb{P}_N f = \frac{1}{N} \sum_{i=1}^N f(W_i)$ for a generic function $f(\cdot)$ and $\sup_{x_1 \in \overline{\mathcal{X}}_1} |R_\tau(x_1)| = o_p((\log(N)Nh^d)^{-1/2})$.

(b) If Assumptions 2.1, 2.2, 3.1, and 3.3 are satisfied, then the representation established in part (a) also holds for the K-fold cross-fitting estimator $\check{\tau}(x_1)$, that is,
$$\check{\tau}(x_1) - \tau_0(x_1) = (\mathbb{P}_N - \mathbb{P}) \Big[\frac{1}{h^d f(x_1)} (\psi(W, \eta_0) - \tau_0(x_1)) \mathcal{K}_h(X_1 - x_1) \Big] + \check{R}_\tau(x_1),$$
where $\sup_{x_1 \in \overline{\mathcal{X}}_1} |\check{R}_\tau(x_1)| = o_p((\log(N)Nh^d)^{-1/2})$.

Theorem 3.1 provides the linear (Bahadur) representations of the nonparametric estimators $\hat{\tau}(x_1)$ and $\check{\tau}(x_1)$ with uniform control of the remainder terms. It serves as a building block for both pointwise and uniform inference about $\tau_0(x_1)$.[10] Starting with the former, we define

$$\sigma_N^2(x_1) = h^d \mathrm{var}\left(\frac{1}{h^d f(x_1)}(\psi(W,\eta_0) - \tau_0(x_1))\mathcal{K}_h(X_1 - x_1)\right),$$

and suppose that $\sigma_N^2(x_1)$ satisfies:

Assumption 3.4. There exists some $\underline{C} > 0$ such that $\min_{x_1 \in \overline{\mathcal{X}}_1} \sigma_N^2(x_1) \geq \underline{C}$ for all N.

Then Theorem 3.1, together with Lyapunov's CLT, implies

$$\frac{\sqrt{Nh^d}\big(\hat{\tau}(x_1) - \tau_0(x_1)\big)}{\sigma_N(x_1)} \xrightarrow{d} \mathcal{N}(0,1) \qquad (15)$$

for any fixed $x_1 \in \overline{\mathcal{X}}_1$. One can estimate the variance $\sigma_N^2(x_1)$ as

$$\hat{\sigma}_N^2(x_1) = \frac{1}{Nh^d \hat{f}^2(x_1; I)} \sum_{i=1}^{n} \big(\psi(W_i, \hat{\eta}(I)) - \hat{\tau}(x_1)\big)^2 \mathcal{K}_h^2(X_{1i} - x_1),$$

and we will show that

$$\sup_{x_1 \in \overline{\mathcal{X}}_1} |\widehat{\sigma}_N(x_1) - \sigma_N(x_1)| = o_p(1) \text{ and}$$
$$\sup_{x_1 \in \overline{\mathcal{X}}_1} |\widehat{\sigma}_N^{-1}(x_1) - \sigma_N^{-1}(x_1)| = o_p(1).$$

Of course, this means that inference in practice can proceed based on (15) with $\hat{\sigma}_N(x_1)$ replacing $\sigma_N(x_1)$. Furthermore, result (15) remains valid if one uses the estimator $\check{\tau}(x_1)$ in place of $\hat{\tau}(x_1)$; in this case $\sigma_N^2(x_1)$ can be estimated as

$$\check{\sigma}_N^2(x_1) = \frac{1}{K}\sum_{k=1}^{K} \check{\sigma}_k^2(x_1), \text{ where}$$
$$\check{\sigma}_k^2(x_1) = \frac{1}{nh^d}\sum_{i \in I_k} \frac{1}{\hat{f}^2(x_1; I_k)} \big(\psi(W_i, \hat{\eta}(I_k^c))$$
$$- \check{\tau}_k(x_1)\big)^2 \mathcal{K}_h^2(X_{1i} - x_1).$$

Theorem 3.2. If Assumptions in Theorems 3.1 and Assumption 3.4 hold, then

$$\sup_{x_1 \in \overline{\mathcal{X}}_1} |\widehat{\sigma}_N(x_1) - \sigma_N(x_1)| = o_p(1),$$
$$\sup_{x_1 \in \overline{\mathcal{X}}_1} |\widehat{\sigma}_N^{-1}(x_1) - \sigma_N^{-1}(x_1)| = o_p(1),$$
$$\sup_{x_1 \in \overline{\mathcal{X}}_1} |\check{\sigma}_N(x_1) - \sigma_N(x_1)| = o_p(1), \text{ and}$$
$$\sup_{x_1 \in \overline{\mathcal{X}}_1} |\check{\sigma}_N^{-1}(x_1) - \sigma_N^{-1}(x_1)| = o_p(1).$$

As can be seen from the proof, the $o_p(1)$ term actually vanishes polynomially in N.

3.2. CATE Estimators Based on First-Stage Lasso Estimators

While the high-level assumptions stated in Section 3.1 can accommodate multiple machine learning procedures for estimating η_0, here we describe the first stage using lasso estimation as a leading example. We now discuss some primitive conditions under which lasso estimation of η_0 will satisfy Assumptions 3.2 and 3.3. Specifically, let $b(X) = (b_1(X), \dots, b_p(X))$ be a dictionary of control terms based on X, where p is potentially larger than the sample size N and can grow with N.[11] Typically, $b(X)$ consists of X, and powers and interactions of the components of X. The lasso approximates the nuisance functions η_0 with linear combinations of the components $b_i(X)$; in particular, for p-vectors β, α, and θ, set

$$\begin{aligned} r_\alpha(x) &:= \mu_0(0,x) - b(x)'\alpha, \\ r_\beta(x) &:= \mu_0(1,x) - b(x)'\beta, \\ r_\theta(x) &:= \pi_0(x) - \Lambda(b(x)'\theta), \end{aligned} \qquad (16)$$

where $\Lambda(\cdot)$ is the logistic c.d.f. A primitive condition that justifies using the lasso is approximate sparsity. Intuitively, this means that it is possible to make the approximation errors r_α, r_β, r_θ small with just a small number of approximating terms, that is, with α, β and θ having only a handful of nonzero components.[12] The coefficients α, β, and θ are estimated by penalized least squares or maximum likelihood, where a penalty is imposed for any nonzero component.

In the lasso computations, we set the tuning parameter to be $2c\sqrt{N}\Phi^{-1}(1 - 0.1/(\log(N)2p))$ and $c\sqrt{N}\Phi^{-1}(1 - 0.1/(\log(N)4p))$ for the conditional mean and propensity score functions estimation, respectively, following Belloni, Chernozhukov, and Hansen (2014b) and Belloni et al. (2017).[13]

To formalize the idea that the dimension p of $b(X)$ is comparable with or larger than the sample size, we let $p = p_N$ be a function of N and allow p_N to grow to infinity as N increases, possibly (much) faster than N. For example, one could set $p_N = O(N^\lambda)$ for any $\lambda > 0$, but even $\log(p_N) = O(N^\lambda)$ is allowed if λ is not too large. The linear approximation errors to the components of η_0, defined in display (16), will be controlled by the sparsity index $s = s_N$, a nondecreasing sequence of positive numbers potentially converging to infinity with N. Also needing control is the upper bound on the components of $b(X)$; to this end, let $\zeta = \zeta_N = \max_{1 \leq j \leq p_N} \|b_j(X)\|_{\mathbb{P},\infty}$, and note that ζ (weakly) increases as $p_N \to \infty$. The following assumption formalizes the notion of approximate sparsity.

Assumption 3.5. Let s_π and s_μ denote the individual sparsity index sequences associated with $\pi_0(\cdot)$ and $\mu_0(j,\cdot)$, respectively. There exist sequences of coefficients $\alpha = \alpha_N$, $\beta = \beta_N$ and $\theta = \theta_N$ such that the linear approximations defined in (16) satisfy the following conditions.

[10] In the online supplement, we provide the linear (Bahadur) representations of the nonparametric estimators $\hat{\beta}(x_1)$ and $\check{\beta}(x_1)$ with uniform control of the remainder terms, which can be of independent interest.

[11] To be fully consistent with the general notation, it would be more precise to denote the dictionary as $X = b(X^*) = (b_1(X^*), \dots, b_p(X^*))$; see the discussion in the second paragraph of Section 2. We opt for simplicity at a small cost in notational consistency.

[12] The linear index structure and approximate sparsity are specific to the lasso; other machine learning methods provide different types of approximations which do not necessarily rely on sparsity.

[13] The constant value is usually chosen as $c = 1.1$. Also in practice, one could use cross-validations to choose the tuning parameter here.

(i) The number of nonzero coefficients is bounded by s_μ and s_π, that is, $\max\{\|\alpha\|_0, \|\beta\|_0\} \le s_\mu$ and $\|\theta\|_0 \le s_\pi$.

(ii) The approximation errors are asymptotically small in the sense that

$$||r_\alpha(X)||_{\mathbb{P},2} + ||r_\beta(X)||_{\mathbb{P},2} = O\left(\sqrt{s_\mu \log(p)/N}\right),$$

$$||r_\alpha(X)||_{\mathbb{P},\infty} + ||r_\beta(X)||_{\mathbb{P},\infty} = O\left(\sqrt{s_\mu^2 \zeta^2 \log(p)/N}\right),$$

$$||r_\theta(X)||_{\mathbb{P},2} = O\left(\sqrt{s_\pi \log(p)/N}\right),$$

$$||r_\theta(X)||_{\mathbb{P},\infty} = O\left(\sqrt{s_\pi^2 \zeta^2 \log(p)/N}\right),$$

where $s_\mu^2 \zeta^2 \log(p)/N \to 0$ and $s_\pi^2 \zeta^2 \log(p)/N \to 0$ (and therefore $s_\mu \log(p)/N \to 0$ and $s_\pi \log(p)/N \to 0$).

Part (i) of Assumption 3.5 states that the number of nonzero coefficients in the $b(X)$-based linear approximations to η_0 is at most s. Part (ii) requires that the approximation errors associated with these linear combinations asymptotically vanish both in L_2 and L_∞ norm. This generally requires $s \to \infty$, but s needs to stay small relative to N in the sense that $s^2\zeta^2 \log(p)/N \to 0$.

Given Assumption 3.5 and additional regularity conditions, results by Belloni et al. (2017) imply that conditions (4) and (5) hold with

$$\begin{aligned}\delta_{1N}^2 \le & \sup_{x_1 \in \overline{\mathcal{X}}_1, j=0,1} \left\|(\hat{\mu}(j, X; I) - \mu_0(j, X))\right\|_{\mathbb{P},2} \\ & \left\|(\hat{\pi}(X; I) - \pi_0(X))\right\|_{\mathbb{P},2} \le (s_\mu s_\pi)^{1/2} \log(p \vee N)/N, \\ \delta_{2N}^2 =& (s_\mu^2 + s_\pi^2)\zeta^2 \log(p \vee N)/N, \text{ and} \\ \delta_{3N}^2 \le & (s_\mu s_\pi)^{1/2} \log(p \vee N)h/N, \end{aligned} \tag{17}$$

where the last inequality holds because $||X_1 - x_1||_2 \mathcal{K}_h(X_1 - x_1) \lesssim h$. Furthermore, Belloni et al. (2017) also established (7) with $\delta_{4N} = s$ and $A_N = p$ for the following function classes:

$$\begin{aligned}\mathcal{G}_N^{(j)} &= \{b(X)'\beta : ||\beta||_0 \le \ell_N s_\mu, \\ & \sup_{x \in \overline{X}} |b(x)'\beta - \mu_0(j, x)| \le M\delta_{2N}\}, \quad j = 0, 1, \\ \mathcal{G}_N^{(\pi)} &= \{\Lambda(b(X)'\theta) : ||\beta||_0 \le \ell_N s_\pi, \\ & \sup_{x \in \overline{X}} |\Lambda(b(x)'\theta) - \pi_0(x)| \le M\delta_{2N}\},\end{aligned}$$

where ℓ_N is some slowly diverging sequence, for example, $\ell_N = \log(\log(N))$ and $M > 0$. (As $\pi_0(\cdot)$, $\mu_0(1, \cdot)$, and $\mu_0(0, \cdot)$ are uniformly bounded, $\mathcal{G}_N^{(0)}, \mathcal{G}_N^{(1)}, \mathcal{G}_N^{\pi}$ have bounded envelope functions.)

Given these results, Assumption 3.2 with first-stage lasso estimation boils down to the following conditions:

$$\min\left(\frac{(s_\mu s_\pi)^{1/2} \log(p \vee N) \log^{1/2}(N)}{(Nh^d)^{1/2}}, \frac{\zeta^2 (s_\mu^2 + s_\pi^2) \log(p \vee N) \log^{1/2}(N) h^{d/2}}{N^{1/2}}\right) = o(1),$$

$$\frac{\zeta^2 (s_\mu + s_\pi)^3 \log^2(p \vee N) \log(N)}{N} = o(1), \quad \text{and}$$

$$\frac{\zeta^2 (s_\mu + s_\pi)^4 \log^3(p \vee N) \log(N)}{N^{2-2/q} h^d} = o(1). \tag{18}$$

These conditions all hold if $\frac{s_\mu s_\pi \log^2(p\vee N) \log(N)}{Nh^d} = o(1)$ and $\zeta^2(s_\mu + s_\pi)^2 \log(p \vee N) \log(N) = o(N^{1-2/q})$. For example, if $q = 4$, $p = O(N^\lambda)$, $\lambda > 0$, and $\zeta = O(N^{1/4})$, then $\max(s_\mu, s_\pi) = o(\sqrt{Nh^d})$ is essentially sufficient for Assumption 3.2, ignoring logarithmic factors of N.

By contrast, Assumption 3.3 holds under substantially weaker sparsity conditions. Given the rates in (17), the l.h.s. of (12) is at most of order $O\big(\sqrt{s_\pi}\sqrt{s_\mu} \log(p)/N\big)$, as $\mathcal{K}_h$ is a bounded function. Hence, Assumption 3.3 essentially reduces to $\sqrt{s_\pi}\sqrt{s_\mu} \log(p)/(Nh^d) = o\big((\log(N)Nh^d)^{-1/2}\big)$. Again, setting $p = O(N^\lambda)$, $\lambda > 0$, and ignoring the logged factors of N gives $s_\pi s_\mu = o(Nh^d)$. This condition is of course satisfied if $s_\pi = s_\mu = o(\sqrt{Nh^d})$, but there can be tradeoffs between the two sparsity indexes. For example, if $s_\pi = O(1)$, that is, the propensity score essentially obeys a finite dimensional model linear in parameters, then $s_\mu = o(Nh^d)$ is possible, that is, $\mu_0(j, \cdot)$ can be a function that is substantially harder to approximate. Given Remark 5 after Assumption 3.3, we can see that our sparsity conditions for Assumption 3.3 are essentially equivalent to those in Chernozhukov and Semenova (2019). On the other hand, Lee, Okui, and Whang (2017) is based on parametric first-stage estimators with the dimension of the regressors fixed. Therefore, they do not need sparsity conditions (though one could regard the parametric assumption as an extreme form of sparsity).

Other types of lasso methods such as the group lasso by Farrell (2015) and the penalized local least squares and maximum likelihood methods by Su, Ura, and Zhang (2019) can also be used. One can verify the rate restrictions in a manner similar to the above.

4. Uniform Inference Based on the Multiplier Bootstrap

Turning to uniform inference, one option is to construct uniform confidence bands analytically similarly to LOW. We provide an alternative method based on the multiplier bootstrap. Our multiplier bootstrap procedure is computationally efficient and takes the nuisance function estimators from the first stage as given and only recomputes the nonparametric regression estimator(s) from the second stage. This step simply involves a random rescaling of the terms in the sums (2) and (3). As lasso estimation is usually time consuming, our procedure is less costly to implement than, say, a standard nonparametric bootstrap requiring new samples from the original data and recomputing the whole estimator.

To describe the procedure formally, we make the following assumption.

Assumption 4.1. The random variable ξ is independent of W with $\mathbb{E}(\xi) = \text{var}(\xi) = 1$, and its distribution has sub-exponential tails.[14]

Assumption 4.1 is standard for multiplier bootstrap inference. For example, a normal random variable with unit mean

[14] A random variable ξ has sub-exponential tails if $\mathbb{P}(|\xi| > x) \le K \exp(-Cx)$ for every x and some constants K and C.

and standard deviation satisfies this assumption. The bootstrap is implemented as follows:

1. Compute the first-stage nuisance function estimates $\hat{\mu}(0, x; I)$, $\hat{\mu}(1, x; I)$, $\hat{\pi}(x; I)$ OR $\hat{\mu}(0, x; I_k^c)$, $\hat{\mu}(1, x; I_k^c)$, $\hat{\pi}(x; I_k^c)$, $k = 1, \ldots, K$.
2. Draw an iid sequence $\{\xi_i\}_{i=1}^N$ from the distribution of ξ.
3. Choose the number of bootstrap replications B, for example, $B = 1000$. Compute $\hat{\tau}^b(x_1)$ by the local linear regression, for $b = 1, \ldots, B$,

$$(\hat{\tau}^b(x_1), \hat{\beta}^b(x_1)) = \arg\min_{a,b} \sum_{i \in I} \xi_i \big[\psi(W_i, \hat{\eta}(I)) - a - (X_{1i} - x_1)'b\big]^2 \mathcal{K}_h(X_{1i} - x_1),$$

or $\check{\tau}^b(x_1) = \frac{1}{K}\sum_{k=1}^K \hat{\tau}_k^b(x_1)$, where for $k = 1, \ldots, K$,

$$(\hat{\tau}_k^b(x_1), \hat{\beta}_k^b(x_1)) = \arg\min_{a,b} \sum_{i \in I_k} \xi_i \big[\psi(W_i, \hat{\eta}(I_k^c)) - a - (X_{1i} - x_1)'b\big]^2 \mathcal{K}_h(X_{1i} - x_1).$$

The following theorem is the bootstrap version of Theorem 3.1, and it forms the basis of our inference procedure.

Theorem 4.1.

(a) If Assumptions 2.1, 2.2, 3.1, 3.2, and 4.1 are satisfied, then

$$\hat{\tau}^b(x_1) - \hat{\tau}(x_1) = (\mathbb{P}_N - \mathbb{P})\Big[\frac{\xi - 1}{h^d f(x_1)}(\psi(W, \eta_0) - \tau_0(x_1))\mathcal{K}_h(X_1 - x_1)\Big] + R_\tau^b(x_1),$$

where $\sup_{x_1 \in \overline{\mathcal{X}}_1} |R_\tau^b(x_1)| = o_p((\log(N)Nh^d)^{-1/2})$.

(b) If Assumptions 2.1, 2.2, 3.1, 3.3, and 4.1 are satisfied, the representation established in part (a) also holds for $\check{\tau}^b(x_1) - \check{\tau}(x_1)$, that is,

$$\check{\tau}^b(x_1) - \check{\tau}(x_1) = (\mathbb{P}_N - \mathbb{P})\Big[\frac{\xi - 1}{h^d f(x_1)}(\psi(W, \eta_0) - \tau_0(x_1))\mathcal{K}_h(X_1 - x_1)\Big] + \check{R}_\tau^b(x_1),$$

where $\sup_{x_1 \in \overline{\mathcal{X}}_1} |\check{R}_\tau^b(x_1)| = o_p((\log(N)Nh^d)^{-1/2})$.

Theorem 4.1 justifies the validity of the multiplier bootstrap in implying that $\sqrt{Nh^d}(\hat{\tau}^b(x_1) - \hat{\tau}_0(x_1))$ converges in distribution to the limiting distribution of $\sqrt{Nh^d}(\hat{\tau}(x_1) - \tau_0(x_1))$ conditional on the sample path (data) with probability 1. Therefore, if Assumption 3.4 also holds, then, conditional on data,

$$\frac{\sqrt{Nh^d}\big(\hat{\tau}^b(x_1) - \hat{\tau}(x_1)\big)}{\widehat{\sigma}_N(x_1)} \xrightarrow{d} \mathcal{N}(0, 1). \qquad (19)$$

The same statements of course hold true if $\check{\tau}^b(x_1)$ and $\check{\tau}(x_1)$ replaces $\hat{\tau}^b(x_1)$ and $\hat{\tau}(x_1)$, respectively.[15] In addition to pointwise inference, the uniform control of the error term $R_N^b(\cdot)$ in Theorem 4.1 makes it possible to employ the multiplier bootstrap for uniform inference. For the rest of the article, we focus on the inference of $\tau_0(x_1)$. The uniform inference of $\beta_0(x_1)$ can be implemented in the same manner. We propose the following algorithm.

Uniform Confidence Band Implementation Procedure.

1. Compute $\hat{\tau}(x_1)$ and $\hat{\sigma}_N(x_1)$ for a suitably fine grid over $\overline{\mathcal{X}}_1$.
2. Compute $\hat{\tau}^b(x_1)$ over the same grid for $b = 1, \ldots, B$ while generating a new set of iid $\mathcal{N}(1, 1)$ random variables $\{\xi_i^b\}_{i=1}^N$ in each step b.
3. For $b = 1, \ldots, B$, compute

$$M_b^{\text{1-sided}} = \sup_{x_1 \in \overline{\mathcal{X}}_1} \frac{\sqrt{Nh^d}(\hat{\tau}^b(x_1) - \hat{\tau}(x_1))}{\widehat{\sigma}_N(x_1)},$$

$$M_b^{\text{2-sided}} = \sup_{x_1 \in \overline{\mathcal{X}}_1} \frac{\sqrt{Nh^d}\big|\hat{\tau}^b(x_1) - \hat{\tau}(x_1)\big|}{\widehat{\sigma}_N(x_1)},$$

where the supremum is approximated by the maximum over the chosen grid points.
4. Given a confidence level $1 - \alpha$, find the empirical $(1 - \alpha)$ quantile of the sets of numbers $\{M_b^{\text{1-sided}} : b = 1, \ldots, B\}$ and $\{M_b^{\text{2-sided}} : b = 1, \ldots, B\}$. Denote these quantiles as $\widehat{C}_\alpha^{\text{1-sided}}$ and $\widehat{C}_\alpha^{\text{2-sided}}$, respectively.
5. The uniform confidence bands are constructed as

$$I_L = \Big\{\Big(\hat{\tau}(x_1) - \widehat{C}_\alpha^{\text{1-sided}}\frac{\widehat{\sigma}_N(x_1)}{\sqrt{Nh^d}}, \infty\Big) : x_1 \in \overline{\mathcal{X}}_1\Big\},$$

$$I_R = \Big\{\Big(-\infty, \hat{\tau}(x_1) + \widehat{C}_\alpha^{\text{1-sided}}\frac{\widehat{\sigma}_N(x_1)}{\sqrt{Nh^d}}\Big) : x_1 \in \overline{\mathcal{X}}_1\Big\},$$

$$I_2 = \Big\{\Big(\hat{\tau}(x_1) - \widehat{C}_\alpha^{\text{2-sided}}\frac{\widehat{\sigma}_N(x_1)}{\sqrt{Nh^d}},\ \hat{\tau}(x_1) + \widehat{C}_\alpha^{\text{2-sided}}\frac{\widehat{\sigma}_N(x_1)}{\sqrt{Nh^d}}\Big) : x_1 \in \overline{\mathcal{X}}_1\Big\}.$$

The following theorem formally states the asymptotic validity of the confidence regions proposed above.

Theorem 4.2. If Assumptions 2.1, 2.2, 3.1, 3.2, 3.4, and 4.1 are satisfied, then

$$\lim_{N \to \infty} \mathbb{P}\big(\tau_0 \in I_L\big) = \lim_{N \to \infty} \mathbb{P}\big(\tau_0 \in I_R\big) = \lim_{N \to \infty} \mathbb{P}\big(\tau_0 \in I_2\big) = 1 - \alpha.$$

Remarks.

1. Theorem 4.2 states that the random confidence bands I_R, I_L, and I_2 contain the *entire* function τ_0 with the prescribed probability $1 - \alpha$ in large samples.
2. If the grid in step 1 is chosen to be a single point x_1, then the algorithm provides pointwise confidence intervals $I_L(x_1)$, $I_R(x_1)$, and $I_2(x_2)$.
3. One can construct uniform confidence bands for τ_0 based on the cross-fitting estimator $\check{\tau}$ following the exact same steps as above; of course, one needs to replace $\hat{\tau}$, $\hat{\tau}^b$, and $\hat{\sigma}_N$ with $\check{\tau}$, $\check{\tau}^b$, and $\check{\sigma}_N$, respectively.
4. It is also possible to construct the uniform confidence band by approximating the supremum of the empirical process via a Gumbel distribution. However, the Gumbel approximation is accurate only up to the logarithmic rate, as pointed out by Lee, Okui, and Whang (2017). The bootstrap approximation proposed in this article has the advantage that the approximation error has a geometric rate of decline and the quality of the approximation is better than that of Gumbel.[16] We also

[15] In the online supplement, we also show similar results regarding $\hat{\beta}^b(x_1)$ and $\check{\beta}^b(x_1)$ that might be of separate interest.

[16] We thank an anonymous referee for this excellent comment.

note that both the bootstrap and the Gumbel approximations rely on the linear expansions established in Theorem 3.1.

We discuss the bandwidth choice in practice. To obtain our theoretical results, we require undersmoothing to eliminate bias asymptotically. When $d = 1$ as in the simulations, we suggest setting $h_N = \hat{h} \times N^{1/5} \times N^{-2/7}$, where $\hat{h} = 1.06 \cdot \hat{\sigma}_{x_1} N^{-1/5}$ and $\hat{\sigma}_{x_1}$ is the estimated standard deviation of X_1. The formula for $\hat{h}$ corresponds to the rule-of-thumb bandwidth with a Gaussian kernel suggested by Silverman (1986).[17] The bandwidth selection is done with the entire sample even for the cross-fitting method. Also the same selection method is employed in the empirical application.

We investigate the finite sample properties of the proposed high-dimensional CATE estimators and the inference procedure outlined above using Monte Carlo experiments. Because of the space constraint, this material can be found in the online supplement.

5. Empirical Application

In this section, we employ the proposed high-dimensional CATE estimators to analyze the average effect of maternal smoking on birth weight while allowing for virtually unrestricted treatment effect heterogeneity conditional on the mother's age. Birth weight has been associated with health and human capital development throughout life (Black, Devereux, and Salvanes 2007, Almond and Currie 2011), and maternal smoking is considered to be the most important preventable cause of low birth weight (Kramer 1987). In recent studies, AHL and LOW both explored this causal relationship using the CATE approach, and found different degrees of heterogeneity by age. Using observations from 3754 white mothers in Pennsylvania, LOW found that the CATE of smoking is decreasing from 17 to around 29 years of age, but they differ from AHL in that the contrast between young and 30-year-old mothers is still not large.[18]

Our study improves on these previous investigations by considering a much larger pool of covariates and explicitly incorporating a variable selection mechanism into the estimation. This initial pool consists of a vector X of raw covariates as well as technical regressors (powers and interactions) to account for the fact that the functional form of π_0 and μ_0 is unknown. By contrast, AHL assume that a low dimensional parametric model (known up to its coefficients) is correctly specified for π_0, while LOW assume that either π_0 or μ_0 obeys such a model. While we still assume that π_0 and μ_0 are sparse functions, we let a data-driven procedure (lasso) select the most relevant regressors.

5.1. Data Description

We start with the same dataset as AHL, composed of vital statistics collected by the North Carolina State Center Health Services, and extract the records of first-time mothers[19] between 1988 and 2002. The variables include whether the mother smokes (the treatment dummy), the baby's birth weight (the main outcome variable, measured in grams), the parents' socioeconomic information, such as age, education, income, race, etc., as well as the mothers' medical and health records. The dataset includes 45 raw covariates and 591,547 observations in total. Table 1 summarizes the most important pretreatment covariates in the dataset.[20]

5.2. High-Dimensional CATE Estimation

In this section, we estimate the CATE of maternal smoking on the baby's birth weight with mother's age as the conditioning variable. Following AHL and LOW, we estimate CATE separately for black mothers and white mothers. We only report the estimation results for white mothers in this section; the results for the black mothers can be found in the online supplement. The dependent variable Y is the baby's birth weight measured in grams. The treatment dummy D takes on the value 1 if the mother smokes and 0 otherwise. We start from the set of variables displayed in Table 1, and construct an even larger dictionary $b(X)$ by adding polynomial terms to account for the unknown form of the nuisance functions in a flexible way. Specifically, we include, up to degree 3, the powers and interaction terms of key dummy variables and continuous and integer covariates. We then end up with 792 covariates in total.

With such a large set of covariates, it is not clear which variables are important in estimating the CATE function. The true set of variables which belong to the estimating equations is assumed to be sparse, as discussed in the previous sections. We hence apply the lasso method in Belloni et al. (2017) to estimate propensity score (π_0) and conditional mean function (μ_0). We then compute the robust score function ψ for each observation i, and run a local linear regression of ψ_i on mother's age evaluated at numerous grid points in the interval [15, 36] (years of age). We use the cross-fitting variant of the estimator, that is, the nuisance function estimation and the kernel regression take place in different subsamples, and then these roles are rotated. In the empirical study, we use the same K ($= 4$) as in the simulations. Granted that the theoretical property of the proposed K-fold cross-fitting estimator is the same as the full-sample estimator in large samples, we recommend using sample-splitting estimator with $K = 4$ or 5 following the suggestion of Chernozhukov et al. (2018). We refer to the resulting point estimates as HDCATE (HD stands for "high dimensional").

AHL include the mother's age, education, month of first prenatal visit (=10 if prenatal care is foregone), number of prenatal visits, and indicators for the baby's gender, the mother's marital

[17] When $d = 2$ or 3, we suggest setting for $j = 1, \dots, d$, $h_{jN} = \hat{h}_j \times N^{1/(4+d)} \times N^{-2/(4+3d)}$ and $\hat{h}_j = 1.06 \cdot \hat{\sigma}_{x_{1j}} N^{-1/(4+d)}$ and $\hat{\sigma}_{x_{1j}}$ is the estimator of the standard deviation of the jth element of X_1.

[18] As the smoking effect is negative, "decreasing" means that the detrimental effects of smoking become stronger with age.

[19] The motivation for focusing on first-time mothers is discussed in AHL. In effect, the restricted sample enables more credible identification of the causal effect, as there cannot be uncaptured feedback from the previous birth experience to the current one.

[20] We drop some covariates from the analysis for various reasons. For example, the mother's weight gain during pregnancy is arguably not a pretreatment variable, and the Kessner index of prenatal care is basically a function of the number of prenatal visits and the timing of the first visit.

Table 1. Variable definitions.

		Name	Type	Description
Outcome variable		bweight	Real number	Birth weight(g)
Treatment		smoke	Dummy	Whether mother smokes or not
Covariates	Parents Basic Info	mage	Real number[a]	Mother's age
		meduc	Integer	Mother's years of schooling
		fage	Integer	Father's age
		feduc	Integer	Father's years of schooling
		fagemiss	Integer	Whether or not father's age is missing
		married	Dummy	Whether or not mother is married
		popdens	Real number	Population density in mother's zip code (units/km^2)
	Mothers' Medical Care & Health Status	prenatal	Integer	Month of first prenatal visit (=10 if prenatal care is foregone)
		pren_visits	Integer	Number of prenatal visits
		terms	Integer	Previous (terminated) pregnancies
		amnio	Dummy	Did mother take amniocentesis?
		anemia	Dummy	Did mother suffer from anemia?
		diabetes	Dummy	Did mother suffer from gestational diabetes?
		hyperpr	Dummy	Did mother suffer from hypertension?
		ultra	Dummy	Did mother take ultrasound exams?
	Others	male	Dummy	Whether or not baby is male
		drink	Dummy	Mother's alcohol use
		by88-02	Dummy	13 birth year dummies (from 1988 to 2002)

[a] NOTE: mother's age is originally recorded as an integer but for the purposes of this exercise we add a uniform $[-1, 1]$ random number to this value to make it a continuous variable. The main results are robust when we add a uniform $[-0.5, 0.5]$ random number to the age variable. See more empirical results in the online supplement.

status, whether or not the father's age is missing, gestational diabetes, hypertension, amniocentesis, ultrasound exams, previous (terminated) pregnancies, and alcohol use as the confounding factors. The variables selected by our first-step estimation are similar to those used in AHL, with some notable differences.[21] In the propensity score function, we also select father's age, and father's education, besides the ones used in AHL, but not gestational diabetes and amniocentesis. In the conditional mean function, we have father's education, and the rest overlap with that of AHL.

The HDCATE estimates are displayed in Figures 1–3, along with 90%, 95%, and 99% confidence bands, respectively. For a given confidence level, we compute two types of intervals. "HDCATE CB" is the proposed uniform confidence band computed according to the algorithm given in Section 4. "PW CB" is a pointwise confidence band, given for purposes of comparison, where the critical value $\widehat{C}_{\alpha}^{\text{2-sided}}$ is replaced by the corresponding value from the standard normal distribution (e.g., 1.96 for $\alpha = 5\%$), and "LOW CB" is the uniform confidence band by that of LOW. The constant function labeled "ATE" represents the estimated average treatment effect across all ages.

Figures 1–3 show that maternal smoking has a negative effect on birth weight at all ages (the upper bounds of the confidence bands are negative), and the average effect is likely to become *more negative* with age. For example, the point estimates show that for teenage mothers of age 18 or younger the negative effect of smoking is, on average, less than 180 g in absolute value. For mothers around age 24, the same effect is -220 g, and it approaches -250 above 35 years of age.[22] Thus, there is substantial variation in the estimated average treatment effect by age. A potential explanation is that older mothers are likely to have smoked for a longer period, and the detrimental effects of smoking are cumulative (the smoking dummy does not control

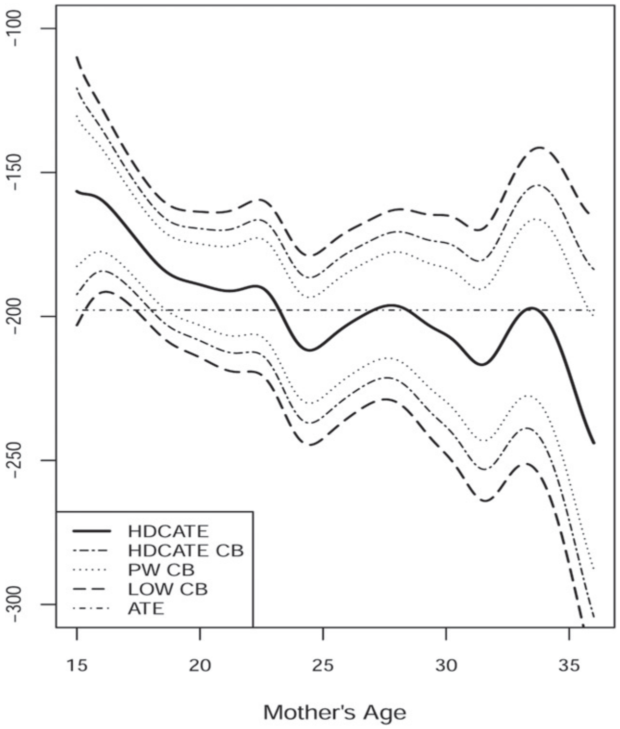

Figure 1. CATE for the effect of smoking on birth weights conditional on mother's age, 99% confidence bands.

[21] Given that we use the cross-fitting method, there are $K = 4$ first-stage estimates, and each has its own variable selection. The reported set of variables selected in the first stage is the union of the selected variables in the four split-sample first stages.

[22] The nonmonotonicities in the point estimate between ages 25 and 35 could be due to undersmoothing and the quickly declining number of first-time mothers toward the top of this age range.

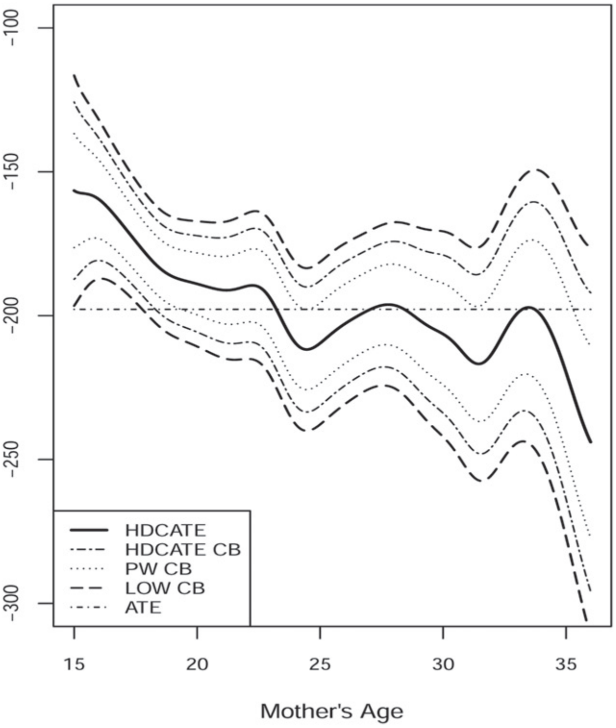

Figure 2. CATE for the effect of smoking on birth weights conditional on mother's age, 95% confidence bands.

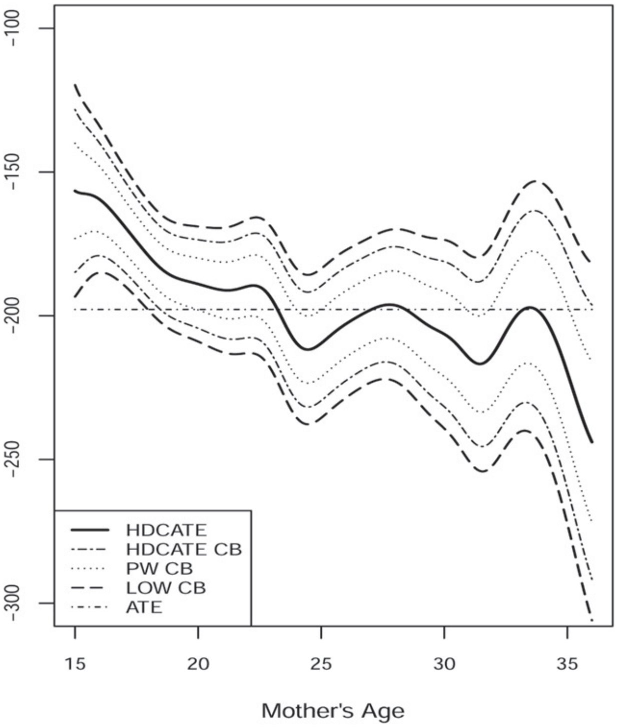

Figure 3. CATE for the effect of smoking on birth weights conditional on mother's age, 90% confidence bands.

for duration or intensity of smoking). The figures also shed some light on the gains from using the proposed method compared with the original study of AHL. The CATE function changes the shape and location of the treatment effect estimates, especially for the younger mothers, and shows that the estimated treatment effects are always significantly negative using our method.[23] Another important difference is that in this study we provide a valid uniform confidence band.

Examining the confidence bands qualifies the analysis of the point estimate in important ways. In Figure 1, the lower bound of the 99% uniform confidence band (dashed line) attains its maximum at around 16 years of age, and the value of this maximum lies just below the minimum of the upper bound attained at around age 24. Thus, it is possible to fit a constant function (at about −185 g) inside the uniform confidence bands. Nevertheless, if one is less conservative and uses the 95% or 90% uniform confidence bands displayed in Figures 2 and 3, respectively, then it is no longer possible to do so. Thus, there is fairly compelling (statistically significant) evidence that the smoking effect becomes more negative at least between the ages of 16 and 24. Based on the pointwise confidence band, there is some evidence of further decline in HDCATE at higher ages but it is possible to fit constant functions even within the 90% uniform confidence bands over the interval $[25, 35]$. (Again note that these bands become rather wide at higher ages due to the relatively small number of observations.) The LOW confidence band is visibly wider than our confidence band, which is consistent with our simulation results.

6. Conclusion

We advance the literature on the estimation of the reduced dimensional CATE function by proposing that the nuisance functions necessary for identification be estimated by flexible machine learning methods, followed by a traditional local linear regression. The asymptotic theory we develop builds on previous work by Belloni et al. (2017) and Chernozhukov et al. (2018). Nevertheless, the theory requires nontrivial modifications to accommodate local linear regression in the second stage. Moreover, CATE is a functional parameter, and our results can be used to conduct uniform inference through a bootstrap procedure. In line with Chernozhukov et al. (2018), we also advocate using the cross-fitting approach to estimate the nuisance functions and conduct the second-stage regression.

Using the proposed methods, we revisited the problem of estimating the average effect of smoking during pregnancy on birth weight as a function of the mother's age. Our results fall in between AHL and LOW in the sense that we do find age-related heterogeneity (unlike LOW), but it is less marked than in the former study. In particular, there is evidence that the negative effect of smoking becomes somewhat more pronounced with age.

Supplementary Materials

The supplementary material contains additional empirical studies, the Monte Carlo simulations as well as detailed technical proofs.

[23] Granted, the choice of bandwidth is different for the two studies, which affects the shape of the estimated heterogeneous treatment curve to some extent, but it is not the key reason for the different results. If we were to use the same bandwidth choice as in AHL, we would still observe the difference, as we mentioned in the main text.

Acknowledgments

The authors thank the anonymous referees, the associate editor, and Christian Hansen (the editor) for many constructive comments. They also thank Songnian Chen, Edward Kennedy, Whitney Newey, Ryo Okui, Yoon-Jae Whang, and participants of California Econometrics Conference (Irvine), Cemmap-WISE Joint Workshop, AMES 2019, for their valuable comments.

Funding

Fan acknowledges financial support from National Natural Science Foundation of China grants 71671149, 71801183, and 71631004 (Key Project). Hsu acknowledges research support from the Ministry of Science and Technology of Taiwan (107-2410-H-001-034-MY3) and the Career Development Award of Academia Sinica, Taiwan. Zhang acknowledges financial support from the Singapore Ministry of Education Tier 2 grant under grant MOE2018-T2-2-169 and the Lee Kong Chian Fellowship.

References

Abrevaya, J., Hsu, Y.-C., and Lieli, R. P. (2015), "Estimating Conditional Average Treatment Effects," *Journal of Business & Economic Statistics*, 33, 485–505. [313]

Ai, C., and Chen, X. (2003), "Efficient Estimation of Models With Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795–1843. [318]

Almond, D., and Currie, J. (2011), "Killing Me Softly: The Fetal Origins Hypothesis," *Journal of Economic Perspectives*, 25, 153–172. [323]

Begun, J. M., Hall, W., Huang, W.-M., and Wellner, J. A. (1983), "Information and Asymptotic Efficiency in Parametric-Nonparametric Models," *The Annals of Statistics*, 11, 432–452. [315]

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017), "Program Evaluation With High-Dimensional Data," *Econometrica*, 85, 233–298. [314,315,317,318,320,321,323,325]

Belloni, A., Chernozhukov, V., and Hansen, C. (2014a), "High-Dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspectives*, 28, 29–50. [314]

——— (2014b), "Inference on Treatment Effects After Selection Among High-Dimensional Controls," *The Review of Economic Studies*, 81, 608–650. [314,320]

Bickel, P., Klaassen, C., Ritov, Y., and Wellner, J. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, New York: Springer-Verlag. [315]

Black, S., Devereux, P., and Salvanes, K. (2007), "From the Cradle to the Labor Market? The Effect of Birth Weight on Adult Outcomes," *The Quarterly Journal of Economics*, 122, 409–439. [323]

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2018), "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference," *Journal of the American Statistical Association*, 113, 767–779. [318]

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017), "Double/Debiased/Neyman Machine Learning of Treatment Effects," *American Economic Review Papers and Proceedings*, 107, 261–265. [314,317]

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018), "Double/Debiased Machine Learning for Treatment and Structural Parameters," *The Econometrics Journal*, 21, C1–C68. [314,319,323,325]

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014), "Gaussian Approximation of Suprema of Empirical Processes," *The Annals of Statistics*, 42, 1564–1597. [315,317]

Chernozhukov, V., and Semenova, V. (2019), "Simultaneous Inference for Best Linear Predictor of the Conditional Average Treatment Effect and Other Structural Functions," Working Paper, Department of Economics, MIT. [314,315,319,321]

Fan, J. (1992), "Design-Adaptive Nonparametric Regression," *Journal of the American Statistical Association*, 87, 998–1004. [317]

——— (1993), "Local Linear Regression Smoothers and Their Minimax Efficiencies," *The Annals of Statistics*, 21, 196–216. [317]

Fan, J., and Gijbels, I. (1992), "Variable Bandwidth and Local Linear Regression Smoothers," *The Annals of Statistics*, 20, 2008–2036. [317]

Farrell, M. H. (2015), "Robust Inference on Average Treatment Effects With Possibly More Covariates Than Observations," *Journal of Econometrics*, 189, 1–23. [315,321]

Farrell, M. H., Liang, T., and Misra, S. (2018), "Deep Neural Networks for Estimation and Inference," arXiv no. 1809.09953. [319]

Firpo, S. (2007), "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 75, 259–276. [315]

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331. [313,315]

Hirano, K., Imbens, G. W., and Ridder, G. (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189. [313,315]

Imbens, G. W., and Wooldridge, J. M. (2009), "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86. [313]

Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. (2017), "Non-Parametric Methods for Doubly Robust Estimation of Continuous Treatment Effects," *Journal of the Royal Statistical Society*, Series B, 79, 1229–1245. [315,318]

Kramer, M. S. (1987), "Intrauterine Growth and Gestational Duration Determinants," *Pediatrics*, 80, 502–511. [323]

Lechner, M. (2019), "Modified Causal Forests for Estimating Heterogeneous Causal Effects," arXiv no. 1812.09487. [314]

Lee, S., Okui, R., and Whang, Y.-J. (2017), "Doubly Robust Uniform Confidence Band for the Conditional Average Treatment Effect Function," *Journal of Applied Econometrics*, 32, 1207–1225. [314,321,322]

Lee, Y.-Y. (2018), "Partial Mean Processes With Generated Regressors: Continuous Treatment Effects and Nonseparable Models," arXiv no. 1811.00157. [317]

Li, Q., and Racine, J. S. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton, NJ: Princeton University Press. [318]

Luedtke, A. R., and van der Laan, M. J. (2016a), "Statistical Inference for the Mean Outcome Under a Possibly Non-Unique Optimal Treatment Strategy," *The Annals of Statistics*, 44, 713. [314]

——— (2016b), "Super-Learning of an Optimal Dynamic Treatment Rule," *The International Journal of Biostatistics*, 12, 305–332. [314]

Newey, W. K. (1994a), "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382. [315]

——— (1994b), "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, 10, 1–21. [317]

Nie, X., and Wager, S. (2017), "Quasi-Oracle Estimation of Heterogeneous Treatment Effects," arXiv no. 1712.04912. [314]

Pfanzagl, J. (1990), *Estimation in Semiparametric Models*, New York: Springer. [315]

Robins, J. M. (2004), "Optimal Structural Nested Models for Optimal Sequential Decisions," in *Proceedings of the Second Seattle Symposium in Biostatistics*, Springer, pp. 189–326. [314]

Robins, J. M., Li, L., Mukherjee, R., Tchetgen, E. T., and van der Vaart, A. (2017), "Minimax Estimation of a Functional on a Structured High-Dimensional Model," *The Annals of Statistics*, 45, 1951–1987. [315]

Robins, J. M., and Rotnitzky, A. (1995), "Semiparametric Efficiency in Multivariate Regression Models With Missing Data," *Journal of the American Statistical Association*, 90, 122–129. [315]

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [313]

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall. [323]

Su, L., Ura, T., and Zhang, Y. (2019), "Non-Separable Models With High-Dimensional Data," *Journal of Econometrics*, 212, 646–677. [315,318,321]

Tsiatis, A. (2007), *Semiparametric Theory and Missing Data*, New York: Springer. [315]

van der Laan, M. J. (2013), "Targeted Learning of an Optimal Dynamic Treatment, and Statistical Inference for Its Mean Outcome," U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper 317. [314]

van der Laan, M. J., and Robins, J. M. (2003), *Unified Methods for Censored Longitudinal Data and Causality*, New York: Springer. [315]

van der Laan, M. J., and Rose, S. (2011), *Targeted Learning: Causal Inference for Observational and Experimental Data*, New York: Springer. [315]

van der Laan, M. J., and Rubin, D. (2006), "Targeted Maximum Likelihood Learning," *The International Journal of Biostatistics*, 2, 11. [315]

van der Vaart, A. W. (2000), *Asymptotic Statistics* (Vol. 3), Cambridge: Cambridge University Press. [315]

Wager, S., and Athey, S. (2018), "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242. [315,319]

Zimmert, M., and Lechner, M. (2019), "Nonparametric Estimation of Causal Heterogeneity Under High-Dimensional Confounding," arXiv no. 1908.08779. [314]