Statistica Sinica

A Simple and Efficient Estimation of Average Treatment Effects in Models with Unmeasured Confounders

Chunrong Ai

Chinese University of Hong Kong, Shenzhen, China

Lukang Huang

Renmin University of China, Beijing, China

Zheng Zhang

Renmin University of China, Beijing, China

Abstract: This paper presents a simple and efficient estimation of the average treatment effect (ATE) and local average treatment effect (LATE) in models with unmeasured confounders. In contrast to existing studies that estimate some unknown functionals in the influence function either parametrically or semiparametrically, we do not model the influence function nonparametrically. Instead, we apply the *calibration* method to a growing number of moment restrictions to estimate the weighting functions nonparametrically and then estimate the ATE and LATE by plugging in. The *calibration* method is similar to the covariate-balancing method in that both methods exploit the moment restrictions. The difference is that the *calibration* method imposes the sample analogue of the moment restrictions, while the covariate-balancing method does not. A simulation study reveals that our estimators have good finite sample performance and out-perform existing alternatives. An application to the empirical analysis of return to education illustrates the practical value of the proposed method.

Key words and phrases: Average treatment effect; Endogeneity; Local average treatment effect; Semiparametric efficiency; Unmeasured confounders

1. Introduction

A common approach to account for individual heterogeneity in the treatment evaluation literature on observational data is to assume that there exist confounders and that conditional on those confounders, there is no systematic selection into the treatment. This assumption is called the *Unconfounded Treatment Assignment* condition (e.g., Rosenbaum and Rubin (1983)). Under this condition, the average treatment effect (ATE) is identified and many estimation methods have been proposed including weighting methods (Hirano, Imbens, and Ridder (2003), Huang and Chan (2017)), imputation methods (Rosenbaum (2002)), regression methods (Chen, Hong, and Tarozzi (2008)), matching methods (Abadie and Imbens, 2006), and covariate-balancing methods (Imai and Ratkovic, 2014).

A key requirement in this literature is that all confounders are measured. If some confounders are unmeasured and left out of the conditioning arguments, the *Unconfounded Treatment Assignment* condition may not hold and consequently the ATE may not be identified. Indeed, the ATE may still be unidentified even if the standard instrumental variable condition is satisfied (e.g., Imbens and Angrist (1994)). Thus, to identify the ATE (or some version of the ATE), additional restrictions must be imposed either on the model specification or on the instrument (or both). For example, Wang and Tchetgen Tchetgen (2018) imposed restrictions on the model specification. They showed that the ATE is identified if (i) there is no additive interaction among the instrument-unmeasured confounders in the treatment probability, conditional on all the confounders and the instrument, or (ii) there is no additive interaction among the treatment status-unmeasured confounders in the expectation of the potential outcomes conditional on all the confounders. They derived the influence function, which contains five unknown functionals. They parameterized all five functionals, estimated those functionals with appropriate parametric approaches, and estimated the ATE by plugging in the estimated functionals. They showed that their ATE estimator is consistent if certain functionals are correctly parameterized and attains the semiparametric efficiency bound if all the functionals are correctly specified.

Alternatively, Imbens and Angrist (1994) imposed restrictions on the instrumental variable. They showed that if the treatment variable is monotone in the instrumental variable for the complier subpopulation, the ATE for the complier subpopulation (i.e., the local average treatment effect (LATE)) is identified. Frölich (2007) extended this local identification result to include confounders. He computed the efficiency bound of the LATE and derived the influence function, which contains four functionals. He then estimated all four functionals nonparametrically and estimated the LATE by plugging in the estimated functionals. He showed that his LATE estimator attains the semiparametric efficiency bound. Donald, Hsu, and Lieli (2014b,a) proposed an inverse probability weighting method for the LATE. However, Kang and Schafer (2007) argued that the inverse probability weighting method is likely to produce extreme weights and unstable estimates.

In this paper, we propose a simple and efficient estimation of both the ATE and the LATE by extending the calibration method developed by Chan, Yam, and Zhang (2016). The calibration method estimates the weighting functions by solving the sample analogue of the moment restrictions. We then estimate the ATE and LATE by plugging in the estimated weighting functions. We show that our estimators of the ATE and LATE are efficient, attaining their respective semiparametric efficiency bound. We note that the covariate-balancing method developed by Imai and Ratkovic (2014) for the weighting functions also exploits the same moment restrictions, although it does not necessarily solve the sample analogue. Imai and Ratkovic (2014) argued that the imposition of those moment restrictions produces covariate-balancing weights that improve the performance of their matching and weighting estimators. Since our method and the covariate-balancing method share the same idea, our method is expected to produce a stable plug-in estimator with good finite sample performance. Furthermore, we argue that the imposition of the sample analogue of the moment restrictions delivers the efficiency of the plug-in estimator.

In contrast to existing methods that estimate some unknown functionals in the influence function either parametrically or semiparametrically, our method has the advantage of not modeling the influence function, thereby avoiding potential model misspecification bias. Even if all methods have the same asymptotic properties, our method may have the advantage of being more stable and showing better finite sample performance because our weights are estimated directly instead of computed as the inverse of the estimated probabilities.

The remainder of the paper is organized as follows. Section 2 describes the basic framework. Section 3 presents the estimation of the ATE and Section 4 derives its large sample properties. Section 5 presents a consistent variance. Section 6 describes the estimation of the LATE and derives its large sample properties. Since the proposed procedure depends on the smoothing parameter, Section 7 presents a data-driven method for determining the smoothing parameter. Section 8 evaluates the finite sample performance of the proposed estimators using a small simulation study, while Section 9 illustrates the practical value of our method by revisiting the return to education study. Section 10 contains a brief discussion. All the technical proofs are relegated to the supplementary material.

2. Basic Framework

2.1 The ATE

Let $D \in \{0,1\}$ denote the binary treatment variable and let $Y(1)$ and $Y(0)$ denote the potential outcomes when an individual is assigned to the treatment and control groups, respectively. The parameter of interest is the ATE $\tau = \mathbb{E}[Y(1) - Y(0)]$. The estimation of τ is complicated by the confounders and the fact that $Y(1)$ and $Y(0)$ are not observed simultaneously. We use X to denote the measured confounders and U to denote the unmeasured confounders. When all the confounders are measured, the following condition identifies τ:

Assumption 1. $(Y(0), Y(1)) \perp D | (X, U)$.

When U is unmeasured and omitted from the conditioning argument, $(Y(0), Y(1)) \perp D|X$ may not hold, resulting in the classical omitted variable problem. To tackle this omitted variable problem, the instrumental variable method is often preferred. Let $Z \in \{0,1\}$ denote the variable satisfying the classical instrumental variable condition:

Assumption 2. *(i) $\forall z, d$, $Y(z,d) = Y(d)$, where $Y(z,d)$ is the response that would be observed if a unit were exposed to d and the instrument had taken value z; (ii) $Z \perp U|X$; and (iii) $Z \not\perp D|X$.*

Wang and Tchetgen Tchetgen (2018) showed that Assumptions 1 and 2 alone do not identify τ. However, if one of the following conditions also holds:

- *Condition (a)*: $\mathbb{E}[D|Z=1,X,U] - \mathbb{E}[D|Z=0,X,U] = \delta^D(X) := \mathbb{E}[D|Z=1,X] - \mathbb{E}[D|Z=0,X]$ and

- *Condition (b)*: $\mathbb{E}[Y(1)-Y(0)|X,U] = \delta^Y(X) := \mathbb{E}[Y(1)-Y(0)|X]$,

then τ is identified as $\tau = \mathbb{E}[\delta(X)]$ with $\delta(X) = \delta^Y(X)/\delta^D(X)$. Wang and Tchetgen Tchetgen (2018) derived the influence function as

$$\begin{aligned}\varphi_{eff}(D,Z,X,Y) =& \frac{2Z-1}{f(Z|X)}\frac{1}{\delta^D(X)}\Big\{Y - D\delta(X) - \mathbb{E}[Y|Z=0,X] \\ &+ \mathbb{E}[D|Z=0,X]\delta(X)\Big\} + \delta(X) - \tau, \end{aligned} \tag{2.1}$$

where $f(Z|X)$ is the conditional probability mass function of Z given X. The semiparametric efficiency bound of τ is $V_{eff} = \mathbb{E}[\varphi_{eff}(D, Z, X, Y)^2]$. Since the influence function depends on five unknown functionals, namely, $\delta(X)$, $\delta^D(X)$, $f(Z|X)$, $\mathbb{E}[Y|Z = 0, X]$, and $\mathbb{E}[D|Z = 0, X]$, they proposed parameterizing all five functionals, estimated the functionals with appropriate parametric methods, and estimated τ by plugging in the estimated functionals. They showed that their estimator of τ is consistent and asymptotically normally distributed if

- $\delta(X)$, $\delta^D(X)$, $\mathbb{E}[Y|Z = 0, X]$, and $\mathbb{E}[D|Z = 0, X]$ are correctly specified,
- $\delta^D(X)$ and $f(Z|X)$ are correctly specified, or
- $\delta(X)$ and $f(Z|X)$ are correctly specified,

and their estimator only attains the semiparametric efficiency bound when all five functionals are correctly specified. However, their estimator is inefficient if any functional is misspecified and inconsistent if $\delta^D(X)$ and $f(Z|X)$ are incorrectly specified.

2.2 The LATE

Wang and Tchetgen Tchetgen (2018) achieved the global identification of the ATE by restricting the model specification. Imbens and Angrist (1994) and Frölich (2007) achieved the local identification of the ATE by restricting the instrumental variable. Specifically, let $D(z) \in \{0,1\}$ denote a binary potential treatment indicator when the instrument takes the value $Z = z$. The observed treatment variable is $D = ZD(1)+(1-Z)D(0)$. Suppose that the treatment status is monotone in instrument (i.e., $P(D(1) < D(0)|X) = 0$). The LATE is identified as

$$\tau_{LATE} = \mathbb{E}[Y(1) - Y(0)|D(1) > D(0)] = \frac{\mathbb{E}[\delta^Y(X)]}{\mathbb{E}[\delta^D(X)]}.$$

Frölich (2007) derived the influence function of τ_{LATE} (see Section 7 of the supplemental material). He proposed the plug-in estimator of τ_{LATE} by estimating the functionals $\mathbb{E}[Y|X, Z = 0]$, $\mathbb{E}[Y|X, Z = 1]$, $\mathbb{E}[D|X, Z = 0]$, and $\mathbb{E}[D|X, Z = 1]$ nonparametrically and showed that his estimator attains the semiparametric efficiency bound. Based on the alternative expression:

$$\tau_{LATE} = \frac{\mathbb{E}[(2Z-1)Y/f(Z|X)]}{\mathbb{E}[(2Z-1)D/f(Z|X)]},$$

Donald, Hsu, and Lieli (2014b) and Donald, Hsu, and Lieli (2014a) proposed estimating $f(Z|X)$ nonparametrically and estimated τ_{LATE} by plugging in. Kang and Schafer (2007) argued that this inverse probability weighting method is sensitive to small values of the estimated $f(Z|X)$.

3. Estimation of the ATE

3.1 Motivation

To motivate our estimation, we rewrite τ as

$$\tau = \mathbb{E}\left[\left\{\frac{2Z-1}{f(Z|X)}\right\}\frac{Y}{\delta^D(X)}\right]. \tag{3.2}$$

In the case of $D = Z$, τ is simplified to $\tau = \mathbb{E}\left[(2D-1)Y/f(D|X)\right]$, which can be estimated using the inverse propensity score method (Hirano, Imbens, and Ridder (2003), Imai and Ratkovic (2014)). In other cases, we have to invert two functions: $f(Z|X)$ and $\delta^D(X)$. In principle, we can replace these functions with some consistent estimates and estimate τ by plugging in. Although there are many methods to estimate these functions, not all the plug-in estimators of τ are efficient (see Hirano, Imbens, and Ridder (2003)). For example, if $f(Z|X)$ and $\delta^D(X)$ are known, the sample average of (3.2) has the asymptotic variance $V_{ineff} = \mathbb{E}\left[\left\{\left(\frac{2Z-1}{f(Z|X)}\right)\frac{Y}{\delta^D(X)} - \tau\right\}^2\right]$, which is greater than the semiparametric efficiency bound V_{eff}. Hahn (1998) and Hirano et al. (2003) established the same result for their models.

Note that $f(Z|X)$ and $\delta^D(X)$ satisfy the following restrictions:

$$\mathbb{E}\left[Zf^{-1}(Z|X)u(X)\right] = \mathbb{E}[u(X)] \;=\; \mathbb{E}\left[(1-Z)f^{-1}(Z|X)u(X)\right] \tag{3.3}$$

$$\text{and } \mathbb{E}\left[\delta^D(X)u(X)\right] = \mathbb{E}\left[D\left\{(2Z-1)f^{-1}(Z|X)u(X)\right\}\right] \tag{3.4}$$

hold for any integrable function $u(X)$. With these restrictions, Wang and Tchetgen Tchetgen (2018) proposed parametric estimators for $f(Z|X)$ and $\delta^D(X)$. Motivated by this insight, we propose an alternative estimation of τ based on the complete model. We first verify that (3.3) and (3.4) identify $f^{-1}(Z|X)$ and $\delta^D(X)$. Let $w(Z|X)$ denote an arbitrary function of (Z, X) and $d(X)$ denote an arbitrary function of X. The following theorem is proven in the supplemental material.

Theorem 1. *For any integrable function $u(X)$,*

$$\mathbb{E}\left[Z \cdot w(Z|X)u(X)\right] = \mathbb{E}[u(X)] = \mathbb{E}\left[(1-Z)w(Z|X)u(X)\right] \tag{3.5}$$

$$\textit{and } \mathbb{E}\left[d(X)u(X)\right] = \mathbb{E}\left[D\left\{(2Z-1)w(Z|X)u(X)\right\}\right] \tag{3.6}$$

hold if and only if $w(Z|X) = f^{-1}(Z|X)$ and $d(X) = \delta^D(X)$ almost surely.

There are two difficulties with identifying restrictions (3.5) and (3.6). First, they hold on the entire functional space. In practice, it is impossible to solve an infinite number of moment restrictions. Second, the unknown functions $f^{-1}(Z|X)$ and $\delta^D(X)$ are infinite dimensional. It is impossible to estimate the infinite dimensional parameter from finite samples. To overcome both difficulties, we follow the sieve literature by approximating the original functional space with a finite dimensional sieve space. Specifically, let $u_K(X) = (u_{K,1}(X), \ldots, u_{K,K}(X))^\top$ denote known basis functions that can approximate any integrable function $u(X)$ arbitrarily well as K goes to infinity (see Hirano et al. (2003)). The linear sieve space spanned by $u_K(X)$ is an approximation of the original functional space. The sieve version of (3.5)–(3.6) is

$$\mathbb{E}\left[Z \cdot w(Z|X)u_K(X)\right] = \mathbb{E}[u_K(X)] = \mathbb{E}\left[(1-Z)w(Z|X)u_K(X)\right], \qquad (3.7)$$

$$\mathbb{E}\left[d(X)u_K(X)\right] = \mathbb{E}\left[D\left\{(2Z-1)w(Z|X)u_K(X)\right\}\right]. \qquad (3.8)$$

Since the sieve space is a subspace of the original functional space, $w(Z|X) = f^{-1}(Z|X)$ and $d(X) = \delta^D(X)$ is still a solution to (3.7) and (3.8) but not the only one. For example, for any globally concave and increasing function $\rho(v)$, let $\lambda_K \in \mathbb{R}^K$ and $\beta_K \in \mathbb{R}^K$ maximize the objective function:

$$G(\lambda, \beta) = \mathbb{E}\left[Z\rho(\lambda^\top u_K(X)) - \lambda^\top u_K(X)\right] + \mathbb{E}\left[(1-Z)\rho(\beta^\top u_K(X)) - \beta^\top u_K(X)\right].$$

Denote $w_K(Z|X) = Z\rho'(\lambda_K^\top u_K(X)) + (1-Z)\rho'(\beta_K^\top u_K(X))$. For any globally concave function $\rho_1(v)$, let $\gamma_K \in \mathbb{R}^K$ maximize the objective function:

$$H(\gamma) = \mathbb{E}\left[\rho_1(\gamma^\top u_K(X)) - D\left\{(2Z-1)w_K(Z|X)\right\} \times \gamma^\top u_K(X)\right].$$

Denote $d_K(X) = \rho_1'(\gamma_K^\top u_K(X))$. Then, $(w_K(Z|X), d_K(X))$ solves (3.7) and (3.8). Since there are infinite choices in $\rho(v)$ and $\rho_1(v)$, there is an infinite number of solutions. While all these solutions converge to $f^{-1}(Z|X)$ and $\delta^D(X)$ as $K \to +\infty$, not all of them satisfy the boundedness condition: $w_K(Z|X) > 1$ and $-1 \le d_K(X) \le 1$. The functions $\rho(v) = v - \exp(-v)$ and $\rho_1(v) = -\log(e^v + e^{-v})$ produce the solution satisfying the boundedness condition and are adopted in this paper.

3.2 Estimation

We now implement the above idea in finite samples. Let $\{Y_i, X_i, Z_i, D_i\}_{i=1}^N$ denote a sample drawn independently from the joint distribution of (Y, X, Z, D). The sample analogue of $G(\lambda, \beta)$ is

$$\begin{aligned}\widehat{G}(\lambda,\beta) =& \frac{1}{N}\sum_{i=1}^N \left\{Z_i\rho(\lambda^\top u_K(X_i)) - \lambda^\top u_K(X_i)\right\} \\ &+ \frac{1}{N}\sum_{i=1}^N \left\{(1-Z_i)\rho(\beta^\top u_K(X_i)) - \beta^\top u_K(X_i)\right\}.\end{aligned} \tag{3.9}$$

Denote $(\hat{\lambda}_K, \hat{\beta}_K) = \arg\max_{\lambda,\beta} \widehat{G}(\lambda, \beta)$ and

$$\widehat{w}_K(Z_i|X_i) = Z_i\rho'(\hat{\lambda}_K^\top u_K(X_i)) + (1-Z_i)\rho'(\hat{\beta}_K^\top u_K(X_i)),\ i = 1, 2, ..., N. \tag{3.10}$$

The sample analogue of $H(\gamma)$ is

$$\widehat{H}(\gamma) = \frac{1}{N}\sum_{i=1}^N \rho_1(\gamma^\top u_K(X_i)) - \frac{1}{N}\sum_{i=1}^N D_i\{(2Z_i - 1)\widehat{w}_K(Z_i|X_i)\} \cdot \gamma^\top u_K(X_i).$$

With $\hat{\gamma}_K = \arg\max_\gamma \widehat{H}(\gamma)$ and $\widehat{d}_K(X_i) = \rho_1'(\widehat{\gamma}^\top u_K(X_i))$ for $i = 1, 2, ..., N$, then $(\widehat{w}_K(Z_i|X_i), \widehat{d}_K(X_i))$ for $i = 1, 2, ..., N$ is a solution to

$$\frac{1}{N}\sum_{i=1}^N Z_i w(Z_i|X_i) u_K(X_i) = \frac{1}{N}\sum_{i=1}^N u_K(X_i) = \frac{1}{N}\sum_{i=1}^N (1 - Z_i) w(Z_i|X_i) u_K(X_i), \quad (3.11)$$

$$\frac{1}{N}\sum_{i=1}^N d(X_i) u_K(X_i) = \frac{1}{N}\sum_{i=1}^N D_i \left\{(2Z_i - 1) w(Z_i|X_i) u_K(X_i)\right\}. \quad (3.12)$$

In Section 4 of the supplemental material, we show that $(\widehat{w}_K(Z|X), \widehat{d}_K(X))$ is a consistent estimator of $(f^{-1}(Z|X), \delta^D(X))$.

The question now is whether the plug-in of $(\widehat{w}_K(Z|X), \widehat{d}_K(X))$ delivers the efficiency of τ. There are two hopeful signs. First, the semiparametric efficiency bound of τ suggests the importance of the restrictions (3.11) and (3.12). $(\widehat{w}_K(Z|X), \widehat{d}_K(X))$ satisfies these restrictions. Second, note that $(\widehat{w}_K(Z|X), \widehat{d}_K(X))$ is the dual solution to a *calibration* problem similar to those studied by Hainmueller (2012), Chan, Yam, and Zhang (2016), and Zhao (2019). The *calibration* method seeks to solve the restrictions (3.11) and (3.12), and at the same time choose the weights closest to the prespecified design weights, say, $\boldsymbol{q} = (q_1, ..., q_N)$. For example, the corresponding *calibration* problem for $\rho(v) = v - \exp(-v)$ and the estimated weights $(\widehat{w}_K(Z_i|X_i), i = 1, 2, ..., N)$ is

$$\min_{w(.|.)} \sum_{i=1}^N D\left(w(Z_i|X_i), q_i\right) \text{ subject to } (3.11),$$

with the uniform weights $\boldsymbol{q} = (2, ..., 2)$ and the distance measure $D(v, 2) =$

$(\upsilon-1)\log(\upsilon-1)-(\upsilon-1)+1$ (see Chan, Yam, and Zhang (2016, Appendix B) for the derivation). The corresponding *calibration* problem for $\rho_1(\upsilon)=-\log(e^{\upsilon}+e^{-\upsilon})$ and the estimated weights $(\widehat{d}_K(X_i), i=1,2,...,N)$ can also be constructed similarly. This equivalence result suggests that our estimated weights have some optimality property that could help the efficiency of the plug-in estimator.

Our method to estimate the weights is similar to the covariate-balancing method proposed by Imai and Ratkovic (2014). For example, they modeled $f(Z|X)$ by $f(Z|X;\varsigma)$ with $\varsigma\in\mathbb{R}^L$ and proposed estimating ς from the moment restriction: $\mathbb{E}\left[Zf^{-1}(1|X;\varsigma)g(X)\right]=\mathbb{E}[g(X)]$, where $g(X)$ is a prespecified M-dimensional vector of known functions with $M\geq L$, using the generalized method of moments or empirical likelihood method. Their estimator of ς may not solve the sample analogue of the above moment restriction when $M>L$ and they considered fixed M. They showed that their covariate-balancing method improves the performance of their matching and weighting estimators. By contrast, we estimate $f^{-1}(1|X)$ nonparametrically by solving the sample analogue of $\mathbb{E}\left[Zf^{-1}(1|X)g(X)\right]=\mathbb{E}[g(X)]$. In addition, the number of moment restrictions M must grow to ensure the consistency of the nonparametric estimate. Despite this difference, our method and the covariate-balancing method share the same idea. In this sense, our method should be viewed as an extension of the covariate-balancing method to nonparametric estimation. Our estimated weights are called covariate-balancing weights and our plug-in estimator is called the covariate-balancing estimator.

Using the covariate-balancing weights, we estimate τ by plugging in

$$\widehat{\tau} = \frac{1}{N} \sum_{i=1}^{N} \frac{\{(2Z_i - 1)\widehat{w}_K(Z_i|X_i)\} Y_i}{\widehat{d}_K(X_i)}.$$

In contrast to the existing estimators of τ, our plug-in estimator $\widehat{\tau}$ has at least two advantages. First, $\widehat{w}_K(Z|X)$ and $\widehat{d}_K(X)$ satisfy the restrictions (3.11) and (3.12), which improves the performance of the estimated ATE. If the number of moment restrictions (i.e., K) is sufficiently large, our method is unlikely to produce extreme weights, thereby improving the finite sample performance of $\widehat{\tau}$ (see Imai and Ratkovic (2014) for a simulation study of how the covariate-balancing method dramatically improves the poor performance of the propensity score matching and weighting estimator, as reported by Kang and Schafer (2007)). Second, our estimator does not require modeling any functionals and is always efficient.

How do the design weights affect the plug-in estimator (see alsoHainmueller (2012, Section 3.3))? For instance, we could use some initial consistent estimate of $f^{-1}(Z|X)$ as the design weights or even $f^{-1}(Z|X)$. We show in Section 8 of the supplementary material that these design weights, uniform or not and estimated or not, do not influence the asymptotic efficiency of τ. This is not surprising since the weighting functions are determined uniquely by the moment restrictions, not by the design weights. In finite samples, Hainmueller (2012) and Zhao (2019) argued in favor of uniform design weights regardless of the true design weights. In light of their argument, we stick to uniform design weights in this paper.

4. Large Sample Properties

To establish the large sample properties of $\widehat{\tau}$, we impose the following assumptions:

Assumption 3. $\mathbb{E}\left[\frac{1}{\delta^D(X)^2}\right] < \infty$ *and* $\mathbb{E}\left[\frac{Y^2}{\delta^D(X)^4}\right] < \infty$.

Assumption 4. *The support $\mathcal{X}$ of the r-dimensional covariate X is a Cartesian product of r compact intervals.*

Assumption 5. *There exist two positive constants $\overline{C}$ and $\underline{C}$ such that*

$$0 < \underline{C} \leq \lambda_{\min}\left(\mathbb{E}\left[u_K(X)u_K^\top(X)\right]\right) \leq \lambda_{\max}\left(\mathbb{E}\left[u_K(X)u_K^\top(X)\right]\right) \leq \overline{C} < \infty,$$

where $\lambda_{\max}\left(\mathbb{E}\left[u_K(X)u_K^\top(X)\right]\right)$ (resp. $\lambda_{\min}\left(\mathbb{E}\left[u_K(X)u_K^\top(X)\right]\right)$) denotes the largest (resp. smallest) eigenvalue of $\mathbb{E}\left[u_K(X)u_K^\top(X)\right]$.

Assumption 6. *There exist three positive constants* $\infty > \eta_1 > \eta_2 > 1 > \eta_3 > 0$ *such that* $\eta_2 \leq f^{-1}(z|x) \leq \eta_1$ *and* $-\eta_3 \leq \delta^D(x) \leq \eta_3$*,* $\forall (z,x) \in \{0,1\} \times \mathcal{X}$*.*

Assumption 7. *There exist* λ_K*,* β_K*, and* γ_K *in* $\mathbb{R}^K$ *and* $\alpha > 0$ *such that for any* $z \in \{0,1\}$*,*

$$\sup_{x \in \mathcal{X}} \left|(\rho'^{-1}\left(f^{-1}(z|x)\right) - z \cdot \lambda_K^\top u_K(x) - (1-z) \cdot \beta_K^\top u_K(x)\right| = O(K^{-\alpha}),$$

$$\sup_{x \in \mathcal{X}} \left|(\rho_1'^{-1}\left(\delta^D(x)\right) - \gamma_K^\top u_K(x)\right| = O(K^{-\alpha}).$$

Assumption 8. $\zeta(K)^4 K^3/N \to 0$ *and* $\sqrt{N} K^{-\alpha} \to 0$*, where* $\zeta(K) = \sup_{x \in \mathcal{X}} \|u_K(x)\|$ *and* $\|\cdot\|$ *is the usual Frobenius norm defined by* $\|A\| = \sqrt{tr(AA^\top)}$ *for any matrix* A*.*

Assumption 3 is needed to bound the asymptotic variance. This condition is satisfied if for each X the correlation between the instrument and treatment variable is bounded away from zero. Assumption 4 restricts the covariates to be bounded. This condition is admittedly restrictive but is commonly imposed in the nonparametric regression literature. Assumption 4 can be relaxed if we restrict the tail distribution of X. For example, Chen, Hong, and Tarozzi (2008, Assumption 3) allowed the support of X to be the entire Euclidean space but imposed $\int_{\mathbb{R}^r}(1+|x|^2)^\alpha f_X(x)dx < \infty$ for some $\alpha > 0$. Assumption 5 rules out near multicollinearity in the approximating basis functions. It can be satisfied by the orthonormalization of the basis functions. A condition of this type is familiar in the sieve regression literature (Chan et al., 2016). Assumption 6 is a matching condition. It requires that individuals in the treatment and control groups can be matched and that individuals in the group when $Z = 1$ and the group when $Z = 0$ can be matched. The second part of the matching condition is satisfied if for each X the correlation between the instrument and treatment variable is bounded away from one. This condition does not imply the first part of Assumption 3. For example, if $\delta^D(X) = X$ with X uniformly distributed over $[-1/2, 1/2]$, then $\delta^D(x)$ satisfies $-\eta_3 \leq \delta^D(x) \leq \eta_3$ with $\eta_3 = 1/2$, but $\mathbb{E}\left[\{\delta^D(X)\}^{-2}\right] = \infty$ does not satisfy the first part of Assumption 3.

Assumption 7 is a condition on the sieve approximation error. It requires that the approximation error shrinks to zero at a polynomial rate. This condition is satisfied by the power series and B -splines with $\alpha = s/r$, where s measures the smoothness of the function to be approximated and r is the dimension of X. A condition of this type is also common in the sieve regression literature (e.g., Hirano et al. (2003), Chan et al. (2016)).

Assumption 8 regulates the growing rate of the smoothing parameter (i.e., K) relative to the sample size. The term $\zeta(K)$ depends on the type of sieve basis $u_K(X)$. For example, $\zeta(K) = O(\sqrt{K})$ for B-splines and

$\zeta(K) = O(K)$ for power series. In the case of power series, Assumption 8 is satisfied by $K = O(N^{\nu})$ for some $r/2s < \nu < 1/7$, which is weaker than $r/[2(s-2r)] < \nu < 1/9$, the condition imposed by Hirano, Imbens, and Ridder (2003). Assumption 8 implies that our method suffers from the curse of dimensionality. Hence, we investigate the potential impact of dimensionality on the performance of our estimator and find that our estimator does not perform well when $r = 10$ (see Section 1.3 of the supplementary material). Such dimensionality poses two problems: the approximation error shrinks to zero slowly and is difficult to balance in small samples. How to deal with both problems is an important and interesting future research avenue.

Under these assumptions, the following theorem is proven in the supplemental material.

Theorem 2. *If Assumptions 3–8 are satisfied, we obtain: (i)* $\hat{\tau} \xrightarrow{p} \tau$ *and (ii)* $\sqrt{N}(\hat{\tau} - \tau) \xrightarrow{d} N(0, V_{eff})$*, where* V_{eff} *is the semiparametric efficiency bound derived by Wang and Tchetgen Tchetgen (2018).*

5. Estimation of V_{eff}

We now present an easy way to compute the variance for $\widehat{\tau}$. The idea is to view $\hat{\theta} = (\hat{\lambda}_K^{\top}, \hat{\beta}_K^{\top}, \hat{\gamma}_K^{\top}, \hat{\tau})^{\top}$ as the moment estimator of $\theta = (\lambda^{\top}, \beta^{\top}, \gamma^{\top}, \tau)^{\top}$

and then apply the variance formula of the moment estimator. Specifically, denote

$$g_1(Z,X;\theta)=\left\{Z\rho'\left(\lambda^\top u_K(X)\right)-1\right\}u_K(X)^\top,$$

$$g_2(Z,X;\theta)=\left\{(1-Z)\rho'\left(\beta^\top u_K(X)\right)-1\right\}u_K(X)^\top,$$

$$g_3(Z,D,X;\theta)=\left\{\rho_1'\left(\gamma^\top u_K(X)\right)-D\left\{Z\cdot\rho'\left(\lambda^\top u_K(X)\right)-(1-Z)\rho'\left(\beta^\top u_K(X)\right)\right\}\right\}u_K(X)^\top,$$

$$g_4(Z,D,X,Y;\theta)=\left\{Z\cdot\rho'\left(\lambda^\top u_K(X)\right)-(1-Z)\rho'\left(\beta^\top u_K(X)\right)\right\}Y/\rho_1'\left(\gamma^\top u_K(X)\right)-\tau.$$

Denote $g(Z,D,X,Y;\theta)=(g_1(Z,X;\theta),g_2(Z,X;\theta),g_3(Z,D,X;\theta),g_4(Z,D,X,Y;\theta))^\top$. Then, $\hat{\theta}$ is the moment estimator solving the moment condition: $N^{-1}\sum_{i=1}^N g(Z_i,D_i,X_i,Y_i;\hat{\theta})=0$. If the smoothing parameter K were fixed, we would compute the covariance matrix of the moment estimator $\hat{\theta}$ as $\widehat{V}_\theta=\hat{L}^{-1}\cdot\widehat{\Omega}\cdot(\hat{L}^{-1})^\top$, where $\hat{L}=N^{-1}\sum_{i=1}^N\partial g(Z_i,D_i,X_i,Y_i;\hat{\theta})/\partial\theta$ and $\hat{\Omega}=N^{-1}\sum_{i=1}^N g(Z_i,D_i\ ,X_i,Y_i;\hat{\theta})\times g(Z_i,D_i,X_i,Y_i;\hat{\theta})^\top$. Let $\mathbf{e}_{3K+1}$ be a $(3K+1)$-dimensional column vector whose last element is 1 and the other elements are 0. Denote

$$\widehat{V}=\mathbf{e}_{3K+1}^\top\left\{\hat{L}^{-1}\cdot\hat{\Omega}\cdot(\hat{L}^{-1})^\top\right\}\mathbf{e}_{3K+1}.$$

Then, $\hat{V}$ would be a consistent estimator of V_{eff}. The next theorem proves that $\hat{V}$ is still a consistent estimator of V_{eff} even if K grows rather than being fixed.

Theorem 3. *Suppose that Assumptions 3–8 are satisfied and* $\mathbb{E}\left[\frac{Y^4}{\delta^D(X)^4}\right]<\infty$ *holds. Then,* $\hat{V}$ *is a consistent estimator of* V_{eff}.

A weighted bootstrap method can be used to estimate the asymptotic variance. This is a useful alternative estimator that could have better finite sample performance than our variance estimator, although at somewhat higher computing costs. For more details, see Cheng (2015).

6. Estimation of the LATE

With the covariate-balancing weights, we estimate the LATE using the plug-in method. Applying the tower law of conditional expectation, we obtain

$$\tau_{LATE} = \frac{1}{\mathbb{E}\left[\delta^D(X)\right]} \cdot \mathbb{E}\left[\left\{\frac{2Z-1}{f(Z|X)}\right\} Y\right].$$

The plug-in estimator is given by

$$\widehat{\tau}_{LATE} = \sum_{i=1}^{N} \left\{(2Z_i - 1)\widehat{w}_K(Z_i|X_i)\right\} Y_i \Big/ \sum_{j=1}^{N} \widehat{d}_K(X_j),$$

and its large sample properties are summarized in the following theorem.

Theorem 4. *Under Assumptions 3–8, we have* $\sqrt{N}(\hat{\tau}_{LATE} - \tau_{LATE}) \xrightarrow{d} N(0, V_{LATE})$*, where* V_{LATE} *is the semiparametric efficiency bound derived by Frölich (2007).*

A consistent estimator of V_{LATE} can be constructed using the same approach as in Section 5.

7. Data-driven Smoothing Parameter

The large sample properties of the proposed estimators permit a wide range of values of K. This presents a dilemma for applied researchers who have only one finite sample and would like to have some guidance on the smoothing parameter. In this section, we present a data-driven approach to determine K. Note that $f(Z|X)^{-1}$ and $\delta^D(X)$ satisfy the following regression equations: $\mathbb{E}\left[Zf^{-1}(Z|X)|X\right] = 1 = \mathbb{E}\left[(1-Z)f^{-1}(Z|X)|X\right]$ and $\mathbb{E}\left[D\{2Z-1\}f^{-1}(Z|X)|X\right] = \delta^D(X)$. We choose the smoothing parameter to minimize the mean squared errors of the above moment conditions. Since there are two unknown functions, we use two smoothing parameters K_1, K_2. Let $(\hat{\lambda}_{K_1}, \hat{\beta}_{K_1})$ denote the maximizer of $\widehat{G}(\lambda,\beta)$ with $u_K(X_i)$ replaced by $u_{K_1}(X_i)$ and $\hat{\gamma}_{K_2}$ denote the maximizer of $\widehat{H}(\gamma)$ with $u_K(X_i)$ replaced by $u_{K_2}(X_i)$ and $\widehat{w}_K(Z_i|X_i)$ replaced with $\widehat{w}_{K_1}(Z_i|X_i)$. The penalized mean-squared-errors are defined by

$$pMSE_1(K_1) = \frac{\sum_{i=1}^N \{Z_i\hat{w}_{K_1}(Z_i|X_i)-1\}^2 + \{(1-Z_i)\hat{w}_{K_1}(Z_i|X_i)-1\}^2}{(1-(K_1^2/N))^2},$$

$$pMSE_2(K_1,K_2) = \frac{\sum_{i=1}^N \left\{D_i(2Z_i-1)\hat{w}_{K_1}(Z_i|X_i) - \hat{d}_{K_2}(X_i)\right\}^2}{(1-K_2^2/N)^2}.$$

We choose K_1 and K_2 to minimize $pMSE_1$ and $pMSE_2$. Specifically, denote the upper bound of K_1 and K_2 by $\bar{K}_1$ and $\bar{K}_2$, respectively. The data-driven K_1 and K_2 are computed by $\hat{K}_1 = \arg\min_{K_1\in\{1,\ldots,\bar{K}_1\}} pMSE_1(K_1)$,

$\hat{K}_2 = \arg\min_{K_2 \in \{1,\ldots,\bar{K}_2\}} pMSE_2(\hat{K}_1, K_2)$.

8. Simulation Studies

In this section, we conduct and report a simulation study with a one-dimensional covariate to evaluate the finite sample performance of our ATE estimator. We consider a scenario identical to the design in Wang and Tchetgen Tchetgen (2018), which contains a univariate covariate. The scenario of multivariate covariates is relegated to the supplemental material because of space constraints. We also conduct simulation studies of the finite sample performance of the LATE estimator; these results are also reported in the supplemental material. The measured confounders are $X = (1, X_1)$ with X_1 uniformly distributed on $(-1, -0.5) \cup (0.5, 1)$. The unmeasured confounder U is a Bernoulli random variable with mean 0.5. The data-generating process for (Z, D, Y) is parameterized as

$$\Pr(Z = 1 | X, U) = \text{expit}(\gamma^\top X),$$

$$\Pr(D = 1 | Z, X, U) = p_0^D(X) + Z \times \mu^D(X) + \kappa(2U - 1),$$

$$\Pr(Y = 1 | Z, X, U) = p_0^Y(X) + Z \times \mu^Y(X) + \kappa(2U - 1),$$

where $\mu^D(X) = \tanh(\beta^\top X)$ and $\mu^Y(X) = \tanh(\beta^\top X) \times \tanh(\alpha^\top X)$, and for $j \in \{D, Y\}$,

$$p_0^j(X) = \frac{\{\psi^j(X) - (\psi^j(X)^2 + 4\mathrm{OP}^j(X)\{1 - \mu^j(X)\}\{1 - \mathrm{OP}^j(X)\})^{1/2}\}}{2\{\mathrm{OP}^j(X) - 1\}}$$

with $OP^j(X) = \exp(\eta_j^\top X)$ and $\psi^j(X) = \mathrm{OP}^j(X)\{2 - \mu^j(X)\} + \mu^j(X)$. The true values of the model parameters are set to $\alpha = (0.1, 0.5)^\top$, $\beta = (0, -0.5)^\top$, $\gamma = (0.1, -0.5)^\top$, $\zeta = (0, -1)^\top$, $\eta = (-0.5, 1)^\top$, and $\kappa = 0.1$. The ATE in this case is $\tau = 0.087$.

In each Monte Carlo run, a sample of 500 observations and 1000 observations are generated from the above data-generating process. To evaluate the performance of our estimator (also called CBE to stand for the covariate-balancing estimator hereafter)) relative to the alternative estimators proposed in the literature, we compute all the estimators from each sample and repeat the Monte Carlo run 500 times.

The alternative estimators are the naive estimator, multiply robust estimator (MR) and bounded multiply robust estimator (B-MR) proposed by Wang and Tchetgen Tchetgen (2018), and estimator proposed by Hirano, Imbens, and Ridder (2003) (HIR hereafter). The details of calculations are given below.

1. The CBE is computed with the power series and data-driven smooth-

ing parameters (K_1, K_2), where the upper bounds $\overline{K}_1$ and $\overline{K}_2$ are set at 3 in this scenario;

2. The naive estimator is computed as the difference in the group means between the treatment and control groups;

3. The MR and B-MR are computed using the procedures proposed by Wang and Tchetgen Tchetgen (2018). Wang and Tchetgen Tchetgen (2018) considered several cases in which some or all functionals are misspecified (i.e., in addition to their models $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$, we add the case in which all the functions are correctly specified (*All*) and the case in which all the functions are misspecified ($None$)).

4. The HIR estimator is computed as

$$\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{2Z_i-1}{\widehat{E}(D|Z_i=1,X_i)-\widehat{E}(D|Z_i=0,X_i)}\right\}\frac{Y_i}{\widehat{E}(Z_i|X_i)},$$

where $\widehat{E}(Z_i|X_i)$ is the fitted value from the logit regression of Z_i on $u_{K_1}(X_i)$, $\widehat{E}(D|Z_i=0,X_i)$ is the fitted value from the logit regression of D_i on $u_{K_2}(X_i)$ using the subsample $Z_i=0$, and $\widehat{E}(D|Z_i=1,X_i)$ is the fitted value from the logit regression of D_i on $u_{K_2}(X_i)$ using the subsample $Z_i=1$. The smoothing parameters K_1 and K_2 are determined by the data-driven method proposed in Section 7.

Table 1: Simulation results of the estimated ATEs

$N = 500$				
Estimators	Bias	Stdev	RMSE	CP
Naive	-0.057	0.045	0.073	0.77
MR (All)	-0.014	0.146	0.147	0.942
MR ($\mathcal{M}_1$)	-0.011	0.147	0.147	0.942
MR ($\mathcal{M}_2$)	-0.021	0.159	0.161	0.952
MR ($\mathcal{M}_3$)	16.14	363.67	364.02	0.976
MR (None)	-22.90	342.10	342.87	0.982
B-MR (All)	-0.006	0.151	0.151	0.942
B-MR ($\mathcal{M}_1$)	-0.028	0.171	0.174	0.952
B-MR ($\mathcal{M}_2$)	-0.031	0.210	0.212	0.958
B-MR ($\mathcal{M}_3$)	-0.007	0.146	0.146	0.942
B-MR (None)	0.110	0.641	0.651	1
HIR	-0.014	0.161	0.162	0.954
CBE	0.003	0.152	0.152	0.96
$N = 1000$				
Estimators	Bias	Stdev	RMSE	CP
Naive	-0.056	0.031	0.064	0.586
MR (All)	-0.002	0.100	0.100	0.96
MR ($\mathcal{M}_1$)	0.000	0.100	0.100	0.96
MR ($\mathcal{M}_2$)	-0.008	0.120	0.120	0.954
MR ($\mathcal{M}_3$)	-43.92	754.17	755.45	0.99
MR (None)	-42.84	725.97	727.23	0.988
B-MR (All)	0.003	0.102	0.102	0.966
B-MR ($\mathcal{M}_1$)	-0.022	0.133	0.135	0.95
B-MR ($\mathcal{M}_2$)	-0.007	0.138	0.138	0.976
B-MR ($\mathcal{M}_3$)	-0.003	0.108	0.108	0.958
B-MR (None)	0.299	0.621	0.689	1
HIR	-0.015	0.120	0.121	0.956
CBE	0.004	0.110	0.110	0.954

Table 1 reports the bias, standard deviation, root mean square error (RMSE), and coverage probability (CP) at the nominal size $\alpha = 0.95$. The table shows several observations. First, the naive estimator is badly biased, which is not surprising since the unmeasured confounder is not controlled for. Second, the B-MR performs as well as the MR when some of the functionals ($\mathcal{M}_1$, $\mathcal{M}_2$) or all the functionals (All) are correctly specified, but performs substantially better than the MR when most of or all the functionals are misspecified. The CPs of both the MR and the B-MR are close to the nominal size in all the cases except when the model is badly misspecified. Despite its out-performance over the MR, the B-MR is still biased when the model is badly misspecified. In most cases, the CP of the HIR is higher than the nominal size. Fourth, the CBE with the data-driven smoothing parameters is unbiased and its CP is around the nominal size, suggesting that the asymptotic theory is a good approximation. Finally, we investigate the stability of the covariate-balancing weights: $\{\widehat{w}_{K_1}(Z_i|X_i), \widehat{d}_{K_2}(X_i)\}_{i=1}^N$. In Section 1.2 of the supplemental material, we plot their empirical distributions in the j^{th} Monte Carlo run for $j \in \{50, 100, 150, 200, 250, 300, 350, 400, 450\}$ and $N = 500$. The plots show that $\{\widehat{w}_{K_1}(1|X_i)\}_{i=1}^N$ are distributed over the interval $[1.25, 2.75]$, $\{\widehat{w}_{K_1}(0|X_i)\}_{i=1}^N$ are distributed over the interval $[1.25, 3.75]$, and $\{\widehat{d}_{K_2}(X_i)\}_{i=1}^N$ are distributed over the interval

$[-0.625, -0.125] \cup [0.125, 0.625]$. In all the cases, the covariate-balancing weights do not contain extreme values.

9. Empirical Application

To evaluate the practical value of our method, we revisit the return to education study of Card (1995). The data are from the National Longitudinal Survey of Young Men, which contains observations on 5525 men aged between 14 and 24 years in 1966. Among them, 3010 provided valid education and wage responses in the 1976 follow-up survey. The parameter of interest is the causal effect of education on earnings. The unmeasured confounder is the preference for education. The treatment variable is education beyond high school (also see Wang and Tchetgen Tchetgen (2018)). Hence, the treatment group includes those who attended college. The earnings variable is wage dichotomized at its median of \$5.375 per hour. Apart from the preference for education, the distance from home to the nearest four-year college is also a deciding factor in attending college. Thus, the dummy for the nearby four-year college is a valid instrument for the treatment variable.

Other measured confounders include race, parents' education, indicators for residence in a metropolitan area in 1966, experience, and experience squared. Race, parents' education, and residence are included because they may affect both the instrument and the outcome. Experience and experience squared, as a measure of underlying ability, are included, as they may modify both the effect of proximity to college on education and the effect of education on earnings. Following Card (1995), the missing values are imputed by the mean. The National Longitudinal Survey is not a representative sample of the US population. This is accounted for by weighting observations using their sampling weights.

We compute the (i) naive estimator, which is the difference between the sample means of those who went to college and those who did not; (ii) two-stage least squares (2SLS) estimator proposed by Angrist and Pischke (2008); (iii) five estimators proposed by Wang and Tchetgen Tchetgen (2018), namely, the regression-based estimator (REG), bounded regression-based estimator (B-REG), inverse probability weighting estimator (IPW), bounded probability weighting estimator (B-IPW), and g-estimator (g) as well as the MR and B-MR; and (iv) CBE. Table 2 reports the point estimates. The confidence intervals are computed using the quantile-based nonparametric bootstrap method, which generates 500 samples through empirical bootstrapping.

Table 2 suggests the following findings. First, the naive approach shows that education has a significantly positive effect on earnings. Since people who pursue more education may have stronger intentions and higher ability to find a job with higher wages than those who pursue less education, the naive estimator could overestimate the effect. The 2SLS approach produces a highly significant and positive effect above 1. This is clearly incorrect since the range of the ATE is $[-1, 1]$ by design. The B-IPW estimate is greater than the naive estimate, which is greater than the true effect. In addition, its confidence interval is too wide, covering the whole range $[-1, 1]$. The MR estimate is unreasonable. The B-REG, g, and B-MR estimates are all highly positive effects, but they are all greater than the naive estimates. By contrast, the CBE estimate is positive and less than the naive estimate. However, it is not significant at the 5% level, suggesting that return to education is positive but small.

10. Discussion

This paper proposes the estimation of the ATE and LATE when some confounders are unmeasured. The proposed estimators do not require modeling any functionals and are consistent and efficient. A small simulation study shows that the proposed estimator has good finite sample performance. An application to education data illustrates the usefulness of the proposed approach.

Table 2: Estimates of the effect of education beyond high school on earnings (dichotomized at the median)

Method	Point estimate	95% confidence interval
Naive	0.122	$(0.084, 0.160)$
2SLS	1.057	$(0.335, 2.236)$
B-REG	0.839	$(0.587, 0.973)$
B-IPW	0.124	$(-1.00, 1.00)$
g	0.682	$(-0.001, 1.00)$
MR	63.74	$(-56.44, 68.52)$
B-MR	0.681	$(0.062, 1.00)$
CBE	0.116	$(-0.085, 0.143)$

This paper focuses on the binary treatment model. However, multivalued or continuous treatments are common in applications. The extension of the idea in this paper to those models is thus worth pursuing in a future project.

Supplementary Material

This supplemental material contains the technical proofs for Theorems 1–4 and partial simulation results, which are not presented in the main text.

Acknowledgements

The authors sincerely thank the Co-Editor Hans-Georg Müller, Associate Editor, and two referees for their constructive suggestions and comments. We also express our gratitude to Yu-Chin Hsu for his helpful comments. Chunrong Ai acknowledges the financial support from the National Natural Science Foundation of China through project 71873138. Lukang Huang acknowledges the financial support from the Renmin University of China. The present work constitutes part of her research study leading to her doctoral thesis. Zheng Zhang acknowledges the financial support from the National Natural Science Foundation of China through project 12001535, and the fund for building world-class universities (disciplines) of the Renmin

University of China.

References

Abadie, A. and G. W. Imbens (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica 74*(1), 235–267.

Angrist, J. D. and J.-S. Pischke (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Card, D. (1995). Using geographic variation in college proximity to estimate the return to schooling. *In Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp (eds L. N. Christofides, R. Swidinsky and E. K. Grant)*, 201–222.

Chan, K. C. G., S. C. P. Yam, and Z. Zhang (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*(3), 673–700.

Chen, X., H. Hong, and A. Tarozzi (2008). Semiparametric efficiency in gmm models with auxiliary data. *Annals of Statistics 36*(2), 808–843.

Cheng, G. (2015). Moment consistency of the exchangeably weighted bootstrap for semiparametric m-estimation. *Scandinavian Journal of Statistics 42*(3), 665–684.

Donald, S. G., Y. Hsu, and R. P. Lieli (2014a). Testing the unconfoundedness assumption via inverse probability weighted estimators of (l)att. *Journal of Business & Economic Statistics 32*(3), 395–415.

Donald, S. G., Y. C. Hsu, and R. P. Lieli (2014b). Inverse probability weighted estimation of local average treatment effects: A higher order mse expansion. *Statistics & Probability Letters 95*(C), 132–138.

Frölich, M. (2007). Nonparametric iv estimation of local average treatment effects with covariates. *Journal of Econometrics 139*(1), 35–75.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica 66*(2), 315–331.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis 20*(1), 25–46.

Hirano, K., G. Imbens, and G. Ridder (2003, July). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

Huang, M.-Y. and K. C. G. Chan (2017). Joint sufficient dimension reduction and estimation of conditional and average treatment effects. *Biometrika 104*(3), 583–596.

Imai, K. and M. Ratkovic (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 243–263.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Kang, J. and J. Schafer (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*(4), 523–539.

Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statistical Science 17*(3), 286–327.

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

Wang, L. and E. Tchetgen Tchetgen (2018). Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(3), 531–550.

Zhao, Q. (2019). Covariate balancing propensity score by tailored loss functions. *Annals of Statistics 47*(2), 965–993.

School of Management and Economics and Shenzhen Finance Institute, Chinese University of Hong Kong, Shenzhen, China

E-mail: chunrongai@cuhk.edu.cn

Institute of Statistics and Big Data, Renmin University of China, Beijing, China

E-mail: huanglukang@ruc.edu.cn

Institute of Statistics and Big Data, Renmin University of China, Beijing, China

E-mail: zhengzhang@ruc.edu.cn