Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

A Covariate Selection Criterion for Estimation of Treatment Effects

Xun Lu

To cite this article: Xun Lu (2015) A Covariate Selection Criterion for Estimation of Treatment Effects, Journal of Business & Economic Statistics, 33:4, 506-522, DOI: 10.1080/07350015.2014.982755

To link to this article: https://doi.org/10.1080/07350015.2014.982755

A Covariate Selection Criterion for Estimation of Treatment Effects

Xun Lu
Department of Economics, Hong Kong University of Science and Technology, Kowloon, Hong Kong (*xunlu@ust.hk*)

We study how to select or combine estimators of the average treatment effect (ATE) and the average treatment effect on the treated (ATT) in the presence of multiple sets of covariates. We consider two cases: (1) all sets of covariates satisfy the unconfoundedness assumption and (2) some sets of covariates violate the unconfoundedness assumption locally. For both cases, we propose a data-driven covariate selection criterion (CSC) to minimize the asymptotic mean squared errors (AMSEs). Based on our CSC, we propose new average estimators of ATE and ATT, which include the selected estimators based on a single set of covariates as a special case. We derive the asymptotic distributions of our new estimators and propose how to construct valid confidence intervals. Our Monte Carlo simulations show that in finite samples, our new average estimators achieve substantial efficiency gains over the estimators based on a single set of covariates. We apply our new estimators to study the impact of inherited control on firm performance.

KEY WORDS: Focused information criterion; Model averaging; Treatment effect; Unconfoundedness.

1. INTRODUCTION

One of the central questions in the semiparametric estimation of the average treatment effect (ATE) and the average treatment effect on the treated (ATT) is how to select proper covariates. Despite its importance, the literature on the selection of covariates for estimating treatment effects is limited. Imbens and Wooldridge (2009, p. 50) pointed out that "a very important set of decisions in implementing all of the methods [of estimating treatment effects] . . . involves the choice of covariates," but "the literature has not been very helpful," and "it is clear that more work needs to be done in this area."

In general, there are two important issues concerning the choice of covariates. First, the choice of covariates crucially depends on the unconfoundedness assumption. If this assumption is violated, the estimators of ATE and ATT would in general be inconsistent. In particular, Rosenbaum (1984), Heckman and Navarro-Lozano (2004), and Wooldridge (2005) warned that conditioning on too many covariates, especially those that are influenced by the treatment, can cause the unconfoundedness assumption to be violated easily. Chalak and White (2012) examined how valid covariates should be chosen to satisfy the unconfoundedness assumption by invoking Reichenbach's (1956) principle of common causes: if two variables are correlated, then either one causes the other or there is a common cause of both. Lu and White (2014, Sec. 5) provided further discussions. Second, even if the unconfoundedness assumption is not a concern, using different sets of covariates may yield different estimation efficiencies (i.e., different asymptotic variances). Hahn (2004) and White and Lu (2011) discussed this issue and show that it is not always desirable to condition on fewer or more covariates. The efficient choice of covariates depends on the underlying structures of potential outcomes, treatments and covariates. Intuitively, we should condition on the proxies of potential outcomes as much as possible, but on the proxies of treatments as little as possible. To analyze the underlying causal structures, we may use causal diagrams in the graphical model literature (see, e.g., Pearl 2009; VanderWeele and Shpitser 2011). All these results in the literature offer useful insights into the choice of covariates. However, in practice, the details of the underlying causal structures are rarely known. The goal of this article is to provide a simple data-driven criterion to select or combine estimators of ATE and ATT in the presence of multiple candidate sets of covariates.

Specifically, the parameters we are interested in are ATE and ATT:

$$\text{ATE} = \mathbb{E}\left(Y\left(1\right) - Y\left(0\right)\right) \text{ and } \text{ATT} = \mathbb{E}\left(Y\left(1\right) - Y\left(0\right)|D = 1\right),$$

where D is a binary treatment and $Y\left(0\right)$ and $Y\left(1\right)$ are potential outcomes corresponding to $D = 0$ and $D = 1$, respectively. To estimate them, we often need to use covariates. In practice, researchers may face multiple sets of covariates. The natural criterion for deciding which set to use is the asymptotic mean squared errors (AMSEs) of the estimators of ATE and ATT. In this article, we provide estimators of the AMSEs, which are referred to as the covariate selection criterion (CSC). Based on the CSC, we can select the best estimator or combine all estimators optimally.

We consider two cases. First, we assume that all candidate sets of covariates are valid in the sense that conditioning on each set of covariates, the unconfoundedness assumption is satisfied, that is, $\left(Y\left(0\right), Y\left(1\right)\right)$ and D are independent conditioning on each set of covariates. Thus using any set of covariates leads to a consistent estimator. In this case, we have multiple consistent estimators. Our CSC is essentially the estimator of the asymptotic variance–covariance matrix. Based on the CSC, we propose a weighted average estimator that achieves the smallest asymptotic variance. This estimator is analogous to Lu and White's (2014) feasible optimally combined GLS (FOGLeSs) estimator in the linear regression context.

Second, we allow some sets of covariates to be locally invalid in the sense that the differences between the parameters of interest (i.e., ATE and ATT) and the parameters that can be

Journal of Business & Economic Statistics
October 2015, Vol. 33, No. 4
DOI: 10.1080/07350015.2014.982755

identified (i.e., those that are entirely based on the joint distribution of the observable data) converge to zero at the rate of $\sqrt{n}$, where n is the sample size. This could be because the difference between the distribution of $(Y(0), Y(1))$ conditioning on the covariates and $D = 1$ and that conditioning on the covariates and $D = 0$ converges to zero at the rate of $\sqrt{n}$. Thus, for any given n, the unconfoundedness assumption is violated. However, at the limit when n approaches infinity, the unconfoundedness assumption holds. This local asymptotic framework is often assumed in the model selection and model averaging literature (see, e.g., Claeskens and Hjort 2008, CH hereafter). We show that by using the locally invalid covariates, the standardized estimators of ATE and ATT converge in distribution to a nonzero mean normal distribution. Thus, the squared asymptotic bias is nonzero and of the same order as the asymptotic variance term. This local asymptotic approach essentially allows us to strike a balance between the bias and variance terms. Our CSC estimates both the bias term and the variance term. Based on our CSC, we propose a weighted average estimator that minimizes the AMSE. We show that the asymptotic distribution of the weighted average estimator is nonstandard (a nonlinear function of a normal vector) and propose a simulation-based method to construct confidence intervals.

If the unconfoundedness assumption is violated globally in the sense that the differences between the parameters of interest and the parameters identified are nonzero constants that do not change with the sample size, then the bias term dominates the variance term. Thus, theoretically speaking, we should never use globally invalid covariates. But even so, in the simulations, we apply our CSC to globally invalid covariates and find that our new average estimators perform reasonably well.

This article is related to the model averaging literature. Broadly speaking, this literature can be classified into the Bayesian model averaging (BMA) stream and the frequentist model averaging (FMA) stream. BMA has a long history (see, e.g., Hoeting et al. 1999 for a review), whereas FMA has a relatively short one. Claeskens and Hjort (2003), Hjort and Claeskens (2003), and CH provided detailed discussions on FMA and proposed a focused information criterion (FIC) that is similar to our CSC. Model averaging is mainly applied to estimate conditional means or conditional quantiles. Hansen (2007) and Hansen and Racine (2012) proposed Mallows' criterion and a jackknife method for averaging linear regressions, respectively. Liu (2015) derived the asymptotic distribution of the average estimator in the local asymptotic framework for linear regressions. Chen, Jacho-Chávez, and Linton (2015) considered an averaging GMM estimator with an increasing number of moment condition. Lu and Su (2015) proposed a jackknife method for averaging quantile regressions. DiTraglia (2012) applied the model averaging idea to the selection of instrument variables and proposes a focused moment selection criterion (FMSC).

In the context of estimation of treatment effects, recently, several papers have discussed the model selection or covariate selection. Vansteelandt, Bekaert, and Claeskens (2010) proposed a focused information criterion for covariate selection for estimation of the marginal causal odds ratio. De Luna, Waernbaum, and Richardson (2011) characterized a minimal subset that satisfies the unconfoundedness assumption from the original reservoir of covariates. Crainiceanu, Dominici, and Parmigiani (2008) proposed a two-stage statistical approach that takes into account the uncertainty of covariate selection. There are also an increasing number of papers that apply the model averaging techniques to the estimation of treatment effects. Wang, Parmigiani, and Dominici (2012, WPD hereafter) developed a model averaging method called Bayesian adjustment for confounding (BAC) that jointly considers the treatment assignment and outcome models. Zigler and Dominici (2014, ZD hereafter) proposed a novel Bayesian model averaging method based on the likelihood that simultaneously considers the propensity score and outcome models. Our article differs from these papers in several aspects. First, WPD and ZD's averaging methods are mainly Bayesian, while ours is frequentist. Second, we work in a local asymptotic framework that allows the unconfoundedness assumption to be violated locally, while WPD and ZD assumed that the largest set of covariates satisfies the unconfoundedness assumption, but some covariates may be redundant. Kitagawa and Muris (2013, KM hereafter) recently proposed a model averaging method for the propensity score weighting estimator for estimation of ATT in a local asymptotic framework. There are several differences between KM and our article. First, KM's model averaging method is based on the Bayes optimal weights with respect to the prior for the local parameters, while ours is "plug-in" -type frequentist model averaging as in CH. Second, we do not consider the specification of the functional form of the regression function or propensity score function, as they are nonparametrically estimated. For the tuning parameter in the nonparametric estimation, we simply use a "rule of thumb." KM assumed that the largest (parametric) model of the propensity score is correctly specified and discussed both the choice of covariates and the specification of the propensity score function. Third, their method focuses on a particular estimator of ATT, while ours can be applied to any estimators of ATE and ATT in the presence of multiple sets of covariates.

This article is organized as follows. In Section 2, we discuss individual estimators for given sets of covariates. In Section 3, we propose the CSC and average estimators when all sets of covariates are valid. In Section 4, we allow some sets of covariates to be locally invalid and propose the corresponding CSC and average estimators. In Section 5, we propose a Hausman-type test to test for the unconfoundedness assumption. In Section 6, we conduct Monte Carlo simulations to demonstrate the superior performance of our new average estimators. In Section 7, we apply our new methods to study the effects of inherited family control on firm performance. Section 8 concludes. All mathematical proofs are gathered in the appendix.

2. INDIVIDUAL ESTIMATORS

Consider a binary treatment D_i and potential outcomes of interest $Y_i(0)$ and $Y_i(1)$, corresponding to the possibilities $D_i = 0$ and $D_i = 1$, respectively. The effects of interest are ATE: $\beta_0 \equiv \mathbb{E}[Y_i(1) - Y_i(0)]$ and ATT: $\gamma_0 \equiv \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1]$. We consider J potential sets of covariates $\mathbf{X}_{1i}, \mathbf{X}_{2i}, \ldots,$ and $\mathbf{X}_{Ji}$, where $\mathbf{X}_{ji}$ $(j = 1, \ldots, J)$ is a vector of dimension k_j. We do not impose any restrictions on the J sets of covariates. For example, k_j's can be different, that is, the dimensions of $\mathbf{X}_{ji}$'s can be different. $\mathbf{X}_{ji}$'s can have common or distinct elements and can be nested or nonnested. To construct the candidate sets of covariates, we often need to apply domain knowledge and use a priori information. The outcome we observe is $Y_i =$

$D_i Y_i(1) + (1 - D_i) Y_i(0)$. We assume the data are independent and identically distributed (iid).

Assumption A.1. We observe an iid sample: $\{Y_i, D_i, \mathbf{X}_{1i}, \mathbf{X}_{2i}, \ldots, \mathbf{X}_{Ji}\}, i = 1, \ldots, n$.

Since we consider an iid sample, for notational simplicity, we suppress the subscript i as long as there is no confusion. Given each set of covariates $\mathbf{X}_j$, we define

$$\beta_j \equiv \mathbb{E}\left[m_{1j}\left(\mathbf{X}_j\right) - m_{0j}\left(\mathbf{X}_j\right)\right]$$

and

$$\gamma_j \equiv \mathbb{E}\left[m_{1j}\left(\mathbf{X}_j\right) - m_{0j}\left(\mathbf{X}_j\right) \mid D = 1\right],$$

where

$$m_{1j}\left(\mathbf{X}_j\right) \equiv \mathbb{E}\left(Y \mid D = 1, \mathbf{X}_j\right)$$

and

$$m_{0j}\left(\mathbf{X}_j\right) \equiv \mathbb{E}\left(Y \mid D = 0, \mathbf{X}_j\right).$$

We also define the propensity score as $p_j\left(\mathbf{X}_j\right) \equiv \Pr\left(D = 1 | \mathbf{X}_j\right)$. Further, let $p \equiv \Pr(D = 1) = \mathbb{E}\left(p_j\left(\mathbf{X}_j\right)\right)$.

Here β_0 and γ_0 are the parameters of interest and β_j and γ_j are the parameters that can be identified in the sense that they can be represented entirely in terms of the joint distribution of the observable data. In general, $\beta_j \neq \beta_0$ and $\gamma_j \neq \gamma_0$. However, they are both equal under the key unconfoundedness assumption: conditional on the covariates $\mathbf{X}_j$, $(Y(0), Y(1))$ and D are independent. In this section, we consider the estimators of β_j and γ_j for the J sets of covariates. Note that we do not make the key unconfoundedness assumption in this section.

There are several estimators of β_j and γ_j in the literature. As our CSC is not restricted to any particular estimator, we describe two common ones. These two estimators are efficient in the sense that they both achieve their semiparametric efficiency bounds for the given set of covariates $\mathbf{X}_j$. Let $\hat{\beta}_j$ and $\hat{\gamma}_j$ denote the estimators of β_j and γ_j, respectively. Imbens, Newey, and Ridder (2007) modified Hahn's (1998) estimators and proposed the following imputation estimators:

$$\hat{\beta}_j = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}_{1j}\left(\mathbf{X}_{ji}\right) - \hat{m}_{0j}\left(\mathbf{X}_{ji}\right)\right)$$

and

$$\hat{\gamma}_j = \frac{\frac{1}{n}\sum_{i=1}^{n} D_i \cdot \left(\hat{m}_{1j}\left(\mathbf{X}_{ji}\right) - \hat{m}_{0j}\left(\mathbf{X}_{ji}\right)\right)}{\frac{1}{n}\sum_{i=1}^{n} D_i}, \quad (1)$$

where $\hat{m}_{1j}(.)$ and $\hat{m}_{0j}(.)$ are nonparametric estimators of $m_{1j}(.)$ and $m_{0j}(.)$, respectively. They can be kernel-based Nadaraya–Watson (NW) estimators or sieve-based estimators (see, e.g., Li and Racine 2007). Hirano, Imbens, and Ridder (2003, HIR) proposed the following propensity score weighted (PSW) estimators:

$$\hat{\beta}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i \cdot \left(\frac{D_i}{1 - \hat{p}_j\left(\mathbf{X}_{ji}\right)} - \frac{1 - D_i}{1 - \hat{p}_j\left(\mathbf{X}_{ji}\right)}\right)$$

and

$$\hat{\gamma}_j = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i \cdot \left(D_i - \frac{\hat{p}_j\left(\mathbf{X}_{ji}\right)}{\left(1 - \hat{p}_j\left(\mathbf{X}_{ji}\right)\right)}\left(1 - D_i\right)\right)}{\hat{p}}, \quad (2)$$

where $\hat{p}_j(.)$ is the nonparametric estimator of $p_j(.)$ and $\hat{p} = n_1/n$, where n_1 is the sample size of the treated.

Define $\beta = (\beta_1, \beta_2, \ldots, \beta_J)'$ and $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_J)'$. Correspondingly, let $\hat{\beta} = \left(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_J\right)'$ and $\hat{\gamma} = (\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_J)'$. We make the following high-level assumption for $\hat{\beta}$ and $\hat{\gamma}$.

Assumption A.2

(i) Suppose that

$$\sqrt{n}\left(\hat{\beta} - \beta\right) = n^{-1/2}\sum_{i=1}^{n}\psi_i + o_p(1),$$

where $\psi_i = (\psi_{1i}, \psi_{2i}, \ldots, \psi_{Ji})$ is a $J \times 1$ random vector with $\mathbb{E}(\psi_{ji}) = 0$ and $\mathbb{E}(\psi_{ji}^2) < \infty$, $j = 1, \ldots, J$.

(ii) Suppose that

$$\sqrt{n}\left(\hat{\gamma} - \gamma\right) = n^{-1/2}\sum_{i=1}^{n}\phi_i + o_p(1),$$

where $\phi_i = (\phi_{1i}, \phi_{2i}, \ldots, \phi_{Ji})$ is a $J \times 1$ random vector with $\mathbb{E}(\phi_{ji}) = 0$ and $\mathbb{E}(\phi_{ji}^2) < \infty$, $j = 1, \ldots, J$.

A.2 is a high-level assumption requiring both $\hat{\beta}$ and $\hat{\gamma}$ to be asymptotically normal with the influence functions ψ_i and ϕ_i, respectively, that is,

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow{d} N\left(0, V_\beta\right) \text{ and } \sqrt{n}\left(\hat{\gamma} - \gamma\right) \xrightarrow{d} N\left(0, V_\gamma\right),$$

where V_β and V_γ are the asymptotic variances: $V_\beta = \mathbb{E}\left(\psi_i \psi_i'\right)$ and $V_\gamma = \mathbb{E}\left(\phi_i \phi_i'\right)$. A.2 can be satisfied under primitive conditions. Hahn (1998), HIR, and Imbens, Newey, and Ridder (2007) discussed those primitive conditions on their estimators. To save space, we omit the details. For both the imputation estimators and PSW estimators, ψ_i and ϕ_i take the following forms:

$$\psi_{ji} = \frac{D_i\left(Y_i - m_{1j}\left(\mathbf{X}_{ji}\right)\right)}{p_j\left(\mathbf{X}_{ji}\right)} - \frac{\left(1 - D_i\right)\left(Y_i - m_{0j}\left(\mathbf{X}_{ji}\right)\right)}{1 - p_j\left(\mathbf{X}_{ji}\right)} + \left[m_{1j}\left(\mathbf{X}_{ji}\right) - m_{0j}\left(\mathbf{X}_{ji}\right) - \beta_j\right] \text{ and}$$

$$\phi_{ji} = \frac{D_i}{p}\left(Y_i - m_{1j}\left(\mathbf{X}_{ji}\right)\right) - \frac{\left(1 - D_i\right) \cdot p_j\left(\mathbf{X}_{ji}\right)}{p \cdot \left(1 - p_j\left(\mathbf{X}_{ji}\right)\right)}\left(Y - m_{0j}\left(\mathbf{X}_{ji}\right)\right) + \frac{D_i}{p}\left[m_{1j}\left(\mathbf{X}_{ji}\right) - m_{0j}\left(\mathbf{X}_{ji}\right) - \gamma_j\right]$$

for $j = 1, \ldots, J$. Given the data, they are readily estimated as

$$\hat{\psi}_{ji} = \frac{D_i\left(Y_i - \hat{m}_{1j}\left(\mathbf{X}_{ji}\right)\right)}{\hat{p}_j\left(\mathbf{X}_{ji}\right)} - \frac{\left(1 - D_i\right)\left(Y_i - \hat{m}_{0j}\left(\mathbf{X}_{ji}\right)\right)}{1 - \hat{p}_j\left(\mathbf{X}_{ji}\right)} + \left[\hat{m}_{1j}\left(\mathbf{X}_{ji}\right) - \hat{m}_{0j}\left(\mathbf{X}_{ji}\right) - \hat{\beta}_j\right] \text{ and}$$

$$\hat{\phi}_{ji} = \frac{D_i}{\hat{p}}\left(Y_i - \hat{m}_{1j}\left(\mathbf{X}_{ji}\right)\right) - \frac{\left(1 - D_i\right) \cdot \hat{p}_j\left(\mathbf{X}_{ji}\right)}{\hat{p} \cdot \left(1 - \hat{p}_j\left(\mathbf{X}_{ji}\right)\right)}\left(Y - \hat{m}_{0j}\left(\mathbf{X}_{ji}\right)\right) + \frac{D_i}{\hat{p}}\left[\hat{m}_{1j}\left(\mathbf{X}_{ji}\right) - \hat{m}_{0j}\left(\mathbf{X}_{ji}\right) - \hat{\gamma}_j\right].$$

Accordingly, V_β and V_γ are estimated as

$$\hat{V}_\beta = \frac{1}{n}\sum_{i=1}^{n}\hat{\psi}_i\hat{\psi}_i' \text{ and } \hat{V}_\gamma = \frac{1}{n}\sum_{i=1}^{n}\hat{\phi}_i\hat{\phi}_i', \quad (3)$$

where $\hat{\psi}_i = (\hat{\psi}_{1i}, \hat{\psi}_{2i}, \ldots, \hat{\psi}_{Ji})$ and $\hat{\phi}_i = (\hat{\phi}_{1i}, \hat{\phi}_{2i}, \ldots, \hat{\phi}_{Ji})$. We can show that $\hat{V}_\beta$ and $\hat{V}_\gamma$ are consistent estimators of V_β and V_γ, respectively.

For a given $\mathbf{X}_j$, it is desirable to use efficient estimators, such as the imputation estimators and the PSW estimators. However, our CSC also allows inefficient estimators to be used as long as they are asymptotically normal with the convergence rate of $\sqrt{n}$.

3. CSC FOR VALID COVARIATES

In this section, we provide a CSC for the simple case where all the candidate sets of covariates satisfy the unconfoundedness assumption. Following Dawid (1979), we use $\mathcal{X} \perp \mathcal{Y} \mid \mathcal{Z}$ to denote the independence of $\mathcal{X}$ and $\mathcal{Y}$ given $\mathcal{Z}$.

Assumption A.3. (i) $D \perp (Y_0, Y_1) | \mathbf{X}_j$ and (ii) $0 < p\left(\mathbf{X}_j\right) < 1$ for $j = 1, \ldots, J$.

A.3(i) requires that conditioning on each set of covariates, the key unconfoundedness assumption holds. It is certainly a strong assumption and we will relax it in the next section. Lu and White (2014, sec. 5) discussed how to find multiple sets of valid candidate covariates that satisfy A.3(i).

A.3(ii) is the overlap assumption that is commonly made in the treatment effect literature. Note that if $\{\mathbf{X}_j\}_{j=1}^J$ is nested such that $\mathbf{X}_1 \subseteq \mathbf{X}_2 \cdots \subseteq \mathbf{X}_J$, we only require $0 < p(\mathbf{X}_J) < 1$, which implies A.3(ii) by the law of iterated expectations. If the overlap assumption does not hold, the asymptotic variance terms, V_β and V_γ, can be infinite, as shown in Khan and Tamer (2010). Crump et al. (2009) develop a systematic method to address the lack of overlap by focusing on a subpopulation. In principle, if A.3(ii) is violated, we can also apply our methods to a proper subpopulation. For simplicity, however, we maintain this overlap assumption.

Given that we have J sets of covariates that satisfy the unconfoundedness assumption, we have J consistent estimators of β_0 and γ_0. Hence, we want to select the optimal estimator or combine all estimators optimally. Let $w = (w_1, w_2, \ldots, w_J)'$ be a $J \times 1$ weight vector such that $\sum_{j=1}^J w_j = 1$. We propose the following average estimators of β_0 and γ_0:

$$\hat{\beta}(w) = \sum_{j=1}^J w_j \hat{\beta}_j \quad \text{and} \quad \hat{\gamma}(w) = \sum_{j=1}^J w_j \hat{\gamma}_j.$$

We can impose various restrictions on w. For example, if we allow only one element of w to be 1 and set the other elements to 0, then we are selecting a single estimator. Therefore, the single best estimator is a special case of the average estimator. Another common restriction is that the weights are all positive, that is, $0 \le w_j \le 1$. This restriction is often imposed in the model averaging literature (see, e.g., Hansen 2007; Hansen and Racine 2012).

To obtain $\hat{\beta}(w)$ and $\hat{\gamma}(w)$, the key step is to determine the weight w. We choose a weight w that minimizes the AMSEs. It is easy to see that

$$\sqrt{n}\left(\hat{\beta}(w) - \beta_0\right) \xrightarrow{d} N\left(0, w' V_\beta w\right)$$

and

$$\sqrt{n}\left(\hat{\gamma}(w) - \gamma_0\right) \xrightarrow{d} N\left(0, w' V_\gamma w\right).$$

Thus, in this case

$$\mathrm{AMSE}_\beta(w) = w' V_\beta w \text{ and } \mathrm{AMSE}_\gamma(w) = w' V_\gamma w.$$

We propose the following CSC for β_0 and γ_0:

$$\mathrm{CSC}_\beta(w) = w' \hat{V}_\beta w \quad \text{and} \quad \mathrm{CSC}_\gamma(w) = w' \hat{V}_\gamma w.$$

If we do not impose any restrictions on w other than $\sum_{j=1}^J w_j = 1$, it is easy to show that the estimated optimal weights are

$$\hat{w}_\beta^* \equiv \arg\min_w \left(\mathrm{CSC}_\beta(w)\right) = \left(\mathcal{I}' \hat{V}_\beta^{-1} \mathcal{I}\right)^{-1} \mathcal{I}' \hat{V}_\beta^{-1}, \quad (4)$$

and

$$\hat{w}_\gamma^* \equiv \arg\min_w \left(\mathrm{CSC}_\gamma(w)\right) = \left(\mathcal{I}' \hat{V}_\gamma^{-1} \mathcal{I}\right)^{-1} \mathcal{I}' \hat{V}_\gamma^{-1}, \quad (5)$$

where $\mathcal{I}$ is the $J \times 1$ vector of ones. Thus, the average estimators of β_0 and γ_0 are respectively,

$$\hat{\beta}^* \equiv \hat{\beta}\left(\hat{w}_\beta^*\right) = \hat{w}_\beta^{*\prime} \hat{\beta}$$

and

$$\hat{\gamma}^* \equiv \hat{\gamma}\left(\hat{w}_\gamma^*\right) = \hat{w}_\gamma^{*\prime} \hat{\gamma}.$$

It is easy to see that $\hat{w}_\beta^* \xrightarrow{p} \arg\min\left(\mathrm{AMSE}_\beta(w)\right)$ and $\hat{w}_\gamma \xrightarrow{p} \arg\min\left(\mathrm{AMSE}_\gamma(w)\right)$ and the average estimators satisfy

$$\sqrt{n}\left(\hat{\beta}^* - \beta_0\right) \xrightarrow{d} N\left(0, \left(\mathcal{I}' V_\beta^{-1} \mathcal{I}\right)^{-1}\right)$$

and

$$\sqrt{n}\left(\hat{\gamma}^* - \gamma_0\right) \xrightarrow{d} N\left(0, \left(\mathcal{I}' V_\gamma^{-1} \mathcal{I}\right)^{-1}\right).$$

The asymptotic variances can be estimated by $(\mathcal{I}' \hat{V}_\beta^{-1} \mathcal{I})^{-1}$ and $(\mathcal{I}' \hat{V}_\gamma^{-1} \mathcal{I})^{-1}$. We can construct the confidence interval for $\hat{\beta}^*$ and $\hat{\gamma}^*$ based on the asymptotic results above directly. Our average estimators $\hat{\beta}^*$ and $\hat{\gamma}^*$ are similar to the FOGLeSs estimators proposed by Lu and White (2014). They considered combining estimators for core regression coefficients in the linear regression context, whereas here we consider averaging semiparametric estimators of ATE and ATT.

Remark 1. In this simple case, to select the single best estimator among $\hat{\beta}$ (or $\hat{\gamma}$), we can simply choose the jth element of $\hat{\beta}$ (or $\hat{\gamma}$), such that the jth diagonal element of $\hat{V}_\beta$ (or $\hat{V}_\gamma$) is smallest.

Remark 2. To ensure the weights are all positive, we can minimize $\mathrm{CSC}_\beta(w)$ or $\mathrm{CSC}_\gamma(w)$ with the restrictions that $\sum_{j=1}^J w_j = 1$ and $0 \le w_j \le 1$, $j = 1, \ldots, J$. This can be implemented using quadratic programming in standard software such as Matlab.

Remark 3. An alternative way of averaging the estimators in the literature is to use smoothed averaging. Specifically, let $\hat{V}_{\beta,j}$ and $\hat{V}_{\gamma,j}$ be the jth diagonal element of $\hat{V}_\beta$ and $\hat{V}_\gamma$, respectively. Then the smoothed average estimators of β_0 and γ_0 are respectively

$$\sum_{j=1}^J \frac{\exp\left(-\frac{1}{2}\hat{V}_{\beta,j}\right)}{\sum_{k=1}^J \exp\left(-\frac{1}{2}\hat{V}_{\beta,k}\right)} \hat{\beta}_j$$

and

$$\sum_{j=1}^J \frac{\exp\left(-\frac{1}{2}\hat{V}_{\gamma,j}\right)}{\sum_{k=1}^J \exp\left(-\frac{1}{2}\hat{V}_{\gamma,k}\right)} \hat{\gamma}_j,$$

which can be calculated easily.

Remark 4. Here there is no natural ordering of $\{\mathbf{X}_j\}_{j=1}^J$, hence we cannot order the estimators $\{\hat{\beta}_j\}_{j=1}^J$ and $\{\hat{\gamma}_j\}_{j=1}^J$ according to their asymptotic variances easily. Even if $\{\mathbf{X}_j\}_{j=1}^J$ is nested

and thus ordered, ranking the estimators is complicated, as it is not always more efficient to condition on more covariates. As pointed out by Hahn (2004) and White and Lu (2011), sometimes using a smaller set of covariates may result in a more efficient estimator. Whether a large or a small set of covariates should be used depends on the underlying structure of $(D, Y(0), Y(1), \mathbf{X}_1, \ldots, \mathbf{X}_J)$, which may be difficult to judge in practice. For further discussion on this issue in the linear regression context, see Lu and White (2014).

Remark 5. Here we propose to combine estimators. One natural question is whether we can combine covariates. For example, we could use the union of $\{\mathbf{X}_j\}_{j=1}^J$ as one set of covariates to estimate β_0 and γ_0. However, this is not a good strategy for two reasons. First, A.3 does not imply that D and (Y_0, Y_1) are independent conditioning on the union of $\{\mathbf{X}_j\}_{j=1}^J$. It is easy to find counter-examples. In fact, Heckman and Navarro-Lozano (2004) and Wooldridge (2005) explicitly warned that conditioning on a large set of covariates, the unconfoundedness assumption can be violated. Second, even if the unconfoundedness assumption is satisfied for the union of $\{\mathbf{X}_j\}_{j=1}^J$, the efficiency of the estimators may suffer due to the use of a large set of covariates as discussed in Remark 4 above.

4. CSC FOR LOCALLY INVALID COVARIATES

In this section, we consider the case where some sets of candidate covariates violate the unconfoundedness assumption locally. We work on a local asymptotic framework, thus it is necessary to introduce a triangular array of the data.

Assumption A.1′. Let $\{D_i, Y_i, \mathbf{X}_{1ni}, \mathbf{X}_{2ni}, \ldots, \mathbf{X}_{Jni},\ 1 \le i \le n,\ n = 1, 2, \ldots\}$ be the a triangular array of random vectors and for a given n, $\{D_i, Y_i, \mathbf{X}_{1ni}, \mathbf{X}_{2ni}, \ldots, \mathbf{X}_{Jni}\}$, $i = 1, \ldots, n$ are iid.

We assume that the treatment D_i, the potential outcomes $Y_i(0)$ and $Y_i(1)$, and thus the observed outcome Y_i do not depend on the sample size n. Thus the parameters of interest β_0 (ATE) and γ_0 (ATT) do not change with n. Define

$$\beta_{jn} \equiv \mathbb{E}\left[m_{1jn}\left(\mathbf{X}_{jni}\right) - m_{0jn}\left(\mathbf{X}_{jni}\right)\right]$$

and

$$\gamma_{jn} \equiv \mathbb{E}\left[m_{1jn}\left(\mathbf{X}_{jni}\right) - m_{0jn}\left(\mathbf{X}_{jni}\right) \mid D_i = 1\right],$$

where

$$m_{1jn}\left(\mathbf{X}_{jni}\right) = \mathbb{E}\left(Y_i \mid D_i = 1, \mathbf{X}_{jni}\right)$$

and

$$m_{0jn}\left(\mathbf{X}_{jni}\right) = \mathbb{E}\left(Y_i \mid D_i = 0, \mathbf{X}_{jni}\right).$$

For notational simplicity, we omit the i and n subscripts as long as there is no confusion.

Assumption A.4. Suppose that there is a known integer J_0 such that $1 \le J_0 < J$.

(i) For $j = 1, \ldots, J_0$, $D_i \perp (Y_i(0), Y_i(1)) | \mathbf{X}_{jni}$.

(ii) For $j = J_0 + 1, \ldots, J$, there exist unknown constants δ_j and λ_j that do not depend on n such that $\beta_{jn} = \beta_0 + \frac{\delta_j}{\sqrt{n}} + o(n^{-1/2})$ and $\gamma_{jn} = \gamma_0 + \frac{\lambda_j}{\sqrt{n}} + o\left(n^{-1/2}\right)$.

(iii) For $j = 1, \ldots, J$, $0 < p_{jn}(\mathbf{X}_{jni}) < 1$, where $p_{jn}(\mathbf{X}_{jni}) = \mathbb{E}(D_i | \mathbf{X}_{jni})$.

A.4(i) requires at least one known set of covariates to be valid. This allows us to estimate the bias term of the estimators based on the locally invalid covariates. Note that there is no restriction on how to construct the J_0 sets of valid covariates. For example, one such set can be the union of all available covariates or a small subset containing the most relevant covariates. To determine such J_0 sets of valid covariates, unavoidably we need to make some subjective judgments using knowledge of the underlying structures. If all sets of covariates are (locally or globally) invalid, then none of the standardized individual estimators has an asymptotic distribution centered at 0. This renders statistical inference impossible. Note that in the traditional way of estimating ATE and ATT using a single set of covariates, subjective judgment is also required.

Admittedly, the assumption that at least one set of covariates is known to be valid is the major limitation of our approach. However, assumptions like A.4(i) are often made in the model averaging literature. For example, DiTraglia (2012) assumed that there are enough known valid instruments to identify the parameter of interest in his average instrumental variable estimator. Similarly, CH and Liu (2015) assumed that the largest model is correctly specified, thus all the parameters are consistently estimated with zero asymptotic bias using the largest model. All these methods involve some subjective judgments, such as choosing valid instruments or a large model that is assumed to be correctly specified.

Though it is fine to use only the J_0 sets of valid covariates, to reduce the AMSE even more, it is desirable to make further use of locally invalid covariates. A.4(ii) requires the remaining $J - J_0$ sets of covariates to be locally invalid in the sense that the differences between the parameters of interest (β_0 and γ_0) and the parameters identified (β_{jn} and γ_{jn}) converge to 0, as the sample size approaches infinity. Thus in the limit, the unconfoundedness assumption is satisfied. However, for a given n, the unconfoundedness assumption is violated. To better understand A.4(ii), note that

$$\beta_{jn} - \beta_0 = \mathbb{E}[\underbrace{\left[\mathbb{E}\left(Y_i(0) | D_i = 1, \mathbf{X}_{jni}\right) - \mathbb{E}\left(Y_i(0) | D_i = 0, \mathbf{X}_{jni}\right)\right]}_{\equiv b_{1n}(\mathbf{X}_{jni})} \mid D_i = 1\,] \cdot \mathbb{E}(D_i)$$

$$+\mathbb{E}[\underbrace{\left[\mathbb{E}\left(Y_i(1) | D_i = 1, \mathbf{X}_{jni}\right) - \mathbb{E}\left(Y_i(1) | D_i = 0, \mathbf{X}_{jni}\right)\right]}_{\equiv b_{0n}(\mathbf{X}_{jni})} \mid D_i = 0\,] \cdot (1 - \mathbb{E}(D_i))$$

$$\gamma_{jn} - \gamma_0 = \mathbb{E}[\underbrace{\left[\mathbb{E}\left(Y_i(0) | D_i = 1, \mathbf{X}_{jni}\right) - \mathbb{E}\left(Y_i(0) | D_i = 0, \mathbf{X}_{jni}\right)\right]}_{\equiv b_{1n}(\mathbf{X}_{jni})} \mid D_i = 1].$$

White and Chalak (2013) used $b_{1n}\left(\mathbf{X}_{jni}\right)$ and $b_{0n}\left(\mathbf{X}_{jni}\right)$ to measure the degrees of departure from the unconfoundedness for $Y_i(0)$ and $Y_i(1)$ at $\mathbf{X}_{jni}$, respectively. $b_{1n}\left(\mathbf{X}_{jni}\right)$ and $b_{0n}\left(\mathbf{X}_{jni}\right)$ are both zero when the unconfoundedness assumption is satisfied. Essentially, here we require $b_{1n}\left(\mathbf{X}_{jni}\right)$ and $b_{0n}\left(\mathbf{X}_{jni}\right)$ to converge to 0 at the rate of $\sqrt{n}$. One intuitive way of understanding condition A.4(ii) is to think of the locally invalid covariates as being generated as

$$\begin{pmatrix} \mathbf{X}_{(J_0+1)ni} \\ \cdots \\ \mathbf{X}_{Jni} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^*_{(J_0+1)i} \\ \cdots \\ \mathbf{X}^*_{Ji} \end{pmatrix} + \frac{1}{\sqrt{n}} \begin{pmatrix} \Delta x_{(J_0+1)i} \\ \cdots \\ \Delta x_{Ji} \end{pmatrix} + o_p\left(n^{-1/2}\right),$$

where $\mathbf{X}^*_{(J_0+1)i}, \ldots, \mathbf{X}^*_{Ji}$ are some random variables satisfying $(Y_i(0), Y_i(1)) \perp D_i \mid \mathbf{X}^*_{ji}$ for $j = J_0+1, \ldots, J$ and $\Delta x_{(J_0+1)i}, \ldots, \Delta x_{Ji}$ are some other random variables.

A.4(ii) does not literally mean that the real-world data are generated in this way. Instead, as a device, it ensures that the squared-bias term of the estimators is of the same order as the variance term (n^{-1}), which allows us to trade-off the bias and variance terms. This can be thought of as a mimic of the bias-variance trade-off in finite samples. If the unconfoundedness assumption is violated globally in the sense that neither $(\beta_{jn} - \beta_0)$ nor $\left(\gamma_{jn} - \gamma_0\right)$ converges to zero as n approaches infinity, then the bias term will dominate and the globally biased estimators should never be used. This "local asymptotic" approach is commonly used in the model averaging literature (e.g., CH and Liu 2015), in the study of weak instrument variables (e.g., Staiger and Stock 1997), and in the study of local alternatives in hypothesis testing in terms of Pitman drift. A.4(iii) is the overlap assumption discussed in Section 3.

The next lemma shows the behavior of the individual estimators under A.4.

Lemma 1. Suppose that A.1′, A.2, and A.4 hold. Then,

(i) $\sqrt{n}\left(\hat{\beta}-\beta_0\right) \xrightarrow{d} \delta + \mathbf{P} \sim N\left(\delta,\ V_\beta\right)$, where $\mathbf{P}$ is a $J \times 1$ random normal vector with mean zero and covariance–variance matrix V_β and δ is a $J \times 1$ vector: $\delta = \left(0, \ldots, 0, \delta_{J_0+1}, \ldots, \delta_J\right)'$.

(ii) $\sqrt{n}\left(\hat{\gamma}-\gamma_0\right) \xrightarrow{d} \lambda + \mathbf{Q} \sim N\left(\lambda,\ V_\gamma\right)$, where $\mathbf{Q}$ is a $J \times 1$ random normal vector with mean zero and covariance–variance matrix V_γ and λ is a $J \times 1$ vector: $\lambda = (0, \ldots, 0, \lambda_{J_0+1}, \ldots, \lambda_J)'$.

We consider the average estimators

$$\hat{\beta}(w) = \sum_{j=1}^{J} w_j \hat{\beta}_j \quad \text{and} \quad \hat{\gamma}(w) = \sum_{j=1}^{J} w_j \hat{\gamma}_j,$$

which include the selected single best estimator as a special case. It is easy to see that the AMSEs of $\hat{\beta}(w)$ and $\hat{\gamma}(w)$ are, respectively,

$$\mathrm{AMSE}_\beta(w) = w'\left(\delta\delta' + V_\beta\right) w$$

and

$$\mathrm{AMSE}_\gamma(w) = w'\left(\lambda\lambda' + V_\gamma\right) w.$$

To derive our CSC, we need to estimate AMSE_β and AMSE_γ. Note that V_β and V_γ can be estimated consistently, but it turns out that the bias terms δ and λ cannot. For simplicity, we assume that $J_0 = 1$, that is, only the first set of covariates is valid. For the general case of $J_0 > 1$, see Remark 8. Then the natural estimator of the bias term δ_j, $j = 2, \ldots, J$, is $\hat{\delta}_j = \sqrt{n}\left(\hat{\beta}_j - \hat{\beta}_1\right)$, that is, the estimator of δ is

$$\hat{\delta} = \left(0, \hat{\delta}_2, \ldots, \hat{\delta}_J\right)' = \sqrt{n} S \cdot \hat{\beta},$$

where S is a known $J \times J$ matrix:

$$S = \begin{bmatrix} 0 & 0 & 0 & \ldots & 0 \\ -1 & 1 & 0 & \ldots & 0 \\ -1 & 0 & 1 & \ldots & 0 \\ & \cdot & \cdot & \cdot & \\ -1 & 0 & \ldots & 0 & 1 \end{bmatrix}.$$

Similarly, the estimator of λ is

$$\hat{\lambda} = \left(0, \hat{\lambda}_2, \ldots, \hat{\lambda}_J\right)' = \sqrt{n} S \cdot \hat{\gamma},$$

where $\hat{\lambda}_j = \sqrt{n}\left(\hat{\gamma}_j - \hat{\gamma}_1\right)$.

The next proposition describes the behavior of the bias estimators.

Proposition 2. Suppose that A.1′, A.2, and A.4 hold. Then

(i) $\hat{\delta} \xrightarrow{d} \delta + S\mathbf{P} \sim N\left(\delta,\ SV_\beta S'\right)$, where $\mathbf{P}$ is defined as in Lemma 1(i).

(ii) $\hat{\lambda} \xrightarrow{d} \lambda + S\mathbf{Q} \sim N\left(\lambda,\ SV_\gamma S'\right)$, where $\mathbf{Q}$ is defined as in Lemma 1(ii).

Proposition 2 states that the estimator of the bias term $\hat{\delta}$ (or $\hat{\lambda}$) converges in distribution to a normal vector with the mean of δ (or λ). Therefore $\hat{\delta}$ (or $\hat{\lambda}$) is not a consistent estimator of δ (or λ). It is tempting to estimate AMSE_β and AMSE_γ respectively by their sample analogues:

$$w'\left(\hat{\delta}\hat{\delta}' + \hat{V}_\beta\right) w \text{ and } w'\left(\hat{\lambda}\hat{\lambda}' + \hat{V}_\gamma\right) w.$$

However, note that $\hat{\delta}\hat{\delta}' \xrightarrow{d} (\delta + S\mathbf{P})(\delta + S\mathbf{P})'$. The expected value of $(\delta + S\mathbf{P})(\delta + S\mathbf{P})'$ is $\delta\delta' + SV_\beta S'$, and so $\delta\delta'$ should be estimated by $\hat{\delta}\hat{\delta}' - S\hat{V}_\beta S'$. Thus we propose the following CSC for β_0:

$$\mathrm{CSC}_\beta(w) = w'\left(\hat{\delta}\hat{\delta}' - S\hat{V}_\beta S' + \hat{V}_\beta\right) w.$$

Similarly, the CSC for γ_0 is

$$\mathrm{CSC}_\gamma(w) = w'\left(\hat{\lambda}\hat{\lambda}' - S\hat{V}_\gamma S' + \hat{V}_\gamma\right) w.$$

Note that $\hat{\delta}\hat{\delta}' - S\hat{V}_\beta S' + \hat{V}_\beta$ is not guaranteed to be semipositive definite. Thus the solution to the unconstrained minimization problem may not exist. To solve the problem, one simple way is to impose restrictions on the weights such as $0 \le w_j \le 1$. Define

$$\mathbb{W} = \left\{ w : \sum_{j=1}^{J} w_j = 1 \text{ and } 0 \le w_j \le 1 \right\}.$$

Thus, the optimal weights are defined as

$$\hat{w}^*_\beta = \arg\min_{w \in \mathbb{W}} \mathrm{CSC}_\beta(w) \quad \text{and} \quad \hat{w}^*_\gamma = \arg\min_{w \in \mathbb{W}} \mathrm{CSC}_\gamma(w).$$

The next proposition studies the properties of our CSC.

Proposition 3. Suppose that A.1′, A.2, and A.4 hold. Then

(i) $\mathrm{CSC}_\beta(w) \xrightarrow{d} w'\Omega_\beta w$, where $\Omega_\beta = S\cdot((\delta+\mathbf{P})\cdot(\delta+\mathbf{P})' - V_\beta)\cdot S' + V_\beta$. Further, $\mathbb{E}(w'\Omega_\beta w) = \mathrm{AMSE}_\beta(w)$, and $\hat{w}_\beta^* \xrightarrow{d} w_\beta^* = \arg\min_{w\in\mathbb{W}}(w'\Omega_\beta w)$.

(ii) $\mathrm{CSC}_\gamma(w) \xrightarrow{d} w'\Omega_\gamma w$, where $\Omega_\gamma = S\cdot((\lambda+\mathbf{Q})\cdot(\lambda+\mathbf{Q})' - V_\gamma)\cdot S' + V_\gamma$. Further, $\mathbb{E}(w'\Omega_\gamma w) = \mathrm{AMSE}_\gamma(w)$, and $\hat{w}_\gamma^* \xrightarrow{d} w_\gamma^* = \arg\min_{w\in\mathbb{W}}(w'\Omega_\gamma w)$.

Hence, our average estimators are defined as

$$\hat{\beta}^* = \hat{\beta}\left(\hat{w}_\beta^*\right) = \hat{w}_\beta^{*\prime}\hat{\beta} \text{ and } \hat{\gamma}^* = \hat{\gamma}\left(\hat{w}_\gamma^*\right) = \hat{w}_\gamma^{*\prime}\hat{\gamma}.$$

The next proposition shows the asymptotic distribution of the average estimators.

Proposition 4. Suppose that A.1′, A.2, and A.4 hold. Then

$$\sqrt{n}\left(\hat{\beta}^* - \beta_0\right) \xrightarrow{d} w_\beta^{*\prime}\cdot[\delta+\mathbf{P}]$$

and

$$\sqrt{n}\left(\hat{\gamma}^* - \gamma_0\right) \xrightarrow{d} w_\gamma^{*\prime}\cdot[\lambda+\mathbf{Q}].$$

The asymptotic distribution of $\hat{\beta}^*$ (or $\hat{\gamma}^*$) is nonstandard and is a complicated nonlinear function of a normal vector $\mathbf{P}$ (or $\mathbf{Q}$), as w_β^* (or w_γ^*) also depends on $\mathbf{P}$ (or $\mathbf{Q}$). To obtain the confidence interval, we propose to use the simulation-based method adopted in CH. For brevity, we only describe the method for $\hat{\beta}^*$. The method for $\hat{\gamma}^*$ is similar. For any given value of δ, we propose the following algorithm to obtain the $100(1-\alpha)\%$ confidence interval for $\hat{\beta}^*$.

Algorithm 1 (Confidence interval of $\hat{\beta}^*$ for a given δ). Simulate a sufficiently large number (say K) of $J\times 1$ normal vectors $\mathbf{P}^{(k)} \sim N\left(0, \hat{V}_\beta\right)$ $(k=1,\ldots,K)$, and for each k, construct the weights

$$\hat{w}_\beta^{*(k)}(\delta) = \arg\min_{w\in\mathbb{W}} w'\left(S\cdot\left(\left(\delta+\mathbf{P}^{(k)}\right)\cdot\left(\delta+\mathbf{P}^{(k)}\right)' - \hat{V}_\beta\right)\cdot S' + \hat{V}_\beta\right)w,\ k=1,\ldots,K.$$

Let $\hat{a}_\beta(\delta)$ and $\hat{b}_\beta(\delta)$ be the $(\alpha/2)$th and $(1-\alpha/2)$ th quantiles of $\hat{w}_\beta^{*(k)}(\delta)\cdot\left[\delta+\mathbf{P}^{(k)}\right]$, respectively $(k=1,\ldots,K)$. Then the confidence interval is constructed as

$$\widehat{CI}_\beta(\delta) = \left[\hat{\beta}^* - \frac{1}{\sqrt{n}}\hat{b}_\beta(\delta),\ \hat{\beta}^* - \frac{1}{\sqrt{n}}\hat{a}_\beta(\delta)\right].$$

The algorithm is easy to implement given the value of δ. However, we do not know the value of δ. One simple way to overcome this is to use the estimated $\hat{\delta}$. The confidence interval is thus constructed as

$$\widehat{CI}_\beta\left(\hat{\delta}\right) = \left[\hat{\beta}^* - \frac{1}{\sqrt{n}}\hat{b}_\beta\left(\hat{\delta}\right),\ \hat{\beta}^* - \frac{1}{\sqrt{n}}\hat{a}_\beta\left(\hat{\delta}\right)\right]. \quad (6)$$

Alternatively, we can use the two-stage procedure proposed by CH (p. 210). To do so, we first need to construct the confidence regions of the estimators of the bias terms. Specifically, let $\tilde{S}$ be the known $(J-1)\times J$ matrix:

$$\tilde{S} \equiv \begin{bmatrix} -1 & 1 & 0 & \ldots & 0 \\ -1 & 0 & 1 & \ldots & 0 \\ & \cdot & \cdot & \cdot & \\ -1 & 0 & \ldots & 0 & 1 \end{bmatrix}. \quad (7)$$

Define $\tilde{\delta} = (\delta_2, \delta_3, \ldots, \delta_J)'$ and $\widehat{\tilde{\delta}} = (\hat{\delta}_2, \hat{\delta}_3, \ldots, \hat{\delta}_J)'$. Then it is easy to show that the Wald-statistic for testing the null that $\tilde{\delta} = \tilde{\delta}^*$ is

$$F_{\tilde{\delta}}\left(\tilde{\delta}^*\right) \equiv \left(\widehat{\tilde{\delta}} - \tilde{\delta}^*\right)'\cdot\left(\tilde{S}\hat{V}_\beta\tilde{S}'\right)^{-1}\cdot\left(\widehat{\tilde{\delta}} - \tilde{\delta}^*\right).$$

Under the null that $\tilde{\delta} = \tilde{\delta}^*$, $F_{\tilde{\delta}}(\tilde{\delta}^*)$ converges to the χ^2_{J-1} distribution. Thus the $100(1-\tau)\%$ confidence set of $\tilde{\delta}$ can be constructed by inverting the test statistic, that is,

$$\hat{A}_{\tilde{\delta}} = \left(\tilde{\delta}^* : F_{\tilde{\delta}}\left(\tilde{\delta}^*\right) \le \chi^2_{J-1,1-\tau}\right), \quad (8)$$

where $\chi^2_{J-1,1-\tau}$ is the $(1-\tau)$th quantile of the χ^2_{J-1} distribution. We propose the following two-stage algorithm to obtain the confidence interval of $\hat{\beta}^*$.

Algorithm 2 (Two-stage confidence interval of $\hat{\beta}^*$).

Step 1: Find the $100(1-\tau)\%$ confidence set of $\tilde{\delta}$: $\hat{A}_{\tilde{\delta}}$ as in Equation (8).

Step 2: For each $\tilde{\delta}^* \in \hat{A}_{\tilde{\delta}}$, construct the $(\alpha/2)$th and $(1-\alpha/2)$th quantiles of $w_\beta^*([0\ \tilde{\delta}^{*\prime}]')'\times[[0\ \tilde{\delta}^{*\prime}]'+\mathbf{P}]$: $\hat{a}_\beta([0\ \tilde{\delta}^{*\prime}]')$ and $\hat{b}_\beta([0\ \tilde{\delta}^{*\prime}]')$ as in Algorithm 1.

Step 3: Construct the confidence interval as

$$\widehat{CI}_\beta = \left[\hat{\beta}^* - \frac{1}{\sqrt{n}}\max_{\tilde{\delta}^*\in\hat{A}_{\tilde{\delta}}}\left(\hat{b}_\beta\left(\left[0\ \tilde{\delta}^{*\prime}\right]'\right)\right),\ \hat{\beta}^* - \frac{1}{\sqrt{n}}\min_{\tilde{\delta}^*\in\hat{A}_{\tilde{\delta}}}\left(\hat{a}_\beta\left(\left[0\ \tilde{\delta}^{*\prime}\right]'\right)\right)\right].$$

The next proposition shows that the two-stage confidence interval has valid coverage properties.

Proposition 5. Suppose that A.1′, A.2, and A.4 hold. Then

$$\lim_{n\to\infty} P\left(\beta_0 \in \widehat{\mathrm{CI}}_\beta\right) \ge 1-\alpha-\tau.$$

Proposition 5 establishes the validity of the two-stage confidence interval for $\hat{\beta}^*$. However, there are two problems associated with it. First, it can be conservative in the sense that the actual coverage can be larger than the nominal coverage. Second, the computation can be intensive, as it requires calculating $\hat{a}_\beta([0\ \tilde{\delta}^{*\prime}]')$ and $\hat{b}_\beta([0\ \tilde{\delta}^{*\prime}]')$ for each $\tilde{\delta}^*$ in the confidence region $\hat{A}_{\tilde{\delta}}$. In the simulation, we show that the naive approach (Equation (6)) actually performs remarkably well in finite samples. DiTraglia (2012) also reported the excellent performance of the naive approach in other contexts. The confidence intervals for $\hat{\gamma}^*$ can similarly be constructed. Other methods, such as those proposed in Zhang and Liang (2011), can also be used to construct the confidence intervals.

Remark 6. As discussed in Section 3, the single best estimator of β_0 (γ_0) is the jth element of $\hat{\beta}$ $(\hat{\gamma})$ such that $\mathrm{CSC}_\beta(w(j))$ $(\mathrm{CSC}_\gamma(w(j)))$ is smallest, where $w(j)$ is a $J\times 1$ selection vector with the jth element being 1 and others 0. We can also construct the smoothed averaging estimators for β_0 and γ_0

Table 1. Relative efficiency: valid covariates

DGP	n	Valid					CSC average	Smoothed average	CSC selected
		Set 1	Set 2	Set 3	Set 4	Set 5			
		Average treatment effect (β_0)							
1	250	1.000	0.765	1.133	0.771	1.123	0.778	0.803	0.873
	500	1.000	0.847	1.258	0.861	1.251	0.810	0.839	0.916
	1000	1.000	0.922	1.362	0.930	1.377	0.840	0.868	0.946
	5000	1.000	0.943	1.360	0.928	1.390	0.839	0.861	0.938
2	250	1.000	0.738	1.006	0.730	0.996	0.794	0.756	0.858
	500	1.000	0.811	1.141	0.834	1.118	0.820	0.848	0.892
	1000	1.000	0.878	1.218	0.887	1.213	0.836	0.906	0.914
	5000	1.000	0.905	1.240	0.892	1.256	0.833	0.924	0.911
3	250	1.000	0.766	0.977	0.786	0.980	0.742	0.806	0.808
	500	1.000	0.838	1.072	0.825	1.045	0.797	0.860	0.860
	1000	1.000	0.886	1.112	0.884	1.114	0.821	0.878	0.881
	5000	1.000	0.908	1.139	0.897	1.130	0.847	0.886	0.890
		Average treatment effect on the treated (γ_0)							
1	250	1.000	0.614	0.994	0.619	0.983	0.650	0.679	0.733
	500	1.000	0.748	1.175	0.756	1.139	0.730	0.764	0.819
	1000	1.000	0.847	1.280	0.843	1.301	0.773	0.807	0.866
	5000	1.000	0.843	1.282	0.828	1.319	0.755	0.776	0.831
2	250	1.000	0.560	0.853	0.554	0.837	0.615	0.590	0.648
	500	1.000	0.663	1.006	0.680	0.967	0.692	0.700	0.724
	1000	1.000	0.753	1.095	0.750	1.096	0.724	0.774	0.762
	5000	1.000	0.786	1.127	0.770	1.143	0.732	0.800	0.782
3	250	1.000	0.737	0.978	0.753	0.970	0.757	0.801	0.803
	500	1.000	0.818	1.080	0.803	1.042	0.830	0.870	0.872
	1000	1.000	0.861	1.112	0.864	1.116	0.879	0.903	0.904
	5000	1.000	0.868	1.143	0.854	1.149	0.841	0.878	0.879

NOTE: The numbers in the main entries are the MSEs standardized by the MSE of the estimator based on covariate set 1.

respectively as

$$\sum_{j=1}^{J} \frac{\exp\left(-\frac{1}{2}\mathrm{CSC}_{\beta}\left(w\left(j\right)\right)\right)}{\sum_{k=1}^{J}\exp\left(-\frac{1}{2}\mathrm{CSC}_{\beta}\left(w\left(k\right)\right)\right)}\hat{\beta}_j$$

and

$$\sum_{j=1}^{J} \frac{\exp\left(-\frac{1}{2}\mathrm{CSC}_{\gamma}\left(w\left(j\right)\right)\right)}{\sum_{k=1}^{J}\exp\left(-\frac{1}{2}\mathrm{CSC}_{\gamma}\left(w\left(k\right)\right)\right)}\hat{\gamma}_j.$$

The confidence intervals can also be constructed in the way that is similar to Algorithms 1 and 2.

Remark 7. The assumption of locally invalid covariates allows us to trade-off bias and variance terms. If some candidate sets of covariates are indeed globally invalid, then their corresponding elements in the CSC will be of order n and much larger than those of locally invalid and valid sets of covariates. Therefore, in principle, to minimize the CSC, the optimal weights of the estimators based on the globally invalid covariates should be 0 in large samples. So, we can apply the same CSC to globally invalid covariates. In the simulations, we consider globally invalid covariates and find that our average estimators perform reasonably well and sometimes even outperform the estimator based on valid covariates only.

Remark 8. For simplicity, when introducing the CSC for locally invalid covariates, we assume that there is only one set of valid covariates. In practice, however, we may consider multiple sets of valid covariates. In this case, we can construct the CSC average estimator in two steps. First, optimally combine the estimators based only on the sets of valid covariates as in Section 3. Second, estimate the bias term using the average estimator in the first step and construct a similar CSC to combine all estimators. Below, we describe in detail the two steps for estimating β_0. The two-step procedure for γ_0 is similar. We use superscripts (1) and (2) to denote the first-step and the second-step, respectively.

Step 1: Let $\hat{\beta}^{(1)} \equiv (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{J_0})$ be the estimators of β_0 using the J_0 sets of valid covariates $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{J_0}$. Further, let $\hat{V}_{\beta}^{(1)}$ be the estimator of the variance-covariance matrix of $\hat{\beta}^{(1)}$. Let $\hat{w}_{\beta}^{*(1)} \equiv (\hat{w}_{\beta,1}^{*(1)}, \ldots, \hat{w}_{\beta,J_0}^{*(1)})$ be the optimal weights in the first step for estimating β_0. Then, as shown in Section 3,

$$\hat{w}_{\beta}^{*(1)} = \left(\mathcal{I}^{(1)\prime}\left(\hat{V}_{\beta}^{(1)}\right)^{-1}\mathcal{I}^{(1)}\right)^{-1}\mathcal{I}^{(1)\prime}\left(\hat{V}_{\beta}^{(1)}\right)^{-1},$$

where $\mathcal{I}^{(1)}$ is the $J_0 \times 1$ vector of ones. The corresponding optimal average estimator of β_0 is $\hat{\beta}^{*(1)} = \hat{w}_{\beta}^{*(1)\prime}\hat{\beta}^{(1)}$.

Step 2: Let $\hat{\delta}^{(2)} \equiv (0, \ldots, 0, \hat{\delta}_{J_0+1}^{(2)}, \ldots, \hat{\delta}_{J}^{(2)})'$ denote the estimator of the bias term $\delta \equiv (0, \ldots, 0, \delta_{J_0+1}, \ldots, \delta_J)'$ in the second step. We can estimate the bias term based on the first-step estimator $\hat{\beta}^{*(1)}$. That is, $\hat{\delta}_j^{(2)} = \sqrt{n}(\hat{\beta}_j - \hat{\beta}^{*(1)})$, $j = J_0 + 1, \ldots, J$. Then, $\hat{\delta}^{(2)}$ can be written succinctly as

$$\hat{\delta}^{(2)} = \sqrt{n}\hat{S}_{\beta} \cdot \hat{\beta}, \tag{9}$$

Table 2-1. Relative efficiency: locally invalid covariates—Case 1

DGP	n	Valid			Locally invalid		CSC average	Smoothed average	CSC selected
		Set 1	Set 2	Set3	Set 4	Set 5			
				Average treatment effect (β_0)					
1	250	1.000	0.765	1.133	0.933	1.626	0.818	0.850	0.885
	500	1.000	0.847	1.258	1.085	1.982	0.874	0.888	0.922
	1000	1.000	0.922	1.362	1.170	2.194	0.904	0.919	0.952
	5000	1.000	0.943	1.360	1.219	2.432	0.922	0.935	0.972
2	250	1.000	0.738	1.006	0.783	1.226	0.800	0.745	0.863
	500	1.000	0.811	1.141	0.931	1.502	0.845	0.857	0.881
	1000	1.000	0.878	1.218	0.994	1.664	0.873	0.909	0.922
	5000	1.000	0.905	1.240	1.057	1.893	0.896	0.946	0.926
3	250	1.000	0.766	0.977	0.924	1.206	0.769	0.822	0.823
	500	1.000	0.838	1.072	0.976	1.320	0.843	0.881	0.882
	1000	1.000	0.886	1.112	1.027	1.392	0.856	0.886	0.888
	5000	1.000	0.908	1.139	1.046	1.482	0.891	0.911	0.912
				Average treatment effect on the treated (γ_0)					
1	250	1.000	0.614	0.994	0.746	1.302	0.689	0.723	0.745
	500	1.000	0.748	1.175	0.910	1.557	0.786	0.803	0.821
	1000	1.000	0.847	1.280	0.995	1.735	0.831	0.855	0.883
	5000	1.000	0.843	1.282	0.986	1.851	0.829	0.844	0.862
2	250	1.000	0.560	0.853	0.602	0.978	0.633	0.600	0.671
	500	1.000	0.663	1.006	0.735	1.146	0.704	0.706	0.728
	1000	1.000	0.753	1.095	0.818	1.332	0.755	0.778	0.785
	5000	1.000	0.786	1.127	0.868	1.508	0.778	0.816	0.805
3	250	1.000	0.737	0.978	0.928	1.229	0.762	0.809	0.810
	500	1.000	0.818	1.080	0.957	1.285	0.841	0.880	0.881
	1000	1.000	0.861	1.112	0.991	1.310	0.867	0.894	0.895
	5000	1.000	0.868	1.143	0.940	1.299	0.865	0.889	0.890

NOTE: See the note to Table 1.

where $\hat{S}_\beta$ is a $J \times J$ matrix:

$$\hat{S}_\beta \equiv \begin{bmatrix} 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ & & \dots & & & & \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ -\hat{w}_{\beta,1}^{*(1)} & -\hat{w}_{\beta,2}^{*(1)} & \dots & -\hat{w}_{\beta,J_0}^{*(1)} & 1 & \dots & 0 \\ & & \dots & & & & \\ -\hat{w}_{\beta,1}^{*(1)} & -\hat{w}_{\beta,2}^{*(1)} & \dots & -\hat{w}_{\beta,J_0}^{*(1)} & 0 & \dots & 1 \end{bmatrix}.$$

Then the CSC in the second step for estimating β_0 is

$$\mathrm{CSC}_\beta^{(2)}(w) = w' \left(n \cdot \hat{S}_\beta \hat{\beta} \hat{\beta}' \hat{S}_\beta' - \hat{S}_\beta \hat{V}_\beta \hat{S}_\beta' + \hat{V}_\beta \right) w.$$

Hence the corresponding optimal weights in the second step are $\hat{w}_\beta^{*(2)} = \arg\min_{w \in \mathbb{W}} \mathrm{CSC}_\beta^{(2)}(w)$ and the eventual estimator of β_0 is $\hat{\beta}^{*(2)} = \hat{w}_\beta^{*(2)\prime} \hat{\beta}$.

The theoretical property of $\hat{\beta}^{*(2)}$ is similar to that in the case with $J_0 = 1$, given that the weights $\hat{w}_\beta^{*(1)}$ are consistent estimators of the optimal weights that minimize the AMSE of the average estimator based on the J_0 sets of valid covariates in the first step. The confidence interval of the two-step $\hat{\beta}^{*(2)}$ can also be constructed following Algorithms 1 and 2 with minor modifications. For example, for the simple method of Algorithm 1, the only difference is that in Equation (6), we replace $\hat{\delta}$ with $\hat{\delta}^{(2)}$ defined in Equation (9). In Algorithm 2, we use the confidence set of $\hat{\delta}^{(2)}$ instead of that of $\hat{\delta}$.

5. A HAUSMAN-TYPE TEST

We have provided two CSC. One requires all sets of covariate to be valid, while the other allows some sets to be locally invalid. In practice, we need to decide which CSC to use. We may use some domain knowledge to determine whether the unconfoundedness assumption is plausible (see, e.g., Lu and White 2014). Alternatively, we may test the null hypotheses

$$\text{Hypothesis 1: } \mathbb{H}_{10} : \beta_1 = \beta_2 = \cdots = \beta_J$$

and

$$\text{Hypothesis 2: } \mathbb{H}_{20} : \gamma_1 = \gamma_2 = \cdots = \gamma_J.$$

We can construct a Hausman-type test for these two hypotheses. Below, we focus on Hypothesis 1, as the test for Hypothesis 2 is similar. Let $\tilde{S}$ be as defined in Equation (7). Hypothesis 1 is equivalent to $\tilde{S}\beta = 0$. The Wald-test statistic is just

$$\mathcal{W}_\beta = n \cdot \hat{\beta}' \tilde{S}' \left[\tilde{S} \hat{V}_\beta \tilde{S}' \right]^{-1} \tilde{S} \hat{\beta}.$$

The next proposition describes the behavior of the test statistic under the null, the global alternatives and the local alternatives. The global and local alternatives for ATE are defined respectively as

$$\mathbb{H}_{1A} : \mathbb{H}_{10} \text{ is false and } \mathbb{H}_{1a} : \beta_j = \beta_1 + \frac{\delta_j}{\sqrt{n}}$$

$$\text{for } j = 2, \dots, J \text{ and at least one } \delta_j \neq 0.$$

Table 2-2. Relative efficiency: locally invalid covariates—Case 2

DGP	n	Valid			Locally invalid		CSC average	Smoothed average	CSC selected
		Set 1	Set 2	Set 3	Set 4	Set 5			
					Average treatment effect (β_0)				
1	250	1.000	0.936	1.626	0.933	1.626	0.934	0.981	0.993
	500	1.000	1.077	1.975	1.085	1.982	0.975	1.008	1.024
	1000	1.000	1.163	2.195	1.170	2.194	0.988	1.019	1.032
	5000	1.000	1.219	2.413	1.219	2.432	0.997	1.030	1.045
2	250	1.000	0.794	1.238	0.783	1.226	0.902	0.809	0.958
	500	1.000	0.914	1.527	0.931	1.502	0.935	0.945	0.982
	1000	1.000	0.985	1.673	0.994	1.664	0.945	1.003	0.976
	5000	1.000	1.060	1.892	1.057	1.893	0.967	1.075	1.003
3	250	1.000	0.905	1.204	0.924	1.206	0.899	0.952	0.952
	500	1.000	0.992	1.327	0.976	1.320	0.935	0.984	0.985
	1000	1.000	1.034	1.397	1.027	1.392	0.942	0.978	0.979
	5000	1.000	1.053	1.481	1.046	1.482	0.964	1.007	1.007
					Average treatment effect on the treated (γ_0)				
1	250	1.000	0.744	1.287	0.746	1.302	0.876	0.934	0.944
	500	1.000	0.907	1.595	0.910	1.557	0.914	0.958	0.971
	1000	1.000	0.995	1.726	0.995	1.735	0.927	0.965	0.975
	5000	1.000	0.993	1.821	0.986	1.851	0.927	0.972	0.983
2	250	1.000	0.613	0.982	0.602	0.978	0.829	0.700	0.911
	500	1.000	0.721	1.195	0.735	1.146	0.855	0.775	0.914
	1000	1.000	0.817	1.336	0.818	1.332	0.874	0.845	0.924
	5000	1.000	0.877	1.506	0.868	1.508	0.892	0.898	0.937
3	250	1.000	0.903	1.238	0.928	1.229	0.894	0.936	0.936
	500	1.000	0.977	1.309	0.957	1.285	0.934	0.961	0.962
	1000	1.000	0.989	1.308	0.991	1.310	0.936	0.965	0.965
	5000	1.000	0.953	1.287	0.940	1.299	0.927	0.953	0.953

NOTE: See the note to Table 1.

Proposition 6.

(i) Suppose that A.1 and A.2 hold. Then under $\mathbb{H}_{10}$, $\mathcal{W}_\beta \xrightarrow{d} \chi^2_{J-1}$.

(ii) Suppose that A.1 and A.2 hold. Then under $\mathbb{H}_{1A}$, $P\left(\mathcal{W}_\beta > e_n\right) \to 1$, for any nonstochastic sequence $e_n = o(n)$.

(iii) Suppose that A.1′ and A.2 hold. Then under $\mathbb{H}_{1a}$, $\mathcal{W}_\beta \xrightarrow{d} \left(\tilde{\delta} + \tilde{S}\mathbf{P}\right) \cdot \left[\tilde{S} V_\beta \tilde{S}'\right]^{-1} \cdot \left(\tilde{\delta} + \tilde{S}\mathbf{P}\right)'$, where $\tilde{\delta} = (\delta_2, \ldots, \delta_J)$ and $\mathbf{P}$ is as defined in Lemma 1(i).

The same results hold for Hypothesis 2. One practical way to decide whether to use the CSC for valid covariates or the CSC for locally invalid covariates is to test Hypotheses 1 and 2 first. If the tests do not reject, we use the CSC for valid covariates. If they do, we use the CSC for locally invalid covariates. This procedure is practical and similar to that recommended by Lu and White (2014), although it may suffer from the well-known pretest problem.

Remark 9. The above testing procedure can also help to determine J_0 in the case where J_0 might be larger than 1. For example, suppose that the first set of covariates is valid and we are uncertain about the validity of the other sets of covariates. Then, we can test the hypotheses: $\beta_1 = \beta_j$ and $\gamma_1 = \gamma_j$ for $j = 2, \ldots, J$. If we fail to reject the hypotheses for the jth set of covariates, we treat it as valid. Otherwise, we treat it as invalid. After determining $J_0 > 1$ sets of valid covariates, we can follow the two-step procedure described in Remark 8 to construct our CSC average estimator.

6. MONTE CARLO SIMULATIONS

We conduct Monte Carlo simulations to examine the finite sample performance of our method. We consider three cases: (1) all sets of covariates are valid, (2) some sets of covariates are locally invalid, and (3) some sets of covariates are globally invalid. For each case, we consider the following three data generating processes (DGPs):

$$\text{DGP 1}: \; Y(0) = 1 + U + 2\epsilon_Y \text{ and } Y(1) = 2 + U + 2\epsilon_Y,$$

$$\text{DGP 2}: \; Y(0) = \Phi(1+U+2\epsilon_Y) \text{ and } Y(1) = \Phi(2+U+2\epsilon_Y),$$

$$\text{DGP 3}: \; Y(0) = (1 + U + 2\epsilon_Y)^2 \text{ and } Y(1) = (2 + U + 2\epsilon_Y)^2,$$

where Φ is the standard normal CDF, $D = \mathbf{1}\{X^* + 2\epsilon_D > 0\}$, $U = X^* + 2\epsilon_U$, and $(X^*, \epsilon_Y, \epsilon_D, \epsilon_U)$ are independent $N(0, 1)$ random variables. In all three DGPs, X^* is the key confounding factor, as $D \perp (Y(0), Y(1)) \,|\, X^*$.

For all three cases, we consider additional covariates: $X_1 = U + 2\epsilon_1$, $X_2 = D + 2\epsilon_2$, $X_3 = U + 2\epsilon_3$, $X_4 = D + 2\epsilon_4$, where $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$ are independent $N(0, 1)$ random variables. Considering that in practice, researchers often start with a large set of covariates, we use $(X^*, X_1, X_2, X_3, X_4)$ as the benchmark

Table 3-1. Relative efficiency: globally invalid covariates—Case 1

DGP	n	Valid			Globally invalid		CSC average	Smoothed average	CSC selected
		Set 1	Set 2	Set 3	Set 4	Set 5			
					Average treatment effect (β_0)				
1	250	1.000	0.765	1.133	1.329	4.302	0.869	0.899	0.937
	500	1.000	0.847	1.258	2.176	8.320	0.926	0.943	0.986
	1000	1.000	0.922	1.362	4.027	17.293	0.965	0.942	0.982
	5000	1.000	0.943	1.360	17.626	84.588	0.914	0.917	0.961
2	250	1.000	0.738	1.006	1.029	2.910	0.850	0.828	0.911
	500	1.000	0.811	1.141	1.619	5.507	0.912	0.925	0.950
	1000	1.000	0.878	1.218	2.796	11.118	0.963	1.000	0.975
	5000	1.000	0.905	1.240	11.737	54.975	0.901	0.967	0.910
3	250	1.000	0.766	0.977	0.851	1.779	0.760	0.823	0.825
	500	1.000	0.838	1.072	1.087	2.875	0.858	0.912	0.913
	1000	1.000	0.886	1.112	1.606	5.186	0.920	0.966	0.968
	5000	1.000	0.908	1.139	5.068	22.249	0.916	0.919	0.919
					Average treatment effect on the treated (γ_0)				
1	250	1.000	0.614	0.994	0.903	2.931	0.716	0.774	0.804
	500	1.000	0.748	1.175	1.625	6.254	0.856	0.898	0.932
	1000	1.000	0.847	1.280	3.077	13.224	0.953	0.953	0.985
	5000	1.000	0.843	1.282	13.388	64.228	0.846	0.841	0.850
2	250	1.000	0.560	0.853	0.735	2.352	0.649	0.620	0.719
	500	1.000	0.663	1.006	1.281	4.877	0.778	0.765	0.837
	1000	1.000	0.753	1.095	2.381	10.402	0.876	0.868	0.887
	5000	1.000	0.786	1.127	10.652	53.234	0.803	0.868	0.787
3	250	1.000	0.737	0.978	0.615	0.853	0.696	0.769	0.770
	500	1.000	0.818	1.080	0.753	1.241	0.792	0.872	0.873
	1000	1.000	0.861	1.112	0.999	1.965	0.893	0.963	0.964
	5000	1.000	0.868	1.143	2.443	6.897	0.998	1.015	1.016

NOTE: See the note to Table 1.

set of covariates, that is, $\mathbf{X}_1 = (X^*, X_1, X_2, X_3, X_4)$. It is easy to verify that $\mathbf{X}_1$ is valid, as $D \perp (Y(0), Y(1)) | \mathbf{X}_1$.

For individual estimators of β_j and γ_j, we use the imputation estimators in Equation (1), where $\hat{m}_{0j}(.)$ and $\hat{m}_{1j}(.)$ are the sieve estimators. The variance–covariance matrices V_β and V_γ are estimated as in Equation (3), where $\hat{p}_j(.)$ is the sieve logit estimator as in HIR. The sieve space is natural splines (for a review of sieve methods, see Chen 2007). For each dimension, we let the number of terms be $n^{1/4}$ and also include their cross-product terms in the sieve space. We consider the sample size $n = 250$, 500, 1000, and 5000. The number of replications is 5000.

6.1 Valid Covariates

We consider $J = 5$ sets of valid covariates. As discussed above, the first set is $\mathbf{X}_1 = (X^*, X_1, X_2, X_3, X_4)$. The remaining four sets are smaller: $\mathbf{X}_2 = (X^*, X_1)$, $\mathbf{X}_3 = (X^*, X_2)$, $\mathbf{X}_4 = (X^*, X_3)$ and $\mathbf{X}_5 = (X^*, X_4)$. It is easy to see that all five sets of covariates satisfy the unconfoundedness assumption. We consider the estimators using each set of covariates and three data-driven averaging or selected estimators: (1) the CSC average estimator, (2) the smoothed average estimator (see Remark 3), and (3) the CSC selected estimator (see Remark 1). Table 1 reports their mean squared errors (MSEs). We normalize the MSEs by dividing them by the MSE of the estimator using the first set of covariates. For most DGPs and sample sizes, the individual estimators based on covariate sets 2 and 4 perform better than that based on the benchmark covariate set 1, while those based on covariate sets 3 and 5 perform worse. This suggests that it is not always more desirable to use a smaller or larger set. Among all three averaging or selected estimators considered here, our CSC average estimator performs the best. Compared with the estimator based on the benchmark covariate set 1, our CSC average estimator reduces the MSEs by 15%–40%. Our CSC average estimator often beats the infeasible single best estimator, especially when the sample size is relatively large. The smoothed average estimator is the second best. The performance of the selected estimator is similar to that of the single best estimator when the sample size is large.

Table 4 shows the empirical coverage of the 95% confidence intervals of the CSC average estimator based on the asymptotic distribution in Section 3. When the sample size n is small, we often have slight undercoverage. However, as the sample size n increases, the empirical coverage becomes close to the nominal 95%.

6.2 Locally Invalid Covariates

We consider the case where some sets of covariates are locally invalid. Again, we consider five sets of covariates. The first set $\mathbf{X}_1 = (X^*, X_1, X_2, X_3, X_4)$ is the benchmark set of valid

Table 3-2. Relative efficiency: globally invalid covariates—Case 2

DGP	n	Valid	Globally invalid				CSC average	Smoothed average	CSC selected
		Set 1	Set 2	Set 3	Set 4	Set 5			
					Average treatment effect (β_0)				
1	250	1.000	1.317	4.314	1.329	4.302	1.026	1.079	1.086
	500	1.000	2.138	8.368	2.176	8.320	1.124	1.186	1.191
	1000	1.000	4.053	17.249	4.027	17.293	1.187	1.175	1.179
	5000	1.000	17.712	84.595	17.626	84.588	1.023	1.000	1.000
2	250	1.000	1.030	2.915	1.029	2.910	0.997	1.114	1.055
	500	1.000	1.583	5.543	1.619	5.507	1.085	1.382	1.149
	1000	1.000	2.807	11.117	2.796	11.118	1.193	1.650	1.236
	5000	1.000	11.806	54.915	11.737	54.975	1.041	1.374	1.000
3	250	1.000	0.837	1.776	0.851	1.779	0.899	0.958	0.959
	500	1.000	1.103	2.913	1.087	2.875	0.974	1.033	1.033
	1000	1.000	1.621	5.157	1.606	5.186	1.071	1.145	1.146
	5000	1.000	5.095	22.316	5.068	22.249	1.109	1.065	1.065
					Average treatment effect on the treated (γ_0)				
1	250	1.000	0.895	2.968	0.903	2.931	0.938	1.022	1.029
	500	1.000	1.598	6.294	1.625	6.254	1.080	1.178	1.188
	1000	1.000	3.086	13.158	3.077	13.224	1.260	1.340	1.348
	5000	1.000	13.448	64.211	13.388	64.228	1.087	1.014	1.015
2	250	1.000	0.737	2.384	0.735	2.352	0.901	0.880	0.992
	500	1.000	1.248	4.908	1.281	4.877	1.037	1.172	1.143
	1000	1.000	2.386	10.352	2.381	10.402	1.226	1.475	1.321
	5000	1.000	10.714	53.147	10.652	53.234	1.139	1.491	1.036
3	250	1.000	0.607	0.852	0.615	0.853	0.795	0.875	0.875
	500	1.000	0.760	1.267	0.753	1.241	0.852	0.935	0.936
	1000	1.000	1.002	1.951	0.999	1.965	0.949	1.056	1.056
	5000	1.000	2.454	6.914	2.443	6.897	1.213	1.343	1.344

NOTE: See the note to Table 1.

covariates. For the remaining four sets of covariates, we consider two cases:

Case 1: $\mathbf{X}_2 = (X^*, X_1)$, $\mathbf{X}_3 = (X^*, X_2)$, $\mathbf{X}_4 = \left(X^\dagger, X_3\right)$ and $\mathbf{X}_5 = \left(X^\dagger, X_4\right)$, where

$$X^\dagger = X^* + \frac{1}{\sqrt{n}}\left[(1+D)(1+U)+\epsilon^\dagger\right], \qquad (10)$$

and $\epsilon^\dagger$ is a standard normal random variable. Here $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ are three sets of valid covariates. $\mathbf{X}_4$ and $\mathbf{X}_5$ are locally invalid, as the presence of $\frac{1}{\sqrt{n}}[(1+D)(1+U)+\epsilon^\dagger]$ in $X^\dagger$ makes the unconfoundedness assumption invalid for any finite n. However $X^\dagger$ approaches X^* when n goes to infinity. Here, $J_0 = 3$. For the implementation details of our average estimator when $J_0 > 1$, see Remark 8.

Case 2: $\mathbf{X}_2 = (X^\dagger, X_1)$, $\mathbf{X}_3 = (X^\dagger, X_2)$, $\mathbf{X}_4 = (X^\dagger, X_3)$, and $\mathbf{X}_5 = (X^\dagger, X_4)$, where $X^\dagger$ is as defined in Equation (10). In this case, we only have one set of valid covariates: $\mathbf{X}_1$ and four sets of locally invalid covariates: $\mathbf{X}_2$, $\mathbf{X}_3$, $\mathbf{X}_4$, and $\mathbf{X}_5$, that is, $J_0 = 1$.

Table 2-1 presents the relative efficiency of different estimators for Case 1. Compared with the estimator based on the benchmark covariate set 1, in general, the estimator based on covariate set 2 performs better, while those based on covariate sets 3 and 5 perform worse. The relative performance of the estimator based on covariate set 4 depends on the sample size. It is interesting to note that sometimes the estimators based on locally invalid covariates outperform those based on valid covariates. For most DGPs and sample sizes, the performance of our CSC average estimator is the best among all three data-driven methods. The MSE is reduced by about 5%–35% compared with that when the estimator based on the first set of covariates is used. Both the smoothed average estimator and the selected estimator perform better than the estimator based on the benchmark covariate set 1.

Table 2-2 shows the relative performance of the estimators for Case 2. Compared with Case 1, covariate sets 2 and 3 also become locally invalid, thus the performances of the estimators based on them deteriorate. As a consequence, the performances of all three data-driven methods deteriorate. However, our CSC average estimator still outperforms the estimator based on the benchmark covariate set 1, and can reduce the MSEs by about 5%–15%.

Table 4 reports the empirical coverage of the 95% confidence intervals of the CSC average estimator for both cases using the simple method (Equation (6)) with $K = 1000$. The actual coverage is close to the nominal coverage of 95% when the sample size is large.

6.3 Globally Invalid Covariates

Even though our theoretical results do not apply to globally invalid covariates, in practice, our CSC can still be used. We consider two cases involving globally invalid covariates:

Table 4. Empirical coverage of 95% confidence intervals for CSC average estimators

DGP	n	Valid covariates	Locally invalid covariates		Globally invalid covariates	
			Case 1	Case 2	Case 1	Case 2
		Average treatment effect (β_0)				
1	250	0.903	0.901	0.881	0.894	0.877
	500	0.925	0.918	0.908	0.912	0.903
	1000	0.943	0.939	0.929	0.938	0.929
	5000	0.948	0.946	0.939	0.947	0.947
2	250	0.901	0.899	0.883	0.891	0.873
	500	0.922	0.919	0.907	0.915	0.906
	1000	0.940	0.936	0.923	0.935	0.920
	5000	0.948	0.949	0.939	0.948	0.948
3	250	0.910	0.903	0.885	0.902	0.885
	500	0.923	0.919	0.905	0.913	0.908
	1000	0.936	0.935	0.927	0.929	0.924
	5000	0.949	0.951	0.940	0.950	0.944
		Average treatment effect on the treated (γ_0)				
1	250	0.888	0.890	0.863	0.878	0.858
	500	0.916	0.917	0.901	0.902	0.896
	1000	0.931	0.929	0.927	0.926	0.917
	5000	0.949	0.947	0.940	0.950	0.956
2	250	0.886	0.890	0.863	0.887	0.858
	500	0.908	0.911	0.897	0.903	0.891
	1000	0.930	0.934	0.918	0.926	0.910
	5000	0.943	0.945	0.935	0.947	0.946
3	250	0.883	0.889	0.873	0.888	0.878
	500	0.904	0.912	0.906	0.906	0.905
	1000	0.912	0.922	0.922	0.913	0.915
	5000	0.937	0.944	0.942	0.938	0.931

Case 1: $\mathbf{X}_2 = (X^*, X_1)$, $\mathbf{X}_3 = (X^*, X_2)$, $\mathbf{X}_4 = (X_3)$, and $\mathbf{X}_5 = (X_4)$. It is easy to see that $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ are valid, while $\mathbf{X}_4$ and $\mathbf{X}_5$ are globally invalid. Thus, $J_0 = 3$.

Case 2: $\mathbf{X}_2 = (X_1)$, $\mathbf{X}_3 = (X_2)$, $\mathbf{X}_4 = (X_3)$, and $\mathbf{X}_5 = (X_4)$. Here we only have one set of valid covariates: $\mathbf{X}_1$ and four sets of globally invalid covariates: $\mathbf{X}_2$, $\mathbf{X}_3$, $\mathbf{X}_4$, and $\mathbf{X}_5$. Thus, $J_0 = 1$.

Table 3-1 presents the relative efficiency for Case 1. Apparently, when the sample size is relatively large, the estimators based on the globally invalid covariate sets 4 and 5 perform much worse than those based on the valid covariate sets 1, 2, and 3. For example, for DGP 1 (ATE) and $n = 5000$, the MSE based on covariate set 5 is 84 times larger than that based on covariate set 1. However, when the sample size is small, sometimes the estimators based on the globally invalid covariates outperform those based on the valid covariates. Our CSC average estimator performs well, and beats the estimator based on the benchmark covariate set 1 for all DGPs and all sample sizes.

Table 3-2 reports the relative performance for Case 2. Compared with Case 1, covariate sets 2 and 3 also become globally invalid. Thus for large sample sizes, the estimators based on covariate sets 2–5 are all much worse than that based on the benchmark covariate set 1. Even in this case, our average CSC estimator still performs slightly better than or is comparable to the estimator based on the valid covariate set 1 when the sample

Table 5. Weights in CSC average estimators: globally invalid covariates—Case 2

DGP	n	Valid	Globally invalid			
		Set 1	Set 2	Set 3	Set 4	Set 5
		Average treatment effect (β_0)				
1	5000	0.975	0.013	0.000	0.013	0.000
2	5000	0.960	0.020	0.000	0.020	0.000
3	5000	0.871	0.068	0.000	0.061	0.000
		Average treatment effect on the treated (γ_0)				
1	5000	0.945	0.028	0.000	0.028	0.000
2	5000	0.924	0.038	0.000	0.038	0.000
3	5000	0.536	0.232	0.001	0.230	0.001

NOTE: The numbers in the main entries are the mean weights of the individual estimators in the CSC average estimators over 5000 replications.

Table 6-1. Outcome and treatment

	Name	Description
Y	Difference in OROA	the three year average of industry- and performance-adjusted OROA after CEO transitions *minus* the three-year average before CEO transitions
D	Family CEO	= 1 if the incoming CEO of the firm is related by blood or marriage to the departing CEO, to the founder, larger shareholder = 0 otherwise

Table 6-2. Covariates (X)

	Name	Description
X_1	Ln sales	logarithm of sales one year prior to the CEO transition
X_2	Prior OROA	industry-adjusted OROA one year prior to the CEO transition
X_3	Prior M-B	industry-adjusted market-to-book (M-B) ratio one year prior to the CEO transition
X_4	Board ownership	the fraction of ownership held by officers and directors
X_5	Family directors	= 1 if the fraction of family to total directors is higher than the median in the sample = 0 otherwise
X_6	Mean pre-transition OROA transition OROA	three-year pre-transition average of the industry- and performance-adjusted OROA
X_7	Selective college	= 1 if the college attended by the incoming CEO is not "very competitive" or higher in Barron's ranking = 0 otherwise
X_8	Graduate school	= 1 if the incoming CEO attended a graduate school = 0 otherwise
X_9	Age promoted	the age when the incoming CEO is appointed
X_{10}	Woman	= 1 if the incoming CEO is a woman = 0 otherwise
X_{11}	R&D expenses	= 1 if the firm reported positive R&D expenses the year prior to the CEO transition =0 otherwise
X_{12}	Nonretirements	=1 if the departing CEO was not reported to leave the firm due to a "retirement" =0 otherwise
X_{13}	Early succession	=1 if the departing CEO left his position before 65 = 0 otherwise
X_{14}	Departing CEO	=1 if the departing CEO continued as chairman after the CEO transition =0 otherwise
X_{15}	CEO ownership	the ownership share of the incoming CEO

size is small, and performs only slightly worse when the sample size is large. Apparently, the performance of our average estimator deteriorates as we add more sets of globally invalid covariates, but is still within a reasonable range even when the majority sets of covariates are globally invalid.

In Remark 7, we argue that for large samples, our method should assign a weight close to zero to the estimators based on globally invalid covariates. We can confirm this through simulations. Table 5 reports the mean weights of each individual estimator in our average estimators (over 5000 replications) for Case 2 with $n = 5000$. For most DGPs, our method assigns very small weights to the estimators based on the globally invalid covariates. For example, for ATE in DGP 1, the first estimator based on the benchmark valid covariates receives a weight of 97.5%, while the other four estimators based on the globally invalid covariates all receive a weight close to 0.

Table 4 shows the empirical coverage of the 95% confidence intervals for both cases. Again, the actual coverage approaches the nominal coverage of 95% when the sample size increases.

The simulation results above suggest that our method can be applied to globally invalid covariates, even though our theory applies to locally invalid covariates only.

Remark 10. So far, we have only considered the case where the number of candidate sets of covariates (J) is fixed. As a referee suggests, it would be interesting to consider the case where J increases with the sample size (n). Therefore, we also let $J = J_n = \lfloor n^{1/3} \rfloor$ for all cases considered above, where $\lfloor \cdot \rfloor$ denotes the integer part of $\cdot$. Our average estimator performs similarly well in the increasing J case. To save space, the detailed results are not reported here, but are available upon request. The asymptotic theory for the increasing J case is, however, quite involved. For example, the variance–covariance matrices V_β and V_γ become high dimensional. To estimate them, some regularization is often needed (see, e.g., Bickel and Levina 2007). We leave the rigorous theoretical analysis for the increasing J case for future research.

Remark 11. In an earlier version of this article, we also considered a case where the benchmark set of covariates is small. However, as an associate editor and a referee kindly point out that in practice, researchers often use a large set of covariates. Therefore we consider the DGPs above. The results for the case where a small set of covariates is used as a benchmark are available upon request.

7. EMPIRICAL APPLICATIONS

We apply our method to study the effects of inherited family control on firm performance as in Pérez-González (2006) and Lu and White (2014). Specifically, we are interested in whether firms with familiarly related incoming chief executive officers (CEOs) underperform relative to firms with unrelated incoming CEOs in terms of the firms' operating return on assets (OROA). We use the data on 335 management transitions of publicly traded U.S. corporations as in Pérez-González (2006). Both Pérez-González (2006) and Lu and White (2014) used linear regressions. Here, we consider the semiparametric estimation

Table 7. Sets of covariates

Set	Covariates
Set 1 ($\mathbf{X}_1$)	$\{X_2, X_4, X_5\}$
Set 2 ($\mathbf{X}_2$)	$\{X_1, X_2, X_3, X_4, X_5, X_6\}$
Set 3 ($\mathbf{X}_3$)	$\{X_2, X_4, X_5, X_7, X_8, X_9, X_{10}\}$
Set 4 ($\mathbf{X}_4$)	$\{X_2, X_4, X_5, X_{11}\}$
Set 5 ($\mathbf{X}_5$)	$\{X_2, X_4, X_5, X_{12}, X_{13}, X_{14}, X_{15}\}$

of ATE and ATT. Tables 6-1 and 6-2 are the same as those in Lu and White (2014) and describe the outcomes (Y), treatment (D) and 15 covariates (X).

We consider the five candidate sets of covariates listed in Table 7, which are exactly the same as those in Lu and White (2014). We first assume that all five sets of covariates are valid. Then we assume some covariates are locally invalid. Lu and White (2014) called the first set of covariates ($\mathbf{X}_1$) core covariates, thus we treat it as the only valid set of covariates and treat all other sets of covariates as locally invalid. (We also try treating other sets of covariates as valid covariates and find similar results.) Lu and White (2014) provided a detailed discussion on how these five sets of covariates are chosen. For brevity, we omit the details here.

The estimation of ATE and ATT is implemented in the same way as in the simulation. In this application, we have many binary covariates. Given the relatively small sample size, these binary variables enter the sieve terms linearly (i.e., without interacting with those continuous covariates). Table 8 reports the estimation results. For ATE, all individual estimates are negative and significant at the 5% significance level. Our CSC average estimate assuming all sets of covariates are valid is −0.0244 with the 95% confidence interval of [−0.0380, −0.0107]. The CSC average estimate assuming that covariate sets 2–5 are locally invalid is −0.0253 with the 95% confidence interval of [−0.0395, −0.0100].

For ATT, all the individual estimates are negative except that based on covariate set 3. However, they are not significant at the 5% level except that based on covariate set 5. If we assume that all covariates are valid, then our CSC average estimate is −0.0173 with the 95% confidence interval of [−0.0360, 0.0014]. If we assume that covariate sets 2–5 are locally invalid, our CSC average estimate is −0.0217 with the 95% confidence interval of [−0.0458, 0.0003]. This suggests that ATT is not significant at the 5% level.

Note that in general, our CSC average estimator has narrower confidence intervals than the individual estimators. This suggests the efficiency gains of our new estimator. We also implement the Hausman-type test for Hypotheses 1 and 2

Table 8. Estimation of ATE and ATT

	ATE (β_0)			ATT (γ_0)		
	Estimate	CI	Length of CI	Estimate	CI	Length of CI
			Individual estimates			
Set 1	−0.0238	[−0.0405, −0.0070]	0.0335	−0.0216	[−0.0463, 0.0032]	0.0495
Set 2	−0.0230	[−0.0380, −0.0080]	0.0300	−0.0205	[−0.0422, 0.0011]	0.0433
Set 3	−0.0253	[−0.0402, −0.0105]	0.0297	0.0018	[−0.0235, 0.0270]	0.0505
Set 4	−0.0230	[−0.0399, −0.0061]	0.0339	−0.0197	[−0.0424, 0.0030]	0.0454
Set 5	−0.0212	[−0.0411, −0.0013]	0.0399	−0.0255	[−0.0490, −0.0020]	0.0470
			CSC: assuming valid covariates			
CSC average	−0.0244	[−0.0380, −0.0107]	0.0273	−0.0173	[−0.0360, 0.0014]	0.0374
Smoothed average	−0.0233	[−0.0386, −0.0079]	0.0307	−0.0171	[−0.0381, 0.0038]	0.0419
CSC selected	−0.0253	[−0.0402, −0.0105]	0.0297	−0.0205	[−0.0422, 0.0011]	0.0433
			CSC: assuming locally invalid covariates			
CSC average	−0.0253	[−0.0395, −0.0100]	0.0295	−0.0217	[−0.0458, 0.0003]	0.0461
Smoothed average	−0.0233	[−0.0380, −0.0080]	0.0300	−0.0174	[−0.0410, −0.0008]	0.0403
CSC selected	−0.0253	[−0.0418, −0.0095]	0.0322	−0.0205	[−0.0434, 0.0032]	0.0465

NOTE: CI stands for 95% confidence interval.

(see Section 5) for ATE and ATT, respectively. The p-values for ATE and ATT are 0.974 and 0.134, respectively.

In summary, we find that the ATE estimates are around -0.025 and are significant at the 5% level. This is consistent with the linear regression results in Pérez-González (2006) and Lu and White (2014), in which case ATE and ATT are the same. However, we find that the magnitudes of ATT are in general smaller than those of ATE, and in most cases, ATT is not significant at the 5% level.

8. CONCLUSION

We provide a data-driven criterion to select the optimal estimator or combine estimators optimally for estimating the average treatment effect and the average treatment effect on the treated. This criterion can be applied to both the case where all sets of covariates are valid and the case where some sets of covariates are locally invalid. We introduce new average estimators that outperform the optimal estimators based on a single set of covariates. We derive the asymptotic distribution of the new estimators and propose the construction of valid confidence intervals. Simulations show that our new estimators perform well in finite samples. The new method is applied to study the effects of inherited control on firm performance, and the results attest to the efficiency gains of our new estimators.

9. MATHEMATICAL PROOFS

Proof of Lemma 1. By Assumption A.2, $\sqrt{n}\left(\hat{\beta}-\beta\right) = n^{-1/2}\sum_{i=1}^{n}\psi_i + o_p(1)$. This implies that

$$\sqrt{n}\left(\hat{\beta}-\begin{pmatrix}\beta_0\\ \dots\\ \beta_0\\ \beta_0+\frac{\delta_{J_0+1}}{\sqrt{n}}\\ \dots\\ \beta_0+\frac{\delta_J}{\sqrt{n}}\end{pmatrix}\right) = n^{-1/2}\sum_{i=1}^{n}\psi_i + o_p(1),$$

i.e.,

$$\sqrt{n}\left(\hat{\beta}-\beta_0\right) = \delta + n^{-1/2}\sum_{i=1}^{n}\psi_i + o_p(1) \xrightarrow{d} N\left(\delta,\ V_\beta\right),$$

by the central limit theorem. Thus part (i) is proved. Part (ii) can be proved analogously. □

Proof of Proposition 2. By Assumption A.2, we have $\sqrt{n}\left(\hat{\beta}-\beta\right) = n^{-1/2}\sum_{i=1}^{n}\psi_i + o_p(1)$. This implies that $\sqrt{n}\left(S\hat{\beta}-S\beta\right) = n^{-1/2}S\cdot\sum_{i=1}^{n}\psi_i + o_p(1)$ by the continuous mapping theorem. Note that

$$\sqrt{n}S\beta = \sqrt{n}\begin{bmatrix}0&0&0&\dots&0\\ -1&1&0&\dots&0\\ -1&0&1&\dots&0\\ &\cdot&\cdot&\cdot&\\ -1&0&\dots&0&1\end{bmatrix}\cdot\begin{pmatrix}\beta_0\\ \beta_0+\frac{\delta_2}{\sqrt{n}}\\ \dots\\ \beta_0+\frac{\delta_J}{\sqrt{n}}\end{pmatrix} = \sqrt{n}\begin{pmatrix}0\\ \frac{\delta_2}{\sqrt{n}}\\ \dots\\ \frac{\delta_J}{\sqrt{n}}\end{pmatrix} = \delta.$$

Thus $\sqrt{n}S\hat{\beta}-\delta = n^{-1/2}S\cdot\sum_{i=1}^{n}\psi_i + o_p(1)$, that is,

$$\hat{\delta} = \sqrt{n}S\hat{\beta} = \delta + n^{-1/2}S\cdot\sum_{i=1}^{n}\psi_i + o_p(1) \xrightarrow{d} \delta + S\mathbf{P} \sim N\left(\delta,\ SV_\beta S'\right),$$

by the central limit theorem. This proves part (i). Part (ii) can be proved analogously. □

Proof of Proposition 3. By Proposition 2, $\hat{\delta} \xrightarrow{d} \delta+S\mathbf{P}$. By the continuous mapping theorem, $\hat{\delta}\hat{\delta}' \xrightarrow{d} (\delta+S\mathbf{P})(\delta+S\mathbf{P})'$. We can show that $\hat{V}_\beta \xrightarrow{p} V_\beta$. Thus, by the continuous mapping theorem,

$$\mathrm{CSC}_\beta(w) = w'\left(\hat{\delta}\hat{\delta}' - S\hat{V}_\beta S' + \hat{V}_\beta\right)w \xrightarrow{d} w'((\delta+S\mathbf{P})(\delta+S\mathbf{P})' - SV_\beta S' + V_\beta)w = w\Omega_\beta w'.$$

Note that

$$\begin{aligned}\mathbb{E}\left(w\Omega_\beta w'\right) &= w'\mathbb{E}\left((\delta+S\mathbf{P})(\delta+S\mathbf{P})' - SV_\beta S' + V_\beta\right)w\\ &= w'\mathbb{E}\left(\delta\delta'+S\mathbf{P}\delta'+\delta\mathbf{P}'S'+S\mathbf{P}\mathbf{P}'S' - SV_\beta S'+V_\beta\right)w\\ &= w'\left(\delta\delta' + S\mathbb{E}\left(\mathbf{P}\mathbf{P}'\right)S' - SV_\beta S' + V_\beta\right)w\\ &= w'\left(\delta\delta' + V_\beta\right)w = \mathrm{AMSE}_\beta(w)\end{aligned}$$

based on the fact that $\mathbf{P}$ is a zero mean normal vector and $\mathbb{E}(\mathbf{P}\mathbf{P}') = V_\beta$. Given that $\mathrm{CSC}_\beta(w) \xrightarrow{d} w\Omega_\beta w'$, by the continuous mapping theorem, we have $\hat{w}_\beta^* \xrightarrow{d} w_\beta^*$. The proof of part (i) is complete. The proof of part (ii) is analogous. □

Proof of Proposition 4. As shown in Proposition 2 and Proposition 3, $\hat{w}_\beta^* \xrightarrow{d} w_\beta^*$ and $\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} \delta+\mathbf{P}$. Thus by the continuous mapping theorem, we have

$$\sqrt{n}\left(\hat{\beta}^*-\beta_0\right) = \hat{w}_\beta^{*\prime}\cdot\sqrt{n}\left(\hat{\beta}-\beta_0\right) \xrightarrow{d} w_\beta^{*\prime}\cdot[\delta+\mathbf{P}].$$

This completes the proof of part (i). The proof of part (ii) is analogous. □

Proof of Proposition 5. Given the value of δ, for a sufficiently large K, we have $1-\alpha = \lim_{n\to\infty} P\left(\beta_0 \in \widehat{CI}_\beta(\delta)\right)$. We can further decompose $\lim_{n\to\infty}\Pr\left(\beta_0 \in \widehat{CI}_\beta(\delta)\right)$ into two parts:

$$\begin{aligned}1-\alpha &= \lim_{n\to\infty} P\left(\beta_0 \in \widehat{CI}_\beta(\delta)\right)\\ &= \lim_{n\to\infty} P\left(\beta_0 \in \widehat{CI}_\beta(\delta) \cap \tilde{\delta} \in \hat{A}_{\tilde{\delta}}\right)\\ &\quad + \lim_{n\to\infty} P\left(\beta_0 \in \widehat{CI}_\beta(\delta) \cap \tilde{\delta} \in \hat{A}_{\tilde{\delta}}^c\right)\\ &\le \lim_{n\to\infty} P\left(\beta_0 \in \widehat{CI}_\beta\right) + \lim_{n\to\infty} P\left(\tilde{\delta} \in \hat{A}_{\tilde{\delta}}^c\right),\end{aligned}$$

where the last inequality is due to the definition of $\widehat{CI}_\beta$ and $\{\beta_0 \in \widehat{CI}_\beta(\delta) \cap \tilde{\delta} \in \hat{A}_{\tilde{\delta}}^c\} \subseteq \{\tilde{\delta} \in \hat{A}_{\tilde{\delta}}^c\}$. Thus,

$$\lim_{n\to\infty} P\left(\beta_0 \in \widehat{CI}_\beta\right) \ge 1-\alpha - \lim_{n\to\infty} P\left(\delta \in \hat{A}_{\tilde{\delta}}^c\right) = 1-\alpha-\pi,$$

where the last equality is due to the construction of $\hat{A}_{\tilde{\delta}}^c$. □

Proof of Proposition 6. For part (i), by Assumption A.2, $\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{d} \mathbf{P}$, which implies that $\sqrt{n}\tilde{S}(\hat{\beta}-\beta) \xrightarrow{d} \tilde{S}\mathbf{P}$. Under $\mathbb{H}_{10}$, $\tilde{S}\beta = 0$, thus, $\sqrt{n}\tilde{S}\hat{\beta} \xrightarrow{d} \tilde{S}\mathbf{P}$. We can show that $\hat{V}_\beta \xrightarrow{p}$

V_β. Thus

$$\mathcal{W}_\beta = n \cdot \hat{\beta}' \tilde{S}' \left[\tilde{S} \hat{V}_\beta \tilde{S}'\right]^{-1} \tilde{S}\hat{\beta} \xrightarrow{d} \left(\tilde{S}\mathbf{P}\right)' \left[\tilde{S} V_\beta \tilde{S}'\right]^{-1} \left(\tilde{S}\mathbf{P}\right) \sim \chi^2_{J-1},$$

based on the fact that $\text{var}\left(\tilde{S}\mathbf{P}\right) = \tilde{S} V_\beta \tilde{S}'$. For part (ii), we can show that

$$n^{-1}\mathcal{W}_\beta = \hat{\beta}' \tilde{S}' \left[\tilde{S} \hat{V}_\beta \tilde{S}'\right]^{-1} \tilde{S}\hat{\beta} = \beta' \tilde{S}' \left[\tilde{S} V_\beta \tilde{S}'\right]^{-1} \tilde{S}\beta + o_p(1),$$

and the conclusion follows. For part (iii), by Assumption A.2, $\sqrt{n}\tilde{S}\left(\hat{\beta} - \beta\right) \xrightarrow{d} \tilde{S}\mathbf{P}$, thus $\sqrt{n}\tilde{S}\hat{\beta} - \sqrt{n}\tilde{S}\beta \xrightarrow{d} \tilde{S}\mathbf{P}$. Under $\mathbb{H}_{1a}$, $\sqrt{n}\tilde{S}\beta = \tilde{\delta}$, thus $\sqrt{n}\tilde{S}\hat{\beta} \xrightarrow{d} \tilde{\delta} + \tilde{S}\mathbf{P}$. Therefore,

$$\mathcal{W}_\beta = n \cdot \hat{\beta}' \tilde{S}' \left[\tilde{S} \hat{V}_\beta \tilde{S}'\right]^{-1} \tilde{S}\hat{\beta} \xrightarrow{d} \left(\tilde{\delta} + \tilde{S}\mathbf{P}\right)' \left[\tilde{S} V_\beta \tilde{S}'\right]^{-1} \left(\tilde{\delta} + \tilde{S}\mathbf{P}\right),$$

by the continuous mapping theorem. □

ACKNOWLEDGMENTS

The author gratefully thanks the joint editor Rong Chen, the associate editor, and two anonymous referees for their many helpful comments. The author also thanks Manuel Arellano, Xiaohong Chen, Bruce Hansen, Hidehiko Ichimura, David Jacho-Chávez, Toru Kitagawa, Tong Li, Liangjun Su, and Alan T. K. Wan for their valuable comments and suggestions. The author acknowledges support from the Hong Kong Research Grants Council (RGC) under grant number 643711.

[Received December 2013. Revised October 2014.]

REFERENCES

Bickel, P. J., and Levina, E. (2007), "Regularized Estimation of Large Covariance Matrices," *The Annals of Statistics*, 36, 199–227. [519]

Chalak, K., and White, H. (2012), "Causality, Conditional Independence, and Graphical Separation in Settable Systems," *Neural Computation*, 24, 1611–1668. [506]

Chen, X. (2007), "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, vol. 6, eds. J. J. Heckman and E. E. Leamer, New York: Elsevier. [516]

Chen, X., Jacho-Chávez, D., and Linton, O. (2015), "Averaging an Increasing Number of Moment Condition Estimators," *Econometric Theory,* forthcoming. [507]

Claeskens, G., and Hjort, N. L. (2003), "The Focused Information Criterion," *Journal of the American Statistical Association*, 98, 900–916. [507]

——— (2008), *Model Selection and Model Averaging*, Cambridge: Cambridge University Press. [507]

Crainiceanu, C., Dominici, F., and Parmigiani, G. (2008), "Adjustment Uncertainty in Effect Estimation," *Biometrika*, 95, 635–651. [507]

Crump, R., Hotz, V., Imbens, G., and Mitnik, O. (2009), "Dealing With Limited Overlap in Estimation of Average Treatment Effects," *Biometrika*, 96, 187–199. [509]

Dawid, A. P. (1979), "Conditional Independence in Statistical Theory," *Journal of the Royal Statistical Society,* Series B, 41, 1–31. [509]

De Luna, X., Waernbaum, I., and Richardson, T. S. (2011), "Covariate Selection for the Nonparametric Estimation of an Average Treatment Effect," *Biometrika*, 98, 861–875. [507]

DiTraglia, F. (2012), "Using Invalid Instruments on Purpose: Focused Moment Selection and Averaging for GMM," University of Pennsylvania Working Paper. [507,510,512]

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effect," *Econometrica*, 66, 315–331. [508]

Hahn, J. (2004), "Functional Restriction and Efficiency in Causal Inference," *Review of Economics and Statistics*, 86, 73–76. [506,510]

Hansen, B. (2007), "Least Squares Model Averaging," *Econometrica*, 75, 1175–1189. [507,509]

Hansen, B., and Racine, J. (2012), "Jackknife Model Averaging," *Journal of Econometrics*, 167, 38–46. [507,509]

Heckman, J., and Navarro-Lozano, S. (2004), "Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models," *Review of Economics and Statistics*, 86, 30–57. [506,510]

Hirano, K., Imbens, G., and Ridder, G. (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189. [508]

Hjort, N. L., and Claeskens, G. (2003), "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98, 879–899. [507]

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–417. [507]

Imbens, G., Newey, W., and Ridder, G. (2007), "Mean-Squared-Error Calculations for Average Treatment Effects," Harvard University Working Paper. [508]

Imbens, G., and Wooldridge, J. (2009), "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86. [506]

Khan, S., and Tamer, E. (2010), "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 78, 2021–2042. [509]

Kitagawa, T., and Muris, C. (2013), "Covariate Selection and Model Averaging in Semiparametric Estimation of Treatment Effects," University College London Working Paper. [507]

Li, Q., and Racine, J. (2007), *Nonparametric Econometrics Theory and Practice*, Princeton, NJ: Princeton University Press. [508]

Liu, C. (2015), "Distribution Theory of the Least Squares Averaging Estimator," *Journal of Econometrics*, *Journal of Econometrics*, 186, 142–159. [507,510,511]

Lu, X., and Su, L. (2015), "Jackknife Model Averaging for Quantile Regressions," *Journal of Econometrics*, forthcoming. [507]

Lu, X., and White, H. (2014), "Robustness Check and Robustness Test in Applied Economics," *Journal of Econometrics*, 178, 194–206. [506,509,514,515,519,520,521]

Pearl, J. (2009), *Causality*, Cambridge: Cambridge University Press. [506]

Pérez-González, F. (2006), "Inherited Control and Firm Performance," *American Economic Review*, 96, 1559–1588. [519,521]

Reichenbach, H. (1956), *The Direction of Time*, Berkeley, CA: University of California Press. [506]

Rosenbaum, P. (1984), "The Consequences of Adjusting for a Concomitant Variable That Has Been Affected by Treatment," *Journal of the Royal Statistical Society,* Series A, 147, 656–666. [506]

Staiger, D., and Stock, J. (1997), "Instrumental Variables Regression With Weak Instruments," *Econometrica*, 65, 557–586. [511]

VanderWeele, T. J., and Shpitser, I. (2011), "A New Criterion for Confounder Selection," *Biometrics*, 67, 1406–1413. [506]

Vansteelandt, S., Bekaert, M., and Claeskens, G. (2012), "On Model Selection and Model Misspecification in Causal Inference," *Statistical Methods in Medical Research*, 21, 7–30. [507]

Wang, C., Parmigiani, G., and Dominici, F. (2012), "Bayesian Effect Estimation Accounting for Adjustment Uncertainty," *Biometrics*, 68, 661–671. [507]

White, H., and Chalak, K. (2013), "Identification and Identification Failure for Treatment Effects Using Structural Systems," *Econometric Reviews*, 32, 273–317. [511]

White, H., and Lu, X. (2011), "Causal Diagrams for Treatment Effect Estimation with Application to Efficient Covariate Selection," *Review of Economics and Statistics*, 93, 1453–1459. [506,510]

Wooldridge, J. (2005), "Violating Ignorability of Treatment by Controlling for Too Many Factors," *Econometric Theory*, 21, 1026–1029. [506,510]

Zhang, X., and Liang, H. (2011), "Focused Information Criterion and Model Averaging for Generalized Additive Partial Linear Models," *The Annals of Statistics*, 39, 174–200. [512]

Zigler, C. M., and Dominici, F. (2014), "Uncertainty in Propensity Score Estimation: Bayesian Methods for Variable Selection and Model-Averaged Causal Effects," *Journal of the American Statistical Association*, 109, 95–107. [507]