BOOTSTRAP INFERENCE FOR PROPENSITY SCORE MATCHING

KARUN ADUSUMILLI

Abstract. Propensity score matching, where the propensity scores are estimated in a first step, is widely used for estimating treatment effects. In this context, the naive bootstrap is invalid (Abadie and Imbens, 2008). This paper proposes a novel bootstrap procedure for the propensity score matching estimator, and demonstrates its consistency. The proposed bootstrap is built around the notion of 'potential errors', introduced in this paper. Precisely, each observation is associated with two potential error terms, corresponding to each of the potential states - treated or control - only one of which is realized. Thus, the variability of the estimator stems not only from the randomness of the potential errors themselves, but also from the probabilistic nature of treatment assignment, which randomly realizes one of the potential error terms. The proposed bootstrap takes both sources of randomness into account by resampling the potential errors as a pair as well as re-assigning new values for the treatments. Simulations and real data examples demonstrate the superior performance of the proposed method relative to using the asymptotic distribution for inference, especially when the degree of overlap in propensity scores is poor. General versions of the procedure can also be applied to other causal effect estimators such as inverse probability weighting and propensity score sub-classification, potentially leading to higher order refinements for inference in such contexts.

This version: August 12, 2018. Latest version here.
I am greatly indebted to Taisuke Otsu for invaluable help on this project. I would also like to thank Hao Dong, Dita Eckardt, Kirill Evdokimov, Sokbae (Simon) Lee, Will Matcham, Chris Muris, Jörn-Steffen Pischke, Shan Aman Rana, Peter M. Robinson, Nelson A. Ruiz, Marcia Schafgans, Sorawoot Srisuma, Luke Taylor, Takuya Ura, and seminar participants at the Bristol Econometrics Study Group, Boston University, Georgetown University, Iowa State University, LSE, UC Berkeley, UC San Diego, University of Illinois - Urbana Champaign and University of Pennsylvania for helpful comments.

1. Introduction

Inference on average treatment effects in the presence of confounding is a primary goal of many observational studies. Propensity Score Matching (PSM) is one of the most widely used methods for estimating treatment effects in such a setting. The propensity score is defined as the probability of obtaining treatment conditional on covariates. Under the assumption of selection on observables (i.e the treatment is as good as randomly assigned conditional on the covariates), Rosenbaum and Rubin (1983) show that matching on the propensity score is sufficient to remove confounding. Using the propensity score for matching reduces the dimensionality of the procedure by summarizing the information contained in the covariates in a single variable. Additionally, PSM can be flexibly combined with other strategies such as regression adjustment to further reduce the bias from the match (Abadie and Imbens, 2011; Imbens and Rubin, 2015). Such favourable properties have led to PSM becoming one of most commonly used methods for causal analysis of observational data. See for example Deheija and Wahba (1999), Heckman, Ichimura, Smith and Todd (1998), Lechner (2002) and Smith and Todd (2001) for some important applications and issues arising from its use in economics.

In practice, the propensity scores are usually estimated through a parametric first stage based on a probit or logit model. Furthermore, to reduce the bias from the match, the number of matches is usually held fixed at small values, for example one. This introduces complications for inference since the matching function - defined as the number of times each unit is used as a match - is a highly non-linear function of the data. Abadie and Imbens (2016) show that the matching estimator under the estimated propensity score is consistent and asymptotically normal. Thus inference for the treatment effect can proceed based on a large sample approximation to the normal distribution, using the variance estimate suggested by the authors. At the same time, Abadie and Imbens (2008) show that the standard non-parametric bootstrap based on resampling fails to be consistent in this context. This is because the usual bootstrap procedure fails to reproduce the distribution of the matching function in the true sample.

In this paper, I propose and demonstrate consistency of a bootstrap procedure for matching on the estimated propensity score. Both matching with and without replacement is considered. The proposed bootstrap is built around the concept of 'potential errors', introduced in this paper as a general tool for causal inference. Potential errors formalize the idea that each observation can be associated with two possible error terms, corresponding to each of the potential states - treated or control - only one of which is actually realized. Thus, the variability of the estimator stems not only from the randomness of the potential errors themselves, but also from the probabilistic nature of treatment assignment, which randomly realizes one of the potential error terms. The proposed bootstrap takes both sources of randomness into account by resampling the potential errors as a pair, while also re-assigning new values for the treatments using the estimated propensity score. Implementing the procedure requires the construction of estimates of the error terms under both states. Since I only observe the errors under one of the potential states for any data point, I provide ways to impute these quantities for the other state.

The notion of potential errors is very general, and can be applied to causal effect estimators beyond propensity score matching. The exact form of the potential errors depends on both the estimator and the quantity being estimated (ATE, ATET, etc.), but a unifying theme is that it is possible to obtain the 'error representation'[1]

$$\text{Estimator} - \text{Expected Value} = \text{Average(Realized errors)}.$$

Here, the terminology 'realized errors' refers to the observed values of the potential errors given the treatment status. For many estimators, directly resampling the realized errors suffices for valid inference, see e.g. Otsu and Rai (2017). However, such a strategy doesn't work for propensity score matching since the potential errors are functions of the estimated propensity score, which is itself a random quantity (see, Section 3.4). Taking the estimation of the propensity scores into account requires recreating the randomness of treatment assignment closely, since this determines the variability of the propensity scores. Doing so naturally leads to the proposed bootstrap statistic. Indeed, my bootstrap statistic is simply the average of the new realized errors - obtained after resampling the potential errors and reassigning treatments - and evaluated at propensity scores estimated from the bootstrap sample.

The proposed bootstrap can be easily extended to other causal effect estimators satisfying the error representation, for example inverse probability weighting or propensity score subclassification (see, Section 6.3). Since it recreates all the sources of randomness more faithfully, it generally provides more precise inference compared to asymptotic methods or methods that only resample the realized errors. The gain in accuracy is especially pronounced when there is poor overlap between the propensity scores of the treated and control groups. Poor overlap usually occurs when there is heavy imbalance between the covariate distributions for the treated and control groups. In such situations, some observations gain disproportionate importance, for instance the few control units close to the treated units, and vice versa. The resulting causal estimate is then highly sensitive to possible switches to the treatment status of these observations. Failure to take this into account leads to severe under-estimation of the actual variance, as shown in simulations. By contrast, the proposed bootstrap is more accurate, and constitutes an attractive choice for inference when the overlap is poor.

I demonstrate consistency of this bootstrap procedure using Le Cam's framework of local limit experiments, applied on the bootstrap data generating process. To this end, I extend the techniques of local limit experiments previously employed by Abadie and Imbens (2016), and Andreou and Werker (2011) to obtain limiting distributions of non-smooth statistics to the setup of bootstrap inference. Thus, the techniques may be of independent theoretical interest.

The finite sample performance of the bootstrap is assessed through a number of simulations and real data examples. In almost all cases the bootstrap provides better size control than inference based on the asymptotic distribution. The results also confirm that the proposed bootstrap is particularly effective when the balance of covariates across treated and control samples is poor. Arguably, poor covariate balance is pervasive in observational studies.

[1]For matching estimators, this is equivalent to the martingale representation of Abadie and Imbens (2012).

The theoretical results in this paper build on the properties of matching estimators with finite number of matches, established in an important series of papers by Abadie and Imbens (2006, 2008, 2011, 2012, 2016). When the number of matches is allowed to increase with sample size, as in the kernel matching method of Heckman, Smith and Todd (1997), the resulting estimator is asymptotically linear, and the usual non-parametric bootstrap can be employed. In the context of a fixed number of matches, Otsu and Rai (2016) propose a consistent bootstrap method for the version of nearest neighbor matching based on a distance measure (Euclidean, Mahalanobis etc.) over the full vector of covariates. The proposal of Otsu and Rai (2016) is equivalent to conditioning on both treatments and covariates, and resampling the realized errors in the error representation. However, their consistency result doesn't extend to propensity score matching because conditioning on both treatments and covariates precludes taking into account the effect of the estimation of propensity scores. Alternatives to the bootstrap that do provide consistent inference in this context include subsampling (Politis and Romano, 1994) and m-out-of-n bootstrap (Bickel, Götze and van Zwet, 2012).

2. Setup

The starting point of my analysis is the standard treatment effect model under selection on observables. I follow the same setup as Abadie and Imbens (2016). The aim is to estimate the effect of a binary treatment, denoted by W, on some outcome Y. A value of $W = 1$ implies the subject is treated, while $W = 0$ implies the subject hasn't received any treatment. The causal effect of the treatment is represented in the terminology of potential outcomes (Rubin, 1974). In particular, I introduce the random variables $(Y(0), Y(1))$, where $Y(0)$ denotes the potential outcome under no treatment, and $Y(1)$ denotes the potential outcome under treatment. I also have access to a set of covariates X, where $\dim(X) = k$. The goal is to estimate the average treatment effect

$$\tau = E[Y(1) - Y(0)].$$

In general, estimation of τ suffers from a missing data problem since only one of the potential outcomes is observable as the actual outcome variable, $Y = Y(W)$. To circumvent this, practitioners commonly impose the following identifying assumptions for τ:

Assumption 1. $(Y(1), Y(0))$ *is independent of* W *conditional on* X *almost surely, denoted as* $(Y(1), Y(0)) \perp\!\!\!\perp W \mid X$.

Assumption 2. (Y_i, W_i, X_i) *are i.i.d draws from the distribution of* (Y, W, X).

The first assumption is that of unconfoundedness, which implies that the treatment is as good as randomly assigned conditional on the covariates X. The second assumption implies that the potential outcome for individual i is independent of the treatment status and covariates of the other individuals. This rules out peer effects, for instance.

Define the propensity score, $p(X) = \Pr(W = 1|X)$, as the probability of being treated conditional on the covariates. Let $\bar{\mu}(w, X)$ and $\mu(w, p(X))$ denote the conditional means $E[Y|W = w, X]$ and $E[Y|W = w, p(X)]$ respectively. Additionally, let $\bar{\sigma}^2(w, X) = E[Y^2|W = w, X]$ and

$\sigma^2(w, p(X)) = E[Y^2|W = w, p(X)]$ denote the conditional variances of Y given $W = w$ and X ; and that of Y given $W = w$ and $p(X)$ respectively. In a seminal paper, Rosenbaum and Rubin (1983) show that under Assumption 1, the potential outcomes are also independent of the treatment conditional on the propensity scores, i.e $(Y(1), Y(0)) \perp\!\!\!\perp W \ |p(X)$. Thus, τ can be alternatively identified as

$$\tau = E[\mu(1, p(X)) - \mu(0, p(X))].$$

In the literature a number of propensity score matching techniques have been proposed that exploit the above characterization of τ, see e.g. Rosenbaum (2009) for a detailed survey. In this section, and for much of this paper, I focus on matching with replacement, with a fixed number of matches for each unit, denoted by M. This is arguably the most commonly used matching procedure in economic applications. The case of matching without replacement is discussed in Section 6.2.

Suppose that I have a sample of N observations. The propensity score matching estimator for the average treatment effect, when matching with replacement, is defined as

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}(2W_i - 1)\left(Y_i - \frac{1}{M}\sum_{j \in \mathcal{J}_M(i;p(X))} Y_j\right),$$

where M is the number of matches for each unit, and $\mathcal{J}_M(i; p(X))$ is the set of matches for the individual i. In particular $\mathcal{J}_M(i; p(X))$ represents the set of M individuals from the opposite treatment arm whose propensity scores are closest to i's own, i.e,

$$\mathcal{J}_M(i; p(X)) = \{j = 1, \ldots, N : \ W_j = 1 - W_i, \text{ and } \left(\sum_{l: W_l = 1 - W_i} \mathbb{I}_{[|p(X_i) - p(X_l)| \leq |p(X_i) - p(X_j)|]}\right) \leq M\Bigg\}.$$

Typically the value of M is taken to be quite small, for example $M = 1$, so as to reduce the bias.

The propensity scores are generally not known but have to be estimated. In this paper, I consider parametric estimates for the propensity scores based on a generalized linear model $p(X) = F(X'\theta)$, where θ is a finite dimensional vector parameter, and $F(.)$ is a (known) link function, for instance a logistic or probit function.[2] Let $(\mathbf{W}, \mathbf{X})$ denote the vector of treatments and covariates $(W_1, \ldots, W_N, X_1, \ldots, X_N)$. I denote the true value of θ by θ_0. The latter is estimated through maximum likelihood as

$$\hat{\theta} = \arg\max_{\theta} L(\theta|\mathbf{W}, \mathbf{X}),$$

where

$$L(\theta|\mathbf{W}, \mathbf{X}) = \sum_{i=1}^{N}\left\{W_i \ln F(X_i'\theta) + (1 - W_i)\ln(1 - F(X_i'\theta))\right\},$$

denotes the log-likelihood function evaluated at θ.

[2] I assume that the propensity score is correctly specified. To guard against mis-specification, one could employ an algorithmic procedure for choosing the propensity score that iterates between a PSM specification and balance checking; see, for example, Deheija and Wahba (1999).

Let $\mathcal{J}_M(i;\theta)$ denote the set of M closest matches to observation i for the match based on $F(X'\theta)$, for any given θ, i.e

$$\mathcal{J}_M(i;\theta) = \{j = 1,\ldots,N:\ W_j = 1 - W_i,\ \text{and} \\ \left(\sum_{l:W_l=1-W_i} \mathbb{I}_{[|F(X_i'\theta)-F(X_l'\theta)|\leq|F(X_i'\theta)-F(X_j'\theta)|]}\right) \leq M\Bigg\}.$$

The matching estimator, for the match based on $F(X'\theta)$, is defined as

$$\hat{\tau}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(2W_i - 1)\left(Y_i - \frac{1}{M}\sum_{j\in\mathcal{J}_M(i;\theta)} Y_j\right).$$

Let $K_M(i;\theta)$ denote the number of times observation i is used as a match based on $F(X'\theta)$, i.e

$$K_M(i;\theta) = \sum_{j=1}^{N}\mathbb{I}_{i\in\mathcal{J}_M(j;\theta)}.$$

Then an alternative way to represent $\hat{\tau}(\theta)$ is provided by the error representation

$$\hat{\tau}(\theta) - \tau - B(\theta) = \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i(W_i;\theta), \tag{2.1}$$

where

$$B(\theta) = \frac{1}{N}\sum_{i=1}^{N}(2W_i - 1)\cdot\left(\mu(1-W_i, F(X_i'\theta)) - \frac{1}{M}\sum_{i\in\mathcal{J}_M(i;\theta)}\mu(1-W_i, F(X_i'\theta))\right)$$

denotes the bias from the match based on $F(X'\theta)$, and

$$\varepsilon_i(W_i;\theta) = \left(\mu\left(1, F(X_i'\theta)\right) - \mu\left(0, F(X_i'\theta)\right) - \tau\right) \\ + (2W_i - 1)\left(1 + \frac{K_M(i;\theta)}{M}\right)\left(Y_i - \mu\left(W_i, F(X_i'\theta)\right)\right) \tag{2.2}$$

denotes the effective error term for each observation. The variance is thus determined by the right hand side of equation (2.1). Consequently, this expression is of primary interest in approximating the distribution of $\hat{\tau}(\theta)$.

The matching estimator for τ based on the estimated propensity score is then given by

$$\hat{\tau} \equiv \hat{\tau}(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}(2W_i - 1)\left(Y_i - \frac{1}{M}\sum_{j\in\mathcal{J}_M(i;\hat{\theta})} Y_j\right).$$

Abadie and Imbens (2016) derive the large sample properties of the above estimator. Under some regularity conditions, they find that the bias term $B(\hat{\theta})$ converges in probability to zero at a rate faster than $\sqrt{N}$, and that $\hat{\tau}$ has an asymptotic normal distribution

$$\sqrt{N}(\hat{\tau} - \tau) \xrightarrow{d} N\left(0, \sigma^2 - c' I_{\theta_0}^{-1} c\right),$$

where σ^2 is the asymptotic variance for matching on the known propensity score,

$$c = E\left[\left\{\frac{\text{cov}[X, \mu(1,X)|F(X'\theta_0)]}{F(X'\theta_0)} + \frac{\text{cov}[X, \mu(0,X)|F(X'\theta_0)]}{1 - F(X'\theta_0)}\right\} f(X'\theta_0)\right]$$

with $f(.) \equiv F'(.)$; and for any value of θ, $I(\theta)$ denotes the information matrix evaluated at θ

$$I_\theta \equiv I(\theta) = E\left[\frac{f^2(X'\theta)}{F(X'\theta)(1-F(X'\theta))}XX'\right].$$

The above result illustrates the well known 'Propensity Score Paradox': Matching on the estimated, as opposed to the true propensity scores, in fact reduces the asymptotic variance.

3. Bootstrap procedure

In this section I propose a bootstrap procedure for inference on the propensity score matching estimator. I fix the following notation: For each $w = 0, 1$, define $\mu(w, p; \theta) = E[Y(w)|F(X_i'\theta) = p]$. In what follows, I abuse notation a bit by dropping the index of $\mu(., .; \theta)$ with respect to θ when the context is clear. For $w = 0, 1$, denote[3]

$$\begin{aligned} e_{1i}(\theta) &= \mu(1, F(X_i'\theta)) - \mu(0, F(X_i'\theta)) - \tau; \\ e_{2i}(w; \theta) &= Y_i - \mu(w, F(X_i'\theta)). \end{aligned}$$

Note that the above are distinct in general from the 'true' errors which are defined similarly but evaluated at θ_0.

I present here an informal description of the bootstrap procedure, relegating many of the formal details to the upcoming sub-sections. Given any value of θ, the pair of potential error terms for each observation i are given by

$$\varepsilon_i(w; \theta) \equiv e_{1i}(\theta) + (2w - 1)\left(1 + \frac{\tilde{K}_M(i; w, \theta)}{M}\right) e_{2i}(w; \theta);\ w = 0, 1,$$

where $\tilde{K}_M(i; w, \theta)$ is a potential matching function, denoting the number of times observation i would have been used as a match depending on whether it is in the treated ($w = 1$) or control group ($w = 0$); see Section (3.2) for the formal definition of $\tilde{K}_M(i; w, \theta)$. Clearly, only one of the quantities $\varepsilon_i(w; \theta) : w = 0, 1$ is directly estimable; the other has to be imputed. Let $\hat{\varepsilon}_i(w; \theta)$ denote the estimated or imputed values of $\varepsilon_i(w; \theta)$. I then sample a set of N covariates denoted by X_j^* for $j = 1, \ldots, N$, along with the associated pair of (estimated) potential error terms $\left(\hat{\varepsilon}_{S_j^*}(0; \theta), \hat{\varepsilon}_{S_j^*}(1; \theta)\right)$, where S_j^* denotes the bootstrap index corresponding to the j-th observation in the draw. Subsequently, new bootstrap treatment values are generated using the estimated propensity scores as

$$W_j^* \sim \text{Bernoulli}(F(X_j^{*\prime}\hat{\theta})).$$

Through this procedure I have sampled a new set of realized error terms given by $\varepsilon_j^*(\theta) \equiv \hat{\varepsilon}_{S_j^*}(W_j^*; \theta)$ for $j = 1, \ldots, N$. The bootstrap statistic, $T_N^*\left(\hat{\theta}^*\right)$, is the sample average of these errors, after some appropriate recentering using the function $\Xi^*\left(\hat{\theta}^*\right)$ [4], i.e

$$T_N^*\left(\hat{\theta}^*\right) \equiv \frac{1}{\sqrt{N}}\sum_{j=1}^{N}\left\{\hat{\varepsilon}_{S_j^*}\left(W_j^*; \hat{\theta}^*\right) - \Xi^*\left(\hat{\theta}^*\right)\right\}.$$

[3] I do not index τ with θ since the average treatment effect is independent of the propensity score.
[4] The prcise expression for the re-centering term $\Xi^*(.)$ is provided in Section 3.3.

The errors above are being evaluated at $\hat{\theta}^*$- the bootstrap counterpart of $\hat{\theta}$ - obtained as

$$\hat{\theta}^* = \arg\max_{\theta} L(\theta|\mathbf{W}^*, \mathbf{X}^*).$$

Note that except for a negligible bias term $B(\hat{\theta})$, the construction of the bootstrap statistic closely mirrors the error representation for $\hat{\tau}(\hat{\theta}) - \tau$ given by

$$\hat{\tau}(\hat{\theta}) - \tau - B(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i(W_i; \hat{\theta}).$$

To formalize the above, I require techniques for: (i) constructing estimates of the error terms, $e_{1i}(\theta), e_{2i}(w;\theta)$, for each observation under both treated and control states; and (ii) constructing the potential matching function, $\tilde{K}_M(i; w, \theta)$, for each observation, also under both states. I now consider these in turn.

3.1. **Constructing estimates of error terms.** Denote by $\hat{\mu}(w, F(X_i'\theta))$ the estimates of the conditional expectation function $\mu(w, F(X_i'\theta))$ evaluated at $F(X_i'\theta)$. These can be obtained through non-parametric methods, for example series regression or smoothing splines. I then obtain the residuals

$$\begin{aligned} \hat{e}_{1i}(\theta) &= \hat{\mu}(1, F(X_i'\theta)) - \hat{\mu}(0, F(X_i'\theta)) - \hat{\tau}(\theta); \\ \hat{e}_{2i}(W_i;\theta) &= Y_i - \hat{\mu}(W_i, F(X_i'\theta)). \end{aligned}$$

These residuals serve as proxies for the unobserved terms $e_{1i}(\theta), e_{2i}(W_i;\theta)$, approximating the values of $e_{2i}(w;\theta)$ when $w = W_i$. For the bootstrap procedure, I also need estimates of $\hat{e}_i(w;\theta)$ when $w \neq W_i$. I obtain these through a secondary matching: Define the secondary matching function as

$$\mathcal{J}_w(i) = \begin{cases} i & \text{if } W_i = w \\ \mathcal{J}_{\text{NN}}(i) & \text{if } W_i \neq w, \end{cases}$$

where $\mathcal{J}_{\text{NN}}(.)$ denotes the closest match (or nearest neighbor) to observation i from the opposite treatment arm, with the closeness measured in terms of a distance metric (Euclidean, Mahalanobis etc.) based on the full set of covariates. I then obtain:

$$\hat{e}_{2i}(w;\theta) = \hat{e}_{2\mathcal{J}_w(i)}(w;\theta).$$

The definition of $e_{2i}(w;\theta)$ proceeds in an analogous fashion.

Note that the secondary matching procedure matches on the full set of covariates, as opposed to matching on the propensity scores. This is done to preserve the conditional correlation between X and the error terms e_{1i}, e_{2i}, given the propensity scores. Indeed it is this correlation that helps drive down the asymptotic variance when using the estimated propensity score.

3.2. **Constructing the matching function.** As with the error terms, the bootstrap procedure requires values of the matching function under both treatment and non-treatment, even as only one of them is actually observed. To obtain the value of $\tilde{K}_M(i; w, \theta)$ in the opposite treatment arm (i.e when $w \neq W_i$), I employ another imputation procedure:

Let $\{\pi_1, \ldots, \pi_{q_N-1}\}$ denote the sample q_N-quantiles of $F(X'\hat{\theta})$. I let $q_N \to \infty$ as $N \to \infty$. Set $\pi_0 = 0$ and $\pi_{q_N} = 1$. Denote by $S_w(l)$, the set of all observations with $W_i = w$ in the l-th block, i.e

$$S_w(l) = \{i : \ \pi_{l-1} \leq F(X_i'\hat{\theta}) < \pi_l \cap W_i = w\},$$

and let $S(l) = S_1(l) \cup S_0(l)$. The number of untreated, treated and combined observations in the block l is given by

$$N_0(l) = \#S_0(l); \quad N_1(l) = \#S_1(l); \quad N(l) = N_0(l) + N_1(l),$$

respectively, where for any set A, $\#A$ denotes its cardinality. Suppose now that observation i falls in the block l. If $w = W_i$, I set $\tilde{K}_M(i; w, \theta) = K_M(i; \theta)$. If however $w \neq W_i$, I set $\tilde{K}_M(i; w, \theta)$ to the value $K_M(j; \theta)$, where j is drawn at random from the $S_w(l)$. Formally, denoting by $l(i)$ the block in which observation i resides, I obtain

$$\tilde{K}_M(i; w, \theta) = \begin{cases} K_M(i; \theta) & \text{if } w = W_i \\ \sum_{j \in S_w(l(i))} \{M_j(i) K_M(j; \theta)\} & \text{if } w \neq W_i, \end{cases}$$

where for each i, $\{M_j(i) : j \in S_w(l(i))\} \equiv \mathbf{M}(i)$ is a multinomial random vector with a single draw on $N_w(l(i))$ equal probability cells. These multinomial random variables are drawn independently for each observation i.

Based on these constructions I can define a combined error term excluding the effect of heterogeneity (i.e excluding $e_{1i}(\theta)$) as

$$\hat{\nu}_i(w; \theta) = \left(1 + \frac{\tilde{K}_M(i; w, \theta)}{M}\right) \hat{e}_{2\mathcal{J}_w(i)}(w; \theta).$$

Thus the estimated potential errors are obtained as

$$\hat{\varepsilon}_i(w; \theta) = \hat{e}_{1i}(\theta) + (2w - 1)\hat{\nu}_i(w; \theta).$$

Note that for a given θ, the potential errors and the matching functions used therein only depend on the original set of observations $(\mathbf{W}, \mathbf{X})$ and not the bootstrap draws.

Remark 1. Unlike the error terms, the values of $K_M(i; \theta)$ cannot be imputed through nearest neighbor matching. Doing so renders the bootstrap inconsistent since $K_M(i; \theta)$ and $K_{\text{NN}}(i)$ are correlated (here, $K_{\text{NN}}(i)$ denotes the number of times observation i is used as a match, when closeness is measured in terms of a distance metric on the full set of covariates). Intuitively, a nearest neighbor based imputation over-selects observations that are already matched often, and hence fails to recreate the actual distribution of the matching function. A similar comment also applies to imputing the values through propensity score matching.

Remark 2. Let $\mathcal{F}_K^{(0)}(.)$ and $\mathcal{F}_K^{(1)}(.)$ denote the conditional distribution functions of $K_M(i; \theta)$ for the control and treated groups, given the own-propensity score $F(X_i'\theta)$. Consider the estimator, $\hat{\mathcal{F}}_K^{(w)}$, of $\mathcal{F}_K^{(w)}$ obtained by coarsening/blocking the propensity scores, and using the empirical distribution of $K_M(i; \theta)$ for $w = 0, 1$ within each block. The procedure described in this section is equivalent to drawing a value from the distribution $\hat{\mathcal{F}}_K^{(w)}(F(X_i'\theta))$, independently for each

i, and using it to impute the value of $\tilde{K}_M(i; w, \theta)$ when $w \neq W_i$. Coarsening is motivated by the fact $K_M(i; \theta)$ takes discrete values, which precludes smoothing. Clearly $\hat{\mathcal{F}}_K^{(w)} \equiv \mathcal{F}_K^{(w)}$ if the propensity scores are constant within the blocks. More generally, $\hat{\mathcal{F}}_K^{(w)}$ approaches $\mathcal{F}_K^{(w)}$ as $N \to \infty$ since I let $q_N \to \infty$. The optimal choice of q_N would minimize the variability in propensity scores within blocks while ensuring enough observations in each, thereby estimating $\mathcal{F}_K^{(w)}$ more accurately.

Sampling from $\hat{\mathcal{F}}_K^{(w)}$ also ensures each $K_M(i; \theta)$, for $i = 1, \ldots, N$, is used almost exactly once, on average, in the bootstrap: the term may drop out because $W_i^* \neq W_i$, but this probability is balanced by the number of times it may be used for imputations (for details, see Appendix B). Thus, the original set of matching functions is well reproduced in the bootstrap.

Remark 3. The variables $\mathbf{M} \equiv \{\mathbf{M}(i) : 1 \leq i \leq N\}$ do not enter the bootstrap distribution as the particular realization of $\mathbf{M}$ is fixed throughout the bootstrap procedure. This is equivalent to fixing an observation j that imputes for i in all the bootstrap draws. Thus the bootstrap distribution should be understood as conditional on both $\mathbf{M}$ and the observed data. This necessarily injects some randomness into the critical values obtained from the bootstrap (though the critical values do converge to the true ones almost surely for each sequence $\mathbf{M}$). To address this, I suggest repeating the bootstrap procedure for a number of different realizations of $\mathbf{M}$, and then taking an average (wrt $\mathbf{M}$) of the bootstrap distribution functions; see below.

3.3. **The bootstrap algorithm.** The bootstrap algorithm proceeds as follows.

Step 0: First obtain a set of multinomial probabilities $\mathbf{M}$ based on independent draws for each individual i as described in Section 3.2. Additionally calculate the nearest neighbor matching function $\mathcal{J}_w(i)$ for each i as defined in Section 3.1. Both these values are kept fixed throughout the bootstrap.

Step 1: Obtain new values of covariates $\mathbf{X}^* = (X_i^*, \ldots, X_N^*)$ through a non-parametric bootstrap draw. This involves drawing N independent categorical random variables $\mathbf{S}^* = (S_1^*, \ldots, S_N^*)$.

Step 2: Based on the estimated propensity score, derive new treatment values $\mathbf{W}^* = (W_1^*, \ldots, W_N^*)$ through the random draws

$$W_i^* \sim \text{Bernoulli}(F(X_i^{*\prime}\hat{\theta})).$$

Step 3: Discard bootstrap samples for which $N_0^* \leq M + 1$ or $N_1^* \leq M + 1$, where N_0^* and N_1^* denote the number of control and treated observations in the bootstrap sample. For all the other samples, estimate the bootstrap statistic $\hat{\theta}^*$ using the MLE procedure on $(\mathbf{W}^*, \mathbf{X}^*)$

$$\hat{\theta}^* = \arg\max_{\theta} L(\theta | \mathbf{W}^*, \mathbf{X}^*).$$

Step 4: Based on $\hat{\theta}^*$, obtain the values of matching function $K_M(i; \hat{\theta}^*)$ for each i using the original sample of observations $\mathbf{W}, \mathbf{X}$. Additionally, derive the residuals $\left(\hat{e}_{1i}(\hat{\theta}^*), \hat{e}_{2i}(W_i; \hat{\theta}^*)\right)$, evaluated at $\hat{\theta}^*$, for each i through series regression (or any other nonparametric method) applied on the original sample of observations. From these, along with the values of $\mathbf{M}$ and $\mathcal{J}_w(i)$ from Step 0, determine the values of $\tilde{K}_M(i; w, \hat{\theta}^*)$ and $\hat{\nu}_i(w; \hat{\theta}^*)$ for $i = 1, \ldots, N$ by following the procedures laid down in Sections 3.1. and 3.2.

For the remaining steps, define the new 'bootstrap' realized errors $\varepsilon_i^*(\theta)$ as

$$\begin{aligned}\varepsilon_i^*(\theta) &\equiv \hat{\varepsilon}_{S_j^*}(W_j^*;\theta) \\ &= \hat{e}_{1S_i^*}(\theta) + W_i^*\hat{\nu}_{S_i^*}(1;\theta) - (1-W_i^*)\hat{\nu}_{S_i^*}(0;\theta).\end{aligned}$$

The bootstrap errors $\varepsilon_i^*(\theta)$ need to be re-centered; the expression for this is given by

$$\Xi^*(\theta) = \frac{1}{N}\sum_{k=1}^{N}\left\{\hat{e}_{1k}(\theta) + F(X_k'\theta)\hat{\nu}_k(1;\theta) - \left(1-F(X_k'\theta)\right)\hat{\nu}_k(0;\theta)\right\}.$$

Note that $\Xi^*(\theta) \equiv E_\theta^*[\varepsilon_i^*(\theta)]$, where $E_\theta^*[.]$ denotes the expectation over the probability distribution implied by $\mathbf{S}^*, \mathbf{W}^* \sim \text{Bernoulli}(F(\mathbf{X}^*\theta))$, conditional on the original data (see also Section 3.4 for a detailed explanation). Finally, for each value of θ, define the bootstrap statistic

$$T_N^*(\theta) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{\varepsilon_i^*(\theta) - \Xi^*(\theta)\right\}.$$

Step 5: Evaluate $T_N^*(\theta)$ at the parameter value $\hat{\theta}^*$ to obtain the bootstrap statistic $T_N^*\left(\hat{\theta}^*\right)$. This step utilizes the values of $\tilde{K}_M(i;w,\hat{\theta}^*)$ and $\hat{\nu}_i(w;\hat{\theta}^*)$ obtained in Step 4.

Step 6: Estimate the critical value by $c_{n,\alpha}^* = \inf\{t : F_n^*(t) \geq 1-\alpha\}$, where $F_n^*(.)$ is the empirical distribution of $T_N^*\left(\hat{\theta}^*\right)$. This can be obtained by repeating Steps 1-5 for a set of B bootstrap repetitions.

Step 7: The critical value, $c_{n,\alpha}^*$, in Step 6 is based on a particular realization of $\mathbf{M}$. To reduce the dependence on the latter, repeat Steps 1-6 for L different values of $\mathbf{M}$ and average the resulting empirical distribution functions $F_n^*(.)$ to obtain $\bar{F}_n^*(.)$. The final estimated critical value is then given by $\bar{c}_{n,\alpha}^* = \inf\{t : \bar{F}_n^*(t) \geq 1-\alpha\}$.

3.4. **Discussion.** This section elaborates further on key aspects of the bootstrap procedure.

3.4.1. *Asymptotic Linearity.* Efron and Stein (1981) have shown that an estimator typically needs to be asymptotically linear in the observations for the standard (nonparametric) bootstrap to be valid. However, the matching estimator fails to satisfy asymptotic linearity under the regime of fixed number of matches. Indeed, fixing the number of matches is qualitatively similar to choosing 'small bandwidth asymptotics' for semiparametric estimators, wherein it is known that asymptotic linearity fails (see, e.g Cattaneo, Jansson & Newey, 2016). The same reasoning also implies the standard bootstrap is invalid for the kernel matching estimator of Heckman, Smith and Todd (1997) under small bandwidth asymptotics. Nevertheless, while the matching estimators are not generally asymptotically linear in the observations $(\mathbf{X}, \mathbf{W}, \mathbf{Y})$, they *are* linear in the potential errors, by construction. Thus, by changing the unit of resampling to potential errors (rather than the observations), we can regain bootstrap consistency.

3.4.2. *Randomization of treatments.* A distinctive feature of the bootstrap procedure is the randomization of the treatments, $\mathbf{W}^*$. For many causal effect estimators, such as nearest neighbor matching using the vector of covariates, it suffices to resample the realized errors (see, e.g. Otsu and Rai, 2017). However, such a strategy doesn't work for propensity score matching because the potential and realized errors are functions of the random quantity $F(\mathbf{X}'\hat{\theta})$. The variability of

$\mathbf{W}$ conditional on $\mathbf{X}$ has a first order effect on inference through the estimation of $\hat{\theta}$, necessitating the re-drawing of $\mathbf{W}^*$ in the bootstrap. The precise mechanism is as follows: Suppose that one of the covariates is heavily imbalanced between the treatment and control groups. Then the magnitude of $\hat{\theta}$ corresponding to the covariate increases, and the procedure places greater emphasis on balancing that covariate. This reduces the conditional (on $\mathbf{X}, \mathbf{W}$) bias, eventually showing up as (unconditional) asymptotic variance reduction, see Section 2. But for a fixed $\mathbf{X}$, the level of imbalance depends on the assignment of $\mathbf{W}$; hence the conditional distribution of $\mathbf{W}$ given $\mathbf{X}$ has a large effect on the variability of the estimate.

3.4.3. *Bootstrap Recentering.* An interesting feature of the recentering term, $\Xi^*(\theta)$, is that it is based on taking the bootstrap expectation over $T_N^*(\theta)$ as if $\mathbf{W}^* \sim \text{Bernoulli}(F(\mathbf{X}^*\theta))$, even though in fact $\mathbf{W}^* \sim \text{Bernoulli}(F(\mathbf{X}^*\hat{\theta}))$. If θ were an exogenous parameter, this would mean the bootstrap expectation of $T_N^*(\theta)$ is exactly 0 only when $\theta = \hat{\theta}$. However $T_N^*()$ is evaluated at $\hat{\theta}^*$, itself a function of the bootstrap random variables. In this case the precise form of the recentering ensures $T_N^*\left(\hat{\theta}^*\right)$ converges in distribution to a mean zero random variable. The reasoning is broadly as follows (see proof of Theorem 1 for details): Suppose for the sake of argument that $\mathbf{W}^* \sim \text{Bernoulli}(F(\mathbf{X}^*\theta))$. Together with $\mathbf{S}^*$, this parametrizes the bootstrap probability distribution, denoted by P_θ^*. Under P_θ^* the test statistic $T_N^*(\theta)$ is exactly mean 0, and therefore converges to a mean 0 random variable in distribution. However for values of θ that are sufficiently close to $\hat{\theta}$, such as $\hat{\theta}^*$, the probability distributions P_θ^* and $P_{\hat{\theta}}^*$ largely coincide. Hence $T_N^*(\theta)$ also converges to a mean 0 random variable under $P_{\hat{\theta}}^*$ - the actual bootstrap distribution.

3.4.4. *Imputation.* The imputation step, Step 0, is critical to the bootstrap. Here, prior to the bootstrap draws, each observation is linked with two others from the opposite treatment arm: the first for imputing the errors (cf Section 3.1), and the second for imputing the matching functions (cf Section 3.2). In general these observations do not coincide. However conditional on the propensity score, the variables $K_M(i;\theta), e_{1i}, e_{2i}$ are independent of each other even in the true DGP. Thus the approximation properties of the bootstrap are not adversely affected.

Alternatively, one may choose to sort the observations into blocks based on the full set of covariates rather than the propensity scores as in Section 3.2. Then a single observation, drawn at random from the block, can be used to impute both the errors and the matching functions. However, even with a binary categorization of the covariates, the number of blocks increases as 2^k with the dimension k. Hence even for moderate k (e.g. $k \geq 5$), it is highly likely that many of the blocks only contain observations from a single treatment arm.

4. Asymptotic properties

In this section, I derive the asymptotic properties of the bootstrap procedure outlined in Section 3.1, and demonstrate its consistency. Let P_θ denote the joint distribution of $\{\mathbf{Y}, \mathbf{W}, \mathbf{X}\}$ implied by $W \sim \text{Bernoulli}(F(X_i'\theta))$, the marginal distribution of X, and the conditional distribution of Y given W, X. The corresponding expectation over P_θ is denoted by $E_\theta[.]$. Also, denote by $\tilde{P}_\theta$ the joint probability distribution over both $\{\mathbf{Y}, \mathbf{W}, \mathbf{X}\}$ and $\mathbf{M}$; with $\tilde{E}_\theta[.]$ as

the corresponding expectation. For convenience, I set $P_0 \equiv P_{\theta_0}$ $E_0[.] \equiv E_{\theta_0}[.]$, $\tilde{P}_0 \equiv \tilde{P}_{\theta_0}$ and $\tilde{E}_0[.] \equiv \tilde{E}_{\theta_0}[.]$.

Because the matching function $K_M(i;\theta)$ is highly non-linear in θ, it is not possible to use linearization to derive the asymptotic distribution of $T_N^*(\hat{\theta}^*)$. I therefore obtain the limiting distribution by employing a version of Le Cam's skeleton argument, analogous to the proof technique of Abadie and Imbens (2016). Let $\mathcal{N} \equiv \{\theta : \|\theta - \theta_0\| < \epsilon\}$ denote a neighborhood of θ_0 for some $\epsilon > 0$ arbitrarily small. The following regularity conditions are similar to Abadie and Imbens (2016):

Assumption 3. *(i) $\theta_0 \in int(\Theta)$ with Θ compact, X has bounded support and $E[XX']$ is non-singular; (ii) $F(.)$ is twice continuously differentiable on $\mathbb{R}$ with derivatives $f(.), f'(.)$ strictly bounded and $f(.)$ strictly positive; (iii) for each $\theta \in \mathcal{N}$ the random variable $F(X'\theta)$ is continuously distributed with interval support; and its pdf $g_\theta(.)$ is such that the collection $\{g_\theta : \theta \in \mathcal{N}\}$ is uniformly Lipschitz continuous; (iv) at least one component of X is continuously distributed, has non-zero coefficient in θ_0, and has a continuous density function conditional on the rest of X; (v) for each $\theta \in \mathcal{N}$ and $w = 0, 1$, the functions $\mu(w, p; \theta)$, $Var[\bar{\mu}(w, X)|F(X'\theta) = p]$, $Cov[X, \bar{\mu}(w, X)|F(X'\theta) = p]$ and $E[\bar{\sigma}^2(w, X)|F(X'\theta) = p]$ are Lipschitz continuous in p with the Lipschitz constants independent of θ; furthermore there exists some $\delta > 0$ such that $E[Y^{4+\delta}|W = w, X = x]$ is uniformly bounded.*

Assumption 4. *There exists some $\epsilon > 0$ such that for all θ satisfying $\|\theta - \theta_0\| < \epsilon$, and for any sequence $\theta_N \to \theta$, $E_{\theta_N}[r(Y, W, X)|W, F(X'\theta_N)]$ converges to $E_\theta[r(Y, W, X)|W, F(X'\theta)]$ almost surely, for any $\mathbb{R}^{k+2}$-to-$\mathbb{R}$ bounded and measurable function $r(y, w, x)$ that is continuous in x.*

The above assumptions rule out the case where all the regressors are discrete. In this case the matching estimator reduces to the propensity score sub-classification estimator, inference for which is easily obtained using standard methods. Assumptions 3(i),(ii) ensure that the propensity scores for all the observations are bounded away from zero and one. Khan and Tamer (2010) show that under full support, the usual parametric rate is not attainable, and the rate of convergence depends on the tail behavior of the regressors and error terms. Hence inference in this context would necessarily be at a non-standard rate, and is beyond the scope of this paper.

Assumption 3 is taken almost directly from Abadie and Imbens (2016). The only substantive difference is in Assumptions 3(iii) and 3(v) which demand uniform extensions of related assumptions in Abadie and Imbens (2016) - in the sense of holding uniformly in a neighborhood $\mathcal{N}$ of θ_0. Assumption 4 is similarly stronger than the corresponding one in Abadie and Imbens (2016). However sufficient conditions for the latter (Theorem S.12 in Abadie and Imbens, 2016) also imply the former.

I shall also require assumptions to ensure the residuals $\{\hat{e}_{1i}(\theta), \hat{e}_{2i}(W_i;\theta)\}$ are 'close' to the unobserved errors $\{e_{1i}(\theta), e_{2i}(W_i;\theta)\}$. I impose the following high level condition:

Assumption 5. *Uniformly over all $\theta \in \mathcal{N}$, it holds under P_0,*

$$\frac{1}{N}\sum_{i=1}^{N}\left(\hat{e}_{1i}(\theta) - e_{1i}(\theta)\right)^2 = o_p(N^{-\xi}), \text{ and}$$

$$\frac{1}{N}\sum_{i=1}^{N}\left(\hat{e}_{2i}(W_i;\theta) - e_{2i}(W_i;\theta)\right)^2 = o_p(N^{-\xi}).$$

for some $\xi > 0$.

The assumption posits that the vector of residuals is close to the vector of true errors in terms of the Euclidean metric. For many of the commonly used non-parametric methods such as series or kernel regression, Assumption 5 can be verified under fairly weak continuity conditions, for instance when $\sup_{\theta\in\mathcal{N}}|\partial\mu(w,x;\theta)/\partial x| < \infty$ under $w = 0, 1$. It is usually straightforward to select the tuning parameters for estimation, such as the number of series terms, either visually or through cross-validation. In simulations, low order polynomial series, such as first or second order polynomials, appear to work reasonably well, and constitute an attractive choice in practice.

The final assumption concerns the number of quantile partitions q_N.

Assumption 6. *The number of quantile partitions satisfies $q_N \to \infty$ and $q_N^{2+\eta}/N \to 0$ as $N \to \infty$ for some $\eta > 0$.*

Assumption 6 is fairly weak in that a wide range of choices for q_N are allowed. Here, the choice of q_N determines how close the bootstrap variance estimate $\hat{V}^*$ is to the true variance (due to re-centering, the bootstrap mean is asymptotically 0). Higher values of q_N increase the balance in the propensity scores within the blocks (thus lowering the bias of $\hat{V}^*$), but reduce the number of observations in the treatment and control groups in each block (thus increasing the variance of $\hat{V}^*$), see Remark 2. In fact, this is the same trade-off faced by sub-classification estimators for average treatment effects. In this case, there exists extensive theoretical and empirical literature suggesting that small values of q_N are sufficient to reduce most of the bias due to the stratification of the propensity score (see e.g. Rosenbaum and Rubin, 1984; Imbens and Rubin, 2015). Indeed, under some reasonable conditions, Rosenbaum and Rubin (1984), drawing on previous work by Cochran (1968), find that 4 blocks/sub-classes are sufficient to reduce the bias by over 85%, while having 5 blocks reduces it by more than 90%. These values are independent of sample size since the bias depends solely on q_N. Consequently, following the recommendation of Rosenbaum and Rubin (1984), I suggest a default choice of $q_N = 5$.

Based on the above assumptions, I can derive the asymptotic properties of the bootstrap estimator. Following the techniques of Abadie and Imbens (2016) and Andreou and Werker (2012), I employ the Le Cam skeleton or discretization device for formalizing the theorem. In particular, I discretize both the bootstrap and sample estimators , $\hat{\theta}^*, \hat{\theta}$ along a grid of cubes of length $d/\sqrt{N}$. For instance, if the j-th component of $\hat{\theta}^*$, $\hat{\theta}^*_j$, falls in the q-th cube where $q = \left\lfloor \sqrt{N}\hat{\theta}_j/d \right\rfloor$ with $\lfloor . \rfloor$ being the nearest integer function, then the corresponding component of the discretized estimator is given by $\tilde{\theta}^*_j = dq/\sqrt{N}$. Analogously, I also discretize $\hat{\theta}$ as $\bar{\hat{\theta}} = d\left\lfloor \sqrt{N}\hat{\theta}/d \right\rfloor/\sqrt{N}$. The theoretical results are thus based on using $\bar{\hat{\theta}}$ rather than $\hat{\theta}$ to construct the bootstrap samples. The discretization is only a theoretical device for applying the skeleton

arguments and not necessary in practice; indeed, the theory doesn't specify any minimum grid size d.

Let P^* denote the bootstrap probability distribution conditional on both the observations, $(\mathbf{Y}, \mathbf{W}, \mathbf{X})$, and $\mathbf{M}$. In other words, P^* represents joint probability distribution of $W^* \sim \text{Bernoulli}(F(X_i^{*\prime}\bar{\hat{\theta}}))$ and $\mathbf{S}^*$ conditional on $(\mathbf{Y}, \mathbf{W}, \mathbf{X}, \mathbf{M})$. The asymptotic properties of the bootstrap procedure are summarized in the following theorem:

Theorem 1. *Suppose that Assumptions 1-6 hold. Then for d sufficiently small,*

$$P^*\left(T_N^*\left(\tilde{\theta}^*\right) \leq z\right) \xrightarrow{p} Pr(Z \leq z) + O(d)$$

under $\tilde{P}_0$, where Z is a normal random variable with mean 0 and variance $V = \sigma^2 - c' I_{\theta_0}^{-1} c$.

I refer to Appendix A.1 for the formal proof Theorem 1. The derivation parallels that of Abadie and Imbens (2016) in using Le Cam's skeleton argument to obtain the limiting distribution. Let P_θ^* denote the joint distribution of $W^* \sim \text{Bernoulli}(F(X_i^{*\prime}\theta))$ and $\mathbf{S}^*$, conditional on both the observed data and $\mathbf{M}$. Note that $P^* \equiv P_{\bar{\hat{\theta}}}^*$. I consider the bootstrap distribution of the estimator under a local sequence of bootstrap probability distributions $P_{\theta_N}^*$, indexed by $\theta_N = \hat{\theta} + h/\sqrt{N}$. Here θ_N can be thought of as local 'shift' of the estimated propensity score parameter. More precisely, I aim to characterize the limiting distribution - under the bootstrap sequence of probabilities, $P_{\theta_N}^*$ - of the vector

$$\begin{pmatrix} T_N^*(\theta_N) \\ \sqrt{N}(\hat{\theta}_N^* - \theta_N) \\ \Lambda_N^*\left(\bar{\hat{\theta}}|\theta_N\right) \end{pmatrix},$$

where $\hat{\theta}_N^*$ is the bootstrap estimator of θ under $P_{\theta_N}^*$, and $\Lambda_N^*(\theta|\theta') \equiv \log(dP_\theta^*/dP_{\theta'}^*)$ denotes the difference in log-likelihood of the bootstrap probability distributions evaluated at θ and θ' . The limiting distribution of $T_N^*(\hat{\theta}^*)$ under P^* can then be obtained by invoking Le Cam's third lemma (to switch from $P_{\theta_N}^*$ to the actual bootstrap probability P^*), and using the discretization device. A technical difficulty is that $\hat{\theta}$ is also random under $\tilde{P}_0$. To this end, I extend the proof techniques of Abadie and Imbens (2016).

Theorem 1 assures that the bootstrap statistic $T_N^*(\hat{\theta}^*)$ has the same limiting distribution as the true sample. A practical consequence of this theorem is $c_{n,\alpha}^* \xrightarrow{p} c_\alpha$ under $\tilde{P}_0$, where c_α is the critical value from the asymptotic distribution of $\sqrt{N}\left(\hat{\tau}(\hat{\theta}) - \tau(\theta_0)\right)$. Thus, the bootstrap procedure is consistent.

As noted earlier, a drawback of the above result is that in finite samples the value of $c_{n,\alpha}^*$ depends on the particular realization of $\mathbf{M}$. To reduce this dependence, it is possible to proceed as in Step 7 of the bootstrap procedure (cf Section 3.3) and average the bootstrap empirical distribution over different values of $\mathbf{M}$. The resulting bootstrap critical value is denoted by $\bar{c}_{n,\alpha}$ (see Section 3.3). The following corollary, proved in Appendix A.2, assures that $\bar{c}_{n,\alpha}$ is consistent with respect to P_0 - the probability distribution of the original data.

Corollary 1. *Suppose that Assumptions 1-6 hold. Then $\bar{c}_{n,\alpha} \xrightarrow{p} c_\alpha + O(d)$ under P_0.*

5. On higher order refinements

In this section I argue that the proposed bootstrap provides a closer approximation to the true distribution of the propensity score matching estimator, as compared to the asymptotic normal limit. I focus in particular on the role played by the randomization of the treatment values and matching functions, and their effect on variance estimation. Previous remarks have already emphasized the importance of redrawing $\mathbf{W}^*$ for inference with propensity score matching. Here, I show by examples that the bootstrap can generate second order refinements even with other causal effect estimators, especially when the overlap in propensity scores is poor.

As the first example, consider the estimation of the variance for the unadjusted treatment effect estimator $\hat{\tau}_a = \bar{Y}_t - \bar{Y}_c$, where $\bar{Y}_t$,$\bar{Y}_c$ denote the sample averages of the outcomes for the treated and control groups. The estimator is consistent when the data is obtained from a Bernoulli trial RCT, for example. Neglecting the heterogeneity term $E[Y(1)|X]-E[Y(0)|X]-\tau_0$ for simplicity, the potential errors in this example are given by $e(1;X) = Y(1) - E[Y(1)|X]$ and $e(0;X) = E[Y(0)|X] - Y(0)$. Suppose that both the propensity scores, $p(X_i)$, and the potential errors, $\{e(1;X_i), e(0;X_i)\}$, are known. The asymptotic variance estimate is

$$\hat{V} = \frac{1}{N}\sum_{i=1}^{N} e^2(W_i;X_i).$$

A straightforward extension of the bootstrap procedure can also be used to provide inference for $\hat{\tau}_a$. The resulting bootstrap variance estimate is

$$\hat{V}_{\text{boot}} = \frac{1}{N}\sum_{i=1}^{N}\Big\{p(X_i)e^2(1;X_i) + (1-p(X_i))e^2(0;X_i)\Big\} - \Xi_a^2,$$

where Ξ_a is the re-centering term. Since $\Xi_a^2 = O(N^{-1})$, I neglect this in further analysis. Let $\Delta_1 = \hat{V} - V$, $\Delta_2 = \hat{V}_{\text{boot}} - V$, and $\Delta_3 = \hat{V} - \hat{V}_{\text{boot}}$, where V denotes the true variance of the estimate. It is possible to decompose $\Delta_1 = \Delta_2 + \Delta_3$, where Δ_2 and Δ_3 are asymptotically independent, since $\hat{V}_{\text{boot}} \approx E[\hat{V}|\mathbf{X}]$. This immediately implies $\hat{V}_{\text{boot}}$ is a more accurate estimator of V than $\hat{V}$. The extent of the gain in accuracy can be characterized using anti-concentration inequalities: with high probability, $\Delta_3 \geq cN^{-1/2}$ for some $c > 0$. Also, the superior performance of the bootstrap holds even if the potential errors have to be estimated. Let $\tilde{V}_{\text{boot}}$ denote the bootstrap estimator based on estimates, $\hat{e}(w;X_i)$, of the potential errors. If, for instance, X is univariate, and the conditional means of $Y(1)$ and $Y(0)$ are linear in X, the values of $\{\hat{e}(1;X_i), \hat{e}(0;X_i)\}$ can be obtained from linear regressions, and it follows $\tilde{V}_{\text{boot}} - \hat{V}_{\text{boot}} = O_p(N^{-1})$. More generally, as long as the dimension of X is not high (in particular $k \leq 5$), it can be shown that $\tilde{V}_{\text{boot}} - \hat{V}_{\text{boot}} = o_p(N^{-1/2})$ and the bootstrap variance estimate is preferable.

The above example demonstrates that for any given realization of the observations, the bootstrap variance estimate is typically closer to the truth. This can translate to large gains when the degree of overlap in propensity scores is poor. The following example is based on propensity score matching for concreteness, but the intuition applies to causal effect estimators more broadly (for example, simply replacing the matching function with inverse propensity scores gives the Horvitz-Thomson estimator):

Consider a dataset where the range of propensity scores falls within an arbitrarily narrow interval centered around a (known) value p_0 that is close to 0. This implies the number of treated observations is very low, but they have a disproportionately high influence, being used as matches very often. Suppose now that the conditional variances (i.e $\sigma(w;X) = \mathrm{Var}(Y(w)|X)$) are independent of w, and determined by a single binary covariate X_1 with $\sigma(x_1) \equiv \sigma(w;X)$ taking the values H (high) and L (low) when $x_1 = 0, 1$ respectively. I also suppose that X_1 takes the values $0, 1$ with equal probability. For simplicity I focus on the within sample variance, by neglecting the first term (corresponding to e_{1i}) in equation (2.2). In this example, the Abadie-Imbens variance estimate is

$$\hat{V}_{\mathrm{AI}} = \frac{1}{N}\sum_{W_i=1}\left(1+\frac{K_M(i)}{M}\right)^2\sigma^2(X_{1i}) \;+\; \frac{1}{N}\sum_{W_i=0}\left(1+\frac{K_M(i)}{M}\right)^2\sigma^2(X_{1i}), \tag{5.1}$$

with $\sigma(X_{1i}) = L + X_{1i}(H-L)$. The bootstrap (within-sample) variance estimate is[5]

$$\hat{V}_{\mathrm{boot}} = \frac{1}{N}\sum_{i=1}^{N} p_0\left(1+\frac{\tilde{K}_M(i;1)}{M}\right)^2\sigma^2(X_{1i}) \;+\; \frac{1}{N}\sum_{i=1}^{N}(1-p_0)\left(1+\frac{\tilde{K}_M(i;0)}{M}\right)^2\sigma^2(X_{1i}). \tag{5.2}$$

The Abadie-Imbens variance estimator - particularly the first term in (5.1) - is highly sensitive to the relative proportion of observations with $X_{1i} = 0$ or 1 in the treated group. Thus when the value of p_0 is low and $H \gg L$, the estimator is highly variable, and therefore inaccurate. On the other hand $\hat{V}_{\mathrm{boot}}$ is more stable. This is because of the re-randomization of treatment values for all the observations, due to which $\hat{V}_{\mathrm{boot}}$ only depends on the observed density of X_{1i} for the entire sample - a much less variable quantity.

A second robustness property of the bootstrap stems from the random imputation of the matching functions (c.f Section 3.2). In the previous example, a low value of p_0 implies greater variability in the matching functions for the treatments. Indeed, it can be shown that

$$\mathrm{Var}[K_M(i)|W_i = 1] \approx \frac{M}{2}\left(\frac{1-p_0}{p_0}\right)^2 + M\frac{1-p_0}{p_0}.$$

Suppose that the variances $\sigma(w;X)$ were not exactly known, then both $\hat{V}_{\mathrm{AI}}, \hat{V}_{\mathrm{boot}}$ would be modified by replacing $\sigma^2(X_{1i})$ with estimated (or imputed, in the case of bootstrap) residuals $\hat{e}^2(w;X_i)$. Consequently $\hat{V}_{\mathrm{AI}}$ is heavily influenced by the error terms of those treated observations that are used as a match most often. Since $\max_{W_i=1} K_M(i) \to \infty$ as $p_0 \to 0$, this again implies greater variability and slow rates of convergence for $\hat{V}_{\mathrm{AI}}$. By contrast, the bootstrap also imputes $\tilde{K}_M(i;1)$ for all the control observations from the conditional distribution of $K_M(i)$ given $W_i = 1$. Thus the high values of $K_M(i)$ are paired with a greater range of the error terms from $\{\hat{e}^2(1;X_i) : i = 1, \ldots N\}$, reducing the influence of a few particular observations.

The above arguments demonstrate as much the benefits of the imputation procedures as those of the bootstrap. However there are other advantages specific to the bootstrap as well. For instance, the bootstrap employs the exact values of the matching functions. By contrast,

[5]This is based on neglecting the recentering term which is of the order N^{-1}. Also I have modified the bootstrap to take into account the known values of the variances and propensity scores. Even if these modifications were not made, the error from approximating the resulting bootstrap variance estimator with $\hat{V}_{\mathrm{boot}}$ can be made arbitrarily small compared to the effect of moving the value of p_0 closer to 0.

in the setup of estimated propensity scores, the asymptotic distribution relies on large sample approximations to the same. When the degree of overlap is poor, or when $p_0 \to 0$ in the above example, the rate of convergence of the matching function to its asymptotic approximation can be very slow, as evidenced by the large variances for $K_M(i;\theta)$. As a result the bootstrap would have better approximation properties.

6. Extensions

6.1. **Average treatment effect on the treated.** Thus far this paper has focused on inference for the average treatment effect. An alternative quantity of interest could be the average treatment effect on the treated (ATET), defined as

$$\tau_t(\theta) = E\left[Y_i(1) - Y_i(0) | W_i = 1\right],$$

when the true propensity score is given by $F(X'\theta)$. The estimator is indexed with θ since it can now be a function of the propensity score. The parameter of interest is the quantity $\tau_t(\theta_0)$. In this section we show how the bootstrap procedure can be extended to provide inference for $\tau_t(\theta_0)$.

The matching estimator for the ATET, for a match based on $F(X'\theta)$, is defined as

$$\hat{\tau}_t(\theta) = \frac{1}{N}\sum_{i=1}^{N} W_i \left(Y_i - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i;\theta)} Y_j \right).$$

The large sample properties of this estimator under estimated propensity scores are derived in Abadie and Imbens (2016). The authors show that the bias is asymptotically negligible (i.e $\sqrt{N} B_t(\hat{\theta}) \xrightarrow{p} 0$), and that

$$\sqrt{N}\left(\hat{\tau}_t(\hat{\theta}) - \tau_t(\theta_0)\right) \xrightarrow{d} N\left(0, \sigma_t^2 - c_t' I_{\theta_0}^{-1} c_t + \frac{\partial \tau(\theta_0)}{\partial \theta}' I_{\theta_0}^{-1} \frac{\partial \tau(\theta_0)}{\partial \theta}\right),$$

subject to discretization. I refer to Abadie and Imbens (2016) for the values of σ_t and c_t.

Unlike the ATE, there doesn't exist a straightforward error representation for $\hat{\tau}_t(\hat{\theta})$ due to the fact $\tau_t(\cdot)$ also depends on the true propensity score. However, it can be shown that an error representation does exist for statistics of the form $\hat{\tau}_t(\theta) - \tau_t(\theta)$. This is given by

$$\hat{\tau}_t(\theta) - \tau_t(\theta) - B_t(\theta) = \frac{1}{N_1}\sum_{i=1}^{N} \varepsilon_{t,i}\left(W_i;\theta\right), \tag{6.1}$$

where $B_t(\theta)$ denotes the bias term, and the potential errors are given by

$$\varepsilon_{t,i}\left(w;\theta\right) = w e_{t,1i}(\theta) + e_{2i}(w;\theta) + (1-w)\left(1 + \frac{\tilde{K}_M(i;w,\theta)}{M}\right) e_{2i}(w;\theta)$$

for $w = 0, 1$, with

$$e_{t,1i}(\theta) \equiv \mu(1, F(X_i'\theta)) - \mu(0, F(X_i'\theta)) - \tau_t(\theta).$$

I thus propose a bootstrap procedure based on (6.1). For each θ, denote

$$\hat{e}_{t,1i}(\theta) \equiv \hat{\mu}(1, F(X_i'\theta)) - \hat{\mu}(0, F(X_i'\theta)) - \hat{\tau}_t(\theta).$$

Then by analogy with the ATE, a valid bootstrap statistic for $\sqrt{N}\left(\hat{\tau}_t(\hat{\theta}) - \tau_t(\hat{\theta})\right)$ is given by

$$\dot{T}^*_{t,N}\left(\hat{\theta}^*\right) = \frac{\sqrt{N}}{N_1^*}\sum_{i=1}^{N}\left\{\varepsilon^*_{t,i}\left(\hat{\theta}^*\right) - \Xi^*_t\left(\hat{\theta}^*\right)\right\},$$

where, for each θ, the bootstrap realized errors are given by

$$\varepsilon^*_{t,i}(\theta) = W_i^*\left\{\hat{e}_{t,1S_i^*}(\theta) + \hat{e}_{2S_i^*}(1;\theta)\right\} + (1 - W_i^*)\left\{\hat{e}_{2S_i^*}(0;\theta) - \hat{\nu}_{S_i^*}(0;\theta)\right\},$$

and

$$\begin{aligned}\Xi^*_t(\theta) &\equiv E^*_\theta[\varepsilon^*_{t,i}(\theta)] \\ &= \frac{1}{N}\sum_{k=1}^{N}\left\{F(X_k'\theta)\left(\hat{e}_{t,1k}(\theta) + \hat{e}_{2k}(1;\theta)\right) + \left(1 - F(X_k'\theta)\right)\left(\hat{e}_{2k}(0;\theta) - \hat{\nu}_k(0;\theta)\right)\right\}\end{aligned}$$

is the re-centering term. Thus, to obtain a valid bootstrap for $\sqrt{N}\left(\hat{\tau}_t(\hat{\theta}) - \tau_t(\theta_0)\right)$, it suffices to adjust $\dot{T}^*_{t,N}\left(\hat{\theta}^*\right)$ by the bootstrap counterpart of $\sqrt{N}\left(\tau_t(\hat{\theta}) - \tau_t(\theta_0)\right)$. Typically, this entails adding $\sqrt{N}\left(\hat{\tau}_t(\hat{\theta}^*) - \hat{\tau}_t(\hat{\theta})\right)$ to $\dot{T}^*_{t,N}\left(\hat{\theta}^*\right)$; unfortunately, this is no longer valid since $\hat{\tau}_t(\cdot)$ is not a smooth function of θ. Consequently, I linearize $\sqrt{N}\left(\tau_t(\hat{\theta}) - \tau_t(\theta_0)\right)$ to obtain the bootstrap statistic $T^*_{t,N}\left(\hat{\theta}^*\right)$ for $\sqrt{N}\left(\hat{\tau}_t(\hat{\theta}) - \tau_t(\theta_0)\right)$ as

$$T^*_{t,N}\left(\hat{\theta}^*\right) = \dot{T}^*_{t,N}\left(\hat{\theta}^*\right) + \sqrt{N}\frac{\hat{\partial\tau}}{\partial\theta}\left(\hat{\theta}^* - \hat{\theta}\right) \tag{6.2}$$

where $\frac{\hat{\partial\tau}}{\partial\theta}$ is an estimator for $\frac{\partial\tau(\theta_0)}{\partial\theta}$, given by (see Abadie & Imbens, 2016)

$$\frac{\hat{\partial\tau}}{\partial\theta} = \frac{1}{N_1}\sum_{i=1}^{N} X_i f(X_i'\hat{\theta})\left\{(2W_i - 1)\left(Y_i - Y_{\mathcal{J}_{\text{NN}}(i)}\right) - \hat{\tau}_t(\hat{\theta})\right\}.$$

The empirical distribution, $F^*_{t,n}(.)$, of $T^*_{t,N}\left(\hat{\theta}^*\right)$ can be obtained by a similar algorithm as in Section 3.3. Using $F^*_{t,n}(.)$, and a particular realization of $\mathbf{M}$, the critical value is obtained as $c^*_{t,n,\alpha} = \inf\{u : F^*_{t,n}(u) \geq 1 - \alpha\}$. Alternatively, averaging the empirical distribution $F^*_{t,n}(.)$ over L different values of $\mathbf{M}$ gives $\bar{F}^*_{t,n}(.)$. The resulting critical values are given by $\bar{c}^*_{t,n,\alpha} = \inf\{u : \bar{F}^*_{t,n}(u) \geq 1 - \alpha\}$.

Let $c_{t,\alpha}$ denote the critical value from the asymptotic distribution of $\sqrt{N}\left(\hat{\tau}_t(\hat{\theta}) - \tau_t(\theta_0)\right)$. The following theorem assures that the bootstrap procedure for the ATET is consistent. As with Theorem 1, the formal statement relies on discretization.

Theorem 2. *Suppose that Assumptions 1-6 hold. Then for d sufficiently small,*

$$P^*\left(T^*_{t,N}\left(\tilde{\theta}^*\right) \leq z\right) \xrightarrow{p} Pr(Z_t \leq z) + O(d)$$

under $\tilde{P}_0$, where Z_t is a normal random variable with mean 0 and variance $V_t = \sigma_t^2 - c_t' I_{\theta_0}^{-1} c_t + \frac{\partial\tau(\theta_0)}{\partial\theta}' I_{\theta_0}^{-1}\frac{\partial\tau(\theta_0)}{\partial\theta}$. Furthermore, $\bar{c}_{t,n,\alpha} \xrightarrow{p} c_{t,\alpha} + O(d)$ under P_0.

The proof of the theorem is similar to that of Theorem 1, and therefore omitted. Similar results also hold for related estimators like the average treatment effect on the controls.

Remark 4. In empirical examples pertaining to the ATET, it is frequently the case that $N_1 \ll N_0$. In such cases, the bootstrap resamples would be predominantly dominated by observations from the control arm. However the error terms and matching functions are imputed from the treated variables. Hence the information from the treated sample is still incorporated in each bootstrap draw.

6.2. **Matching without replacement.** In this section I consider matching without replacement as an alternative for estimating the ATET. This has the advantage of having a lower variance, compared to matching with replacement. At the same time, if the pool of controls is sufficiently large, the increase in bias is not substantial. Here I focus on so called optimal-matching (Rosenbaum, 1989), which is one procedure for matching without replacement. However, the proposed bootstrap is applicable more generally, for instance to greedy or sequential matching.

Suppose that the propensity scores are given by $F(\mathbf{X}'\theta)$. The matching indices, $\mathcal{J}_M^{\text{opt}}(i;\theta)$, for optimal-matching are obtained as the ones that minimize the sum of matching discrepancies, i.e

$$\mathcal{J}_M^{\text{opt}}(\cdot;\theta) \in \text{argmin}_{\{J(i):i=1,\ldots,N\}} \sum_{i=1}^{N} W_i \sum_{j\in J(i)} \left\| F(X_i'\theta) - F(X_j'\theta) \right\|,$$

where $J(.) : \{i : W_i = 1\} \mapsto \{i : W_i = 0\}$ is any one-one mapping from the indices of the treated observations to that of the controls. The corresponding matching function is denoted by

$$K_M^{\text{opt}}(i;\theta) = \sum_{j=1}^{N} \mathbb{I}_{i\in\mathcal{J}_M^{\text{opt}}(j;\theta)}.$$

By definition $K_M^{\text{opt}}(i;\theta) \in \{0,1\}$ for every unit i in the treatment group. For matching based on $F(X'\theta)$, the optimal-matching estimator for the ATET is then

$$\hat{\tau}_t^{\text{opt}}(\theta) = \frac{1}{N}\sum_{i=1}^{N} W_i \left(Y_i - \frac{1}{M} \sum_{j\in\mathcal{J}_M^{\text{opt}}(i;\theta)} Y_j \right).$$

The estimators $\hat{\tau}_t^{\text{opt}}(\theta)$ and $\hat{\tau}_t(\theta)$ only differ in employing $K_M^{\text{opt}}(i;\theta)$ instead of $K_M(i;\theta)$ as the matching function. With the estimated propensity score, the quantity of interest is $\hat{\tau}_t^{\text{opt}}(\hat{\theta})$. To obtain its large sample properties, I impose the following condition, based on Abadie and Imbens (2012): (Let $\mathcal{N}$ denote some neighborhood of θ_0)

Assumption 7. *Uniformly over all $\theta \in \mathcal{N}$, it holds under P_0 that as $N_1 \to \infty$,*

$$\frac{1}{\sqrt{N_1}} \sum_{i=1}^{N} W_i \sum_{j\in\mathcal{J}^{opt}(i;\theta)} \left\| F(X_i'\theta) - F(X_j'\theta) \right\| \xrightarrow{p} 0.$$

Assumption 7 is a high level condition ensuring the bias from the optimal matching decays fast enough to 0. Suppose that $g_{1,\theta}$ and $g_{0,\theta}$ denote the conditional pdfs of $F(X'\theta)$ conditional on $W_i = 1$ and $W_i = 0$ respectively. Following the arguments of Abadie and Imbens (2012, Proposition 1), sufficient conditions for Assumption 7 can be provided as: (i) $\sup_{\theta\in\mathcal{N}} g_{1,\theta}, g_{0,\theta} \leq C < \infty$ and $\inf_{\theta\in\mathcal{N}} g_{0,\theta} \geq c > 0$; and (ii) $N_1^r \leq cN_0$ for some $c > 0$ and $r > 1$. Here, the requirement of $N_1 \ll N_0$ is crucial for driving down the bias. Using Assumptions 1-7, it is

possible to derive the limiting distribution of $\hat{\tau}_t^{\text{opt}}(\hat{\theta})$,

$$\sqrt{N}\left(\hat{\tau}_t^{\text{opt}}(\hat{\theta}) - \tau_t(\theta_0)\right) \xrightarrow{d} N\left(0, \sigma_w^2 - c_w' I_{\theta_0}^{-1} c_w + \frac{\partial \tau(\theta_0)}{\partial \theta}' I_{\theta_0}^{-1} \frac{\partial \tau(\theta_0)}{\partial \theta}\right),$$

subject to discretization. The proof of the above, together with the expressions for $\sigma_w^2,\ c_w$, can be obtained by adapting the arguments of Abadie and Imbens (2012, 2016). In general $\sigma_w^2,\ c_w$ are distinct from the corresponding quantities, $\sigma_t^2,\ c_t$, for matching with replacement.

Given the close analogy with $\hat{\tau}_t(\theta)$, it is straightforward to modify the bootstrap procedure of Section 6.1 to obtain valid inference for $\hat{\tau}_t^{\text{opt}}(\hat{\theta})$. The primary difference is that the matching functions are obtained as $K_M^{\text{opt}}(i;\hat{\theta}^*)$ rather than $K_M(i;\hat{\theta}^*)$ in Step 4. Also, only the values of the potential matching function $\tilde{K}_M^{\text{opt}}(i;w,\theta)$ for $w = 0$ need to be known, since the optimal-matching function is defined solely for control variables. The proposed bootstrap test statistic for $\hat{\tau}_t^{\text{opt}}(\hat{\theta})$, denoted by $T_{t,N}^{(\text{opt})*}\left(\hat{\theta}^*\right)$, thus has the same form as $T_{t,N}^*\left(\hat{\theta}^*\right)$, with the sole change being the matches are now given by $K_M^{\text{opt}}(i;\theta)$. Consistency of the bootstrap procedure can be demonstrated by analogous arguments to Theorem 1, using results from Abadie and Imbens (2012).

Theorem 3. *Suppose that Assumptions 1-7 hold. Then for d sufficiently small,*

$$P^*\left(T_{t,N}^{(opt)*}\left(\tilde{\theta}^*\right) \leq z\right) \xrightarrow{p} Pr(Z_t^{(opt)} \leq z) + O(d)$$

under $\tilde{P}_0$, where $Z_t^{(opt)}$ is a normal random variable with mean 0 and variance $V_t^{opt} = \sigma_w^2 - c_w' I_{\theta_0}^{-1} c_w + \frac{\partial \tau(\theta_0)}{\partial \theta}' I_{\theta_0}^{-1} \frac{\partial \tau(\theta_0)}{\partial \theta}$.

6.3. **Other causal effect estimators.** The bootstrap procedure can easily be extended to other causal effect estimators. Indeed, estimators of the ATE that are linear in the outcome variables, for instance propensity score sub-classification or Horvitz-Thompson estimators, have a common structure in terms of an error representation of the form

$$\hat{\tau}^{(c)} - E[\hat{\tau}^{(c)}] = \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i^{(c)}(W_i;\theta),$$

where the potential errors are given by[6]

$$\varepsilon_i^{(c)}(w;\theta) = e_{1i}(\theta) + (2w-1)\Lambda_i(\mathbf{X}'\theta, \mathbf{W}_{-i}, w) e_{2i}(w;\theta).$$

Here, $\Lambda_i(\mathbf{X}'\theta, \mathbf{W}_{-i}, w)$ may interpreted as quantifying the importance of each observation in terms of estimating the ATE, depending on whether it is in the treated ($w = 1$), or control group ($w = 0$). The estimators differ only in the choice of $\Lambda_i(\mathbf{X}'\theta, \mathbf{W}_{-i}, w)$. The propensity score matching estimator sets $\Lambda_i(\mathbf{X}'\theta, \mathbf{W}_{-i}, w) = 1 + \tilde{K}_M(i;w,\theta)$, while setting

$$\Lambda_i^{-1}(\mathbf{X}'\theta, \mathbf{W}_{-i}, w) = wF(X_i'\theta) + (1-w)(1 - F(X_i'\theta))$$

[6]The term $\mathbf{W}_{-i}$ denotes the vector of treatments $\mathbf{W}$ excluding W_i.

gives the Horvitz-Thompson estimator. In a similar vein, the propensity score sub-classification estimator sets

$$\Lambda_i(\mathbf{X}'\theta, \mathbf{W}_{-i}, w) = w\frac{N_1(b_i(\theta)) + N_0(b_i(\theta))}{N_1(b_i(\theta))} + (1-w)\frac{N_1(b_i(\theta)) + N_0(b_i(\theta))}{N_0(b_i(\theta))},$$

where $b_i(\theta)$ denotes the block in which observation i resides when the blocks are obtained by partitioning $F(\mathbf{X}'\theta)$; and $N_1(b)$, $N_0(b)$ denote the number of treated and control observations in block b. A common theme across all choices is that control (treated) units with high (low) propensity scores gain greater importance, to compensate for them being fewer in number.

The techniques in Section 3 provide a template for estimating and imputing the potential error terms, $\varepsilon_i^{(c)}(w;\theta)$. In particular, the values of $e_{2i}(w;\theta)$ for $w = 0, 1$ can be obtained through secondary matching as in Section 3.1. Additionally, the unobserved values of the importance function $\Lambda_i(\mathbf{X}'\theta, \mathbf{W}_{-i}, w)$ can be imputed either through a blocking scheme as in Section 3.2, or directly, if the functional form is known, as in the case of Horvitz-Thompson and propensity score sub-classification estimators. Consequently, given the potential errors, a bootstrap algorithm can be constructed by analogy with Section 3.3; indeed, the bootstrap drawing and re-centering schemes continue to apply.

The consistency of the bootstrap procedure for this more general class of estimators follows by the same reasoning as in Theorem 1.

7. Simulation

In this section I investigate the finite sample performance of the bootstrap procedure outlined in Section 3.3 using simulation exercises. These confirm my theoretical results and demonstrate the accuracy of the bootstrap procedure.

7.1. **Simulation designs.** I consider four different data generating processes. The first DGP (DGP1) is taken from Abadie and Imbens (2016, Supplementary material). I generate a two dimensional vector (X_1, X_2) of covariates by drawing both variables from a uniform $[-1/2, 1/2]$ distribution independently of each other. The potential outcomes are generated as $Y(0) = 3X_1 - 3X_2 + U_0$ and $Y(1) = 5 + 5X_1 + X_2 + U_1$, where U_1 and U_0 are mutually independent standard normal random variables. The propensity score is given by the logistic function

$$p(X) \equiv P(W = 1|X) = \frac{\exp(X_1 + 2X_2)}{1 + \exp(X_1 + 2X_2)},$$

and the treatments are generated as $W \sim \text{Bernoulli}(p(X))$. Finally, the outcome variables are generated as $Y = WY(1) + (1-W)Y(0)$.

The second DGP (DGP2) is similar to the first except that the potential outcomes are generated as $Y(0) = -3X_1 + 3X_2 + U_0$ and $Y(1) = 5 + 7X_1 + 12X_2^2 + U_1$. In this DGP the treatment effect varies more widely with X. Additionally, it also incorporates some non-linearity through the quadratic term in $Y(1)$.

The third DGP (DGP3) is also similar to the first except that the propensity scores are given by

$$P(W = 1|X) = \frac{\exp(X_1 + 7X_2)}{1 + \exp(X_1 + 7X_2)}.$$

The effect of this is to greatly reduce to amount of overlap in the propensity scores between the treated and control samples, as compared to DGP1. For instance, out of a set of 1000 observations, less than 5% of the first 200 observations as ordered by the propensity score are from the treated sample .

The final DGP (DGP4) is adapted from Kang and Schafer (2007). This is chosen for its resemblance with a real data study.[7] For each observation I draw covariates X_1, X_2, X_3, X_4 independently of each other from a standard normal distribution. The potential outcomes are given by $Y(1) = 210 + 27.4X_1 + 13.7X_2 + 13.7X_3 + 13.7X_4 + U_1$ and $Y(0) = U_0$ where U_1 and U_0 are independent standard normal random variables. The propensity scores are given by

$$P(W = 1|X) = \frac{\exp(-X_1 + 0.5X_2 - 0.25X_3 - 0.1X_4)}{1 + \exp(-X_1 + 0.5X_2 - 0.25X_3 - 0.1X_4)}.$$

In all DGPs I consider the case of a single match, i.e $M = 1$. I consider four different sample sizes: $N = 100,\ 200,\ 500,\ 1000$. In all cases, the number bootstrap repetitions is $B = 399$, and the number of Monte-Carlo repetitions is 2500. To ease the computational burden, I only present results for the bootstrap procedure based on a single realization of the multinomial random vector $\mathbf{M}$ (i.e I only follow steps 1-6 of the algorithm in Section 3.3).

7.2. **Choice of tuning parameters.** The procedure requires choosing the type of series regression and the number of quantile partitions q_N. Based on visual inspection, I used third order polynomials for the series regression for all DGPs. In practice the number of series terms can also be chosen through cross-validation. I also experimented with different types of non-parametric estimators and show the procedure is not sensitive to the particular choices employed. The choice of q_N was discussed in Section 4; there I recommended setting a value of $q_N = 5$. Correspondingly, for the baseline results I set $q_N = 5$ throughout. In a separate table I also report results for different choices of q_N.

For the secondary matching (c.f Section 3.1), I employ nearest neighbor matching based on the Euclidean metric. Since in all the DGPs the covariates are standard normal and independent of each other, this is practically equivalent to matching on the Mahalanobis metric.

7.3. **Simulation results.** Table 1 reports the performance of the bootstrap inferential procedure for all the DGPs, along with inference based on the asymptotic distribution. The nominal coverage probability is 0.95. The tuning parameters of the number of series terms and q_N haven been deliberately kept unchanged with sample size to emphasize that the values reported are not due to the particular selection of these parameters. In all cases, the bootstrap critical values are very close to nominal even for relatively small sample sizes, for example $N = 100$.

The bootstrap outperforms inference based on the asymptotic distribution in almost all cases. The performance of the bootstrap is particularly advantageous when the sample size is small, see e.g. the results for $N = 100$; and when the extent of imbalance in propensity scores is high, e.g. DGP3. At the same time, when there is sufficient overlap between the propensity scores,

[7]The original simulation study of Kang and Schafer (2007) uses a version of this DGP to evaluate the performance of different procedures under missing data. Here I adapt it to study average treatment effects.

TABLE 1. Rejection probabilities under the null for various DGPs

		Sample size			
		$N = 100$	$N = 200$	$N = 500$	$N = 1000$
DGP1	*Bootstrap*	0.057	0.044	0.050	0.050
	Asymptotic	0.070	0.059	0.054	0.049
DGP2	*Bootstrap*	0.058	0.052	0.054	0.058
	Asymptotic	0.063	0.056	0.058	0.057
DGP3	*Bootstrap*	0.090	0.069	0.052	0.047
	Asymptotic	0.141	0.100	0.092	0.069
DGP4	*Bootstrap*	0.079	0.066	0.057	0.058
	Asymptotic	0.118	0.077	0.061	0.057

TABLE 2. Rejection probabilities under the null for different non-parametric estimators when $N = 500$

	Non-parametric estimators			
	Linear	Poly-3	Poly-4	Spline
DGP1	0.056	0.050	0.055	0.051
DGP2	0.052	0.054	0.047	0.048
DGP3	0.030	0.052	0.048	0.049
DGP4	0.064	0.057	0.062	0.062

and the effect of estimation of propensity scores is negligible, as in DGP2, there is very little difference between the inferential procedures.

To assess the sensitivity of the bootstrap, I repeated the Monte-Carlo simulations for different choices of tuning parameters. In Table 2, I experiment with different non-parametric specifications to estimate the residuals, namely: linear, third and fourth order polynomials, and cubic smoothing splines (with smoothing parameter 0.99). The bootstrap is quite insensitive to the choice of the specification. I found similar results for the other sample sizes; for brevity I do not report these results.

In Table 3, I repeat the procedure for different values of q_N under the sample sizes $N = 200$ and $N = 500$ for all the DGPs. I find that the bootstrap procedure is largely robust to the actual choice of q_N, except for the value of $q_N = 1$, which corresponds to no partitioning. This is consistent with the observation, made in Section 4, that small values of q_N are sufficient to reduce most of the bias. At the same time, even for larger sample sizes, the reduction in bias is marginal as q_N increases beyond a certain amount. For example, there is not much variability in the results between $q_N = 5$ and $q_N = 8$.

7.4. **Robustness to Mis-specification.** To check the robustness of the inference to mis-specification, I modify the DGPs by using a Probit link function for the true propensity scores, even as the estimation and inferential procedures themselves employ the Logistic regression. Table 4 reports the results of the simulation under various DGPs when $N = 200$ and 500.

TABLE 3. Rejection probabilities under the null for different values of q_N

		Number of quantile partitions			
		$q_N = 1$	$q_N = 2$	$q_N = 5$	$q_N = 8$
DGP1	$N = 200$	0.067	0.057	0.046	0.049
	$N = 500$	0.057	0.054	0.050	0.050
DGP2	$N = 200$	0.072	0.058	0.060	0.048
	$N = 500$	0.070	0.047	0.047	0.060
DGP3	$N = 200$	0.202	0.082	0.069	0.076
	$N = 500$	0.190	0.078	0.052	0.045
DGP4	$N = 200$	0.103	0.079	0.066	0.064
	$N = 500$	0.093	0.078	0.057	0.054

TABLE 4. Rejection probabilities for the null under mis-specification

		Data Generating Process			
		DGP1	DGP2	DGP3	DGP4
N = 200	*Bootstrap*	0.050	0.046	0.107	0.087
	Asymptotic	0.063	0.062	0.141	0.133
N = 500	*Bootstrap*	0.052	0.052	0.091	0.069
	Asymptotic	0.052	0.062	0.142	0.101

While performance of both inferential procedure degrades somewhat, the bootstrap remains much more robust. A reason for this could be that the residuals $\hat{e}_1(.), \hat{e}_2(.,.)$ - obtained under the mis-specified propensity score - still approximate the actual errors under mis-specification.

8. Case study - The LaLonde datasets

The National Supported Work (NSW) demonstration was a randomized evaluation of a job training program, first analyzed by LaLonde (1986), and later the focus of papers by Heckman and Hotz (1989), Deheija and Wahba (1999), Smith and Todd (2005) among others. The original dataset is based on a randomized study. LaLonde (1986) set aside the experimental control group and replaced it with two other sets of observations from the Panel Study of Income Dynamics (PSID) and the Current Population Survey (CPS). In this section I simulate observations resembling the LaLonde experimental and observational datasets, and use them as test cases for analyzing the relative performance of the bootstrap and asymptotic inferential procedures[8].

8.1. **Description of the data and the data generating process.** The datasets comprise of the following pre-treatment variables: age (age), years of education (edu), indicator for high school dropout (nodeg), indicator for married (mar), real earnings (in thousands of dollars) in 1974 (re74), indicator for unemployed in 1974 (un74), real earnings (in thousands of dollars)

[8]LaLonde (1986) replaced the experimental control group to analyze the accuracy of non-experimental statistical methods. Here I abstract away from this issue by explicitly imposing selection on observables in simulations.

in 1975 (re75), indicator for unemployed in 1975 (un75), and finally two indicators for race: (black) and (hispanic). The outcome variable is real earnings in 1978 (Y). For the results in this section I consider only the African American subsample, which comprises the bulk ($> 85\%$) of the original experimental data. This selects $N_0 = 215$ and $N_1 = 156$ control and treated observations respectively from the experimental dataset, for a total of $N = 371$ observations.

For the observational data, I follow LaLonde (1986) in replacing the experimental control group with the subgroup of all men from PSID and CPS samples who were not working when surveyed in the spring of 1976 (denoted as PSID-2 and CPS-2 respectively). I further extract the African-American subsample from these datasets. This selects $N_0 = 99$ and $N_0 = 286$ observations for the control groups based on the PSID and CPS samples, for a total of $N = 255$ and $N = 442$ observations respectively (given the $N_1 = 156$ treated observations).

I simulate observations mimicking the experimental and observational datasets by broadly following the algorithm described in Busso, DiNardo and McRary (2014). Denote by $\tilde{\mathbf{X}}$ the original set of covariates, and let $\mathbf{Z}$ denote the set of variables comprised of an intercept, $\tilde{\mathbf{X}}$, all the squared terms in $\tilde{\mathbf{X}}$, and the following interaction terms: un75×un74, edu×re75, re74×re75. For each simulation draw, I generate N observations using the following procedure: (1) Draw new covariates $\mathbf{X}$ using the population model specified in the next paragraph; (2) Estimate the propensity scores as $p(\mathbf{X}) = F(V'\theta_0)$ where $F(.)$ is the Logistic function, V is a vector of covariates described below, and θ_0 is the parameter vector obtained given by running a Logistic regression on the original datasets; (3) construct $Y_i(0) = Z_i'\delta_0 + \sigma_0\epsilon_i$, where δ_0 is obtained by regressing the control observations of the original datasets with $\mathbf{Z}$, σ_0^2 is the root mean squared error of the regression, and ϵ_{0i} are iid standard normal errors; (4) construct $Y_i(1)$ analogously using the treated observations from the original datasets; (5) construct treatment values as $\mathbf{W} \sim \text{Bernoulli}(p(\mathbf{X}))$; (6) construct outcome values as $Y_i = W_iY_i(1) + (1 - W_i)Y_i(0)$.

Following Busso, DiNardo and McRary (2014), I draw the new covariates $\mathbf{X}$ in the following way: (1) draw the indicator variables mar, un74, un75 by sapling with replacement from the original datasets; (2) fix the pair (mar, un74, un75) as a group and simulate the other variables, i.e (age, edu, re74, re75), from a group-specific multivariate normal distribution, where the distributional parameters are the group means and covariances estimated from the original data; (3) round the values of age, edu to the nearest integer values.

For the experimental data I use a linear specification for the propensity scores with V = (age, edu, nodeg, mar, re74, re75, un74, un75). For the observational designs I employ a somewhat modified version of the propensity score specification used by Deheija and Wahba (1999): V = (age, edu, mar, nodeg, re74, re75, age^2, edu^2, re74^2, re75^2, edu×re74).

The simulations are designed to replicate the broad features of both the experimental and observational datasets. Of particular interest is the degree of overlap in the propensity scores between the treated and control groups. Figure 8.1 presents a representative plot for the simulated datasets. In the experimental design there is a high degree of overlap in the propensity scores which are also bounded away from 0 and 1. On the other hand, the degree of overlap is quite poor in the observational designs with many of the treated observations concentrated

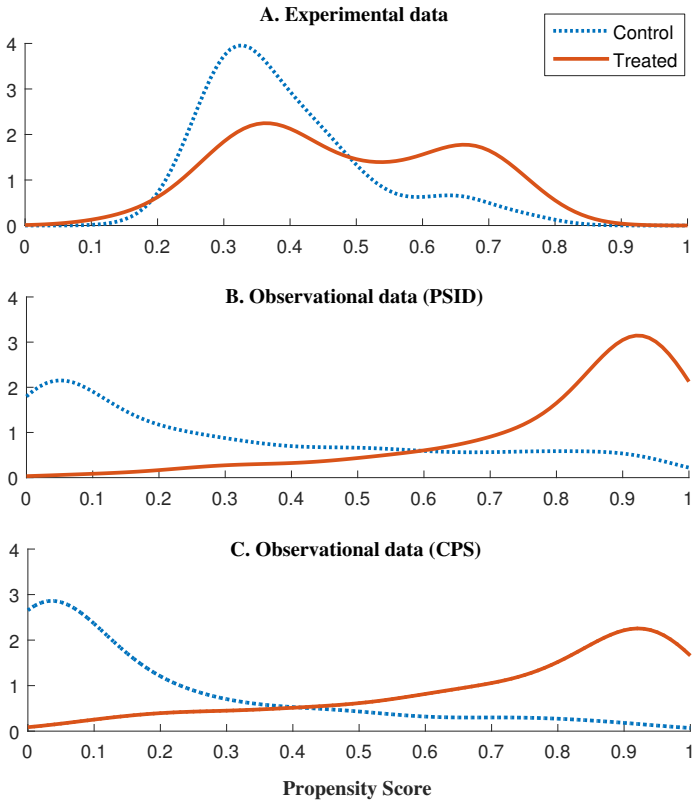

FIGURE 8.1. Representative overlap plots based on kernel density estimates of propensity scores for control (dotted line) and treated units (solid line)

around the propensity score value of 0. This has a significant impact on the performance of inferential methods for matching.

8.2. **Simulation results.** I first describe the bootstrap procedure: For secondary matching (cf Section 3.1), I used nearest neighbor matching based on the Mahalanobis metric, applied over the unique set of covariates in the data, i.e overage, edu, mar, re74, re75.[9] Additionally, based on a visual inspection, I employed a linear specification for the series regression in all designs. For the number of quantile partitions, I employed $q_N = 5$ for the experimental and PSID designs, and $q_N = 4$ for the CPS design. The reason for the lower value of q_N in the latter case is due to the poor overlap in the propensity scores, which results in some cells having no treated observations when q_N is higher.[10] Table 5 reports the performance of the bootstrap and asymptotic inferential procedures for the matching estimator of the ATE. All values are based on 5000 Monte-Carlo repetitions with $B = 399$. The results are provided after bias correction, which in any case is an order of magnitude smaller than the standard deviation.

The first three rows of Table 5 present the simulation results with the same sample sizes as in the original datasets. For the experimental design, both the bootstrap and asymptotic methods provide very similar performance. This is an example in which estimation of the propensity

[9]Indeed the other covariates are defined as functions of these with nodeg=1(edu < 12), un74=1(re74 $= 0$) and un75=1(re75 $= 0$).

[10]In the rather rare instance where one of the cells has no treated observations even with the lower value of q_N, I impute the matching function by drawing treated observations randomly from the neighboring cell.

TABLE 5. Rejection probabilities and average length of confidence intervals (in thousands of dollars) under experimental and observational designs

	Rejection probability		Confidence Interval length		
	Bootstrap	*Asymptotic*	*Bootstrap*	*Asymptotic*	*True*
Experimental (N = 371)	0.061	0.054	3.474	3.528	3.666
Observational (PSID) (N = 255)	0.079	0.214	7.426	5.621	8.335
Observational (CPS) (N = 422)	0.076	0.233	8.669	5.985	9.288
Experimental (N = 150)	0.075	0.075	5.270	5.312	5.756
Observational (PSID) (N = 500)	0.063	0.148	5.984	4.848	6.147
Observational (CPS) (N = 1000)	0.062	0.169	7.242	5.363	7.194

scores hardly affects variance. The asymptotic method appears to be slightly preferable, even if the difference is not statistically significant. This is possibly due to the bias introduced by the nearest-neighbor-matching technique while imputing the error terms.

The performance of the inferential methods declines under both observational designs. Nevertheless, the asymptotic procedure performs considerably worse than the bootstrap, and underestimates the length of the confidence interval by close to 33% of the true length. (I also found that in about 4-5% of the cases, the asymptotic procedure actually reported a negative value for the variance!) By contrast, the bootstrap provides good size control, despite the fact the propensity scores are not bounded away from 0 and 1.

Figure 8.2 plots the estimates of the finite sample distribution (after centering by the true value) using bootstrap and asymptotic methods for representative simulation samples. For the observational data, the estimate from the asymptotic method is highly biased and heavily underestimates the true variance. The bootstrap distribution is much closer to the actual one.

In fact, not only is the asymptotic variance estimate heavily biased for the observational data, it is also highly variable. Figure 8.3 demonstrates this by plotting the finite sample distributions using bootstrap and asymptotic methods for 20 different simulation samples under experimental and PSID designs (the CPS dataset is omitted for brevity). For the PSID data, the bootstrap estimate of the finite sample distribution is much more stable over the different simulation samples. However both methods perform very similarly on the experimental dataset, suggesting that most of the differences in the PSID design are generated by poor overlap. This is consistent with the discussion in Section 5, where I showed that the asymptotic variance estimate is much more sensitive to a few influential observations, as compared to the bootstrap.

In the last three rows of Table 5, I redo the simulation with different sample sizes. Here, I employ $q_N = 5$ for all the designs. For the experimental design, both inferential methods perform

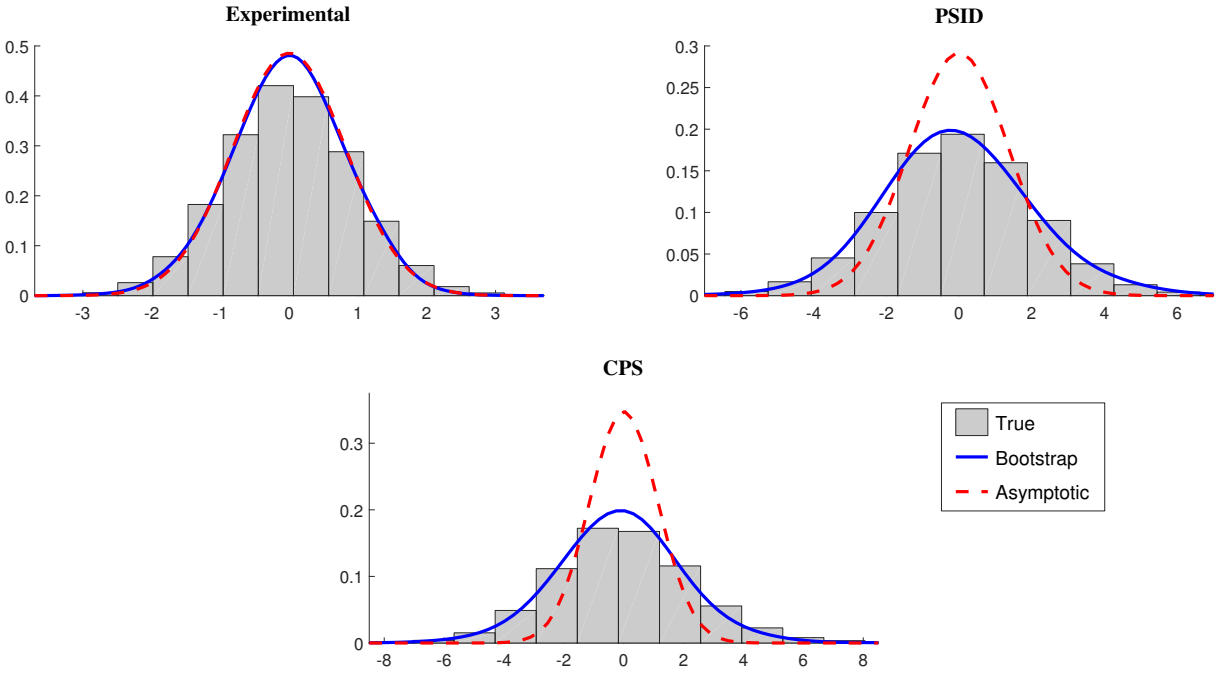

FIGURE 8.2. Estimates of the finite sample distribution using bootstrap (solid blue) and asymptotic methods (dashed red) for representative simulation samples. The bars represent the actual finite sample distribution.

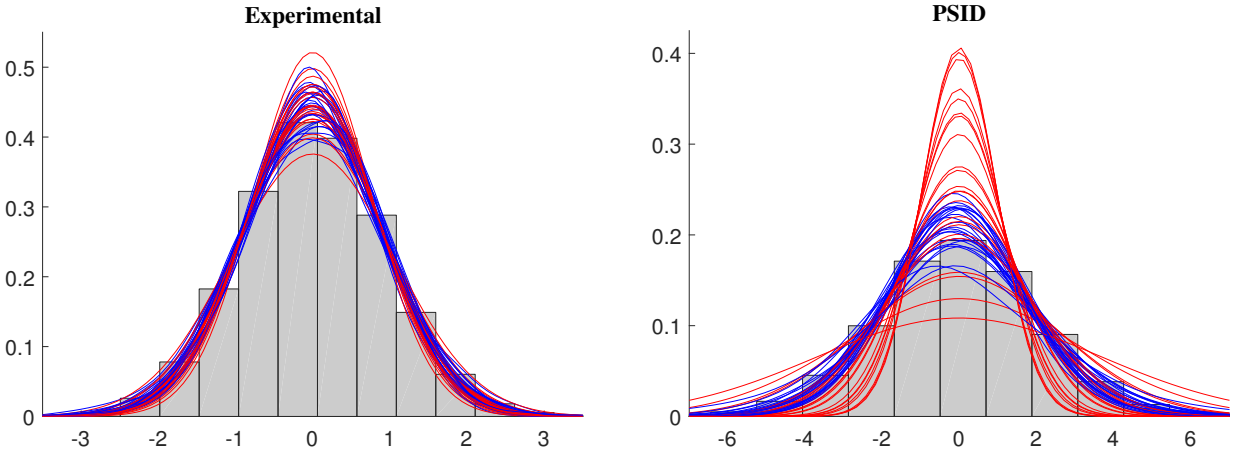

FIGURE 8.3. Estimates of finite sample distributions using bootstrap (blue) and asymptotic methods (red) for 20 different simulation samples. The bars represent the actual finite sample distribution. Note the difference in scaling of the axes.

well even on a sample size that is about half the original one. However, for the observational designs the bootstrap outperforms asymptotic inference by a considerable margin even after doubling the number of observations.

9. CONCLUSION

In this paper, I propose a bootstrap procedure for propensity score matching estimators of the ATE and ATET, and demonstrate its consistency. The procedure can be easily extended to other estimators, including, but not limited to, inverse probability weighting (e.g. Horvitz-Thompson) and propensity score sub-classification. It is built around the concepts of potential

errors and the error representation, introduced in this paper. Both these concepts are also applicable very generally. Together, they constitute a powerful new formalism for describing causal effect estimators.

Simulations and theoretical examples suggest the proposed bootstrap achieves greater accuracy than asymptotic methods, particularly when the overlap in propensity scores is poor. They also highlight the key role played by the (re-)randomization of treatment values in obtaining more precise inference. While beyond the scope of this paper, it would be interesting to formally investigate the higher order properties of the bootstrap procedure.

This paper focuses on treatment effects. However, the techniques and results in this paper may also be useful in other contexts, for instance where the outcome data is missing at random (i.e the propensity for a data point to be missing is only a function of observed covariates).

Appendix A. Proofs of main results

Let $\mathbf{Z} = (\mathbf{Y}, \mathbf{W}, \mathbf{X})$ denote the observed data. Let $\tilde{P}_0$ denote the joint probability distribution of the observed data $\mathbf{Z} = (\mathbf{Y}, \mathbf{W}, \mathbf{X})$ together with $\mathbf{M}$; and $\tilde{E}_0[.]$, the corresponding expectation over $\tilde{P}_0$. I shall reserve the notation $\xrightarrow{p}$ for convergence in probability with respect to $\tilde{P}_0$. I shall also use the notation a.s-$\tilde{P}_0$ for 'almost surely under $\tilde{P}_0$'. As defined in the man text, let P^*_θ denote the joint distribution of $\mathbf{W}^*, \mathbf{S}^*$ conditional on $\mathbf{Z}, \mathbf{M}$, when $W^* \sim \text{Bernoulli}(F(X_i^{*\prime}\theta))$. In other words, this is equivalent to the distribution of the bootstrap sequence of observations (conditional on the data and $\mathbf{M}$) when the treatments, $\mathbf{W}^*$, are constructed using θ instead of $\bar{\hat{\theta}}$. I shall use P^* as a shorthand for $P^*_{\bar{\hat{\theta}}}$.

In the proof I consider local sequences of the form $\theta_N = \bar{\hat{\theta}} + h/\sqrt{N}$ for some vector h. This in turn indexes a local sequence of bootstrap probability distributions $P^*_{\theta_N}$, or P^*_N for simplicity of notation. Let $\mathbf{Z}^*_N = (\mathbf{W}^*_N, \mathbf{X}^*) = f(\mathbf{W}^*_N, \mathbf{S}^*)$ denote the bootstrap observations obtained under P^*_N. I index the observations with N to reflect the fact that the distribution of $\mathbf{Z}^*_N$ as a function of the data depends on θ_N, which varies with N. I shall denote by $\mathcal{L}(.)$ the (unconditional) probability law of some random variable, and by $\mathcal{L}^*_N(.)$ the probability law of a random variable under the bootstrap distribution P^*_N conditional on the data and $\mathbf{M}$. Let $E^*_N[.]$ be the expectation of a random variable with respect to P^*_N.

Let $\Lambda^*_N(\theta|\theta') \equiv \log(dP^*_\theta/dP^*_{\theta'})$ denote the difference in log-likelihood of the bootstrap probability distributions evaluated at θ and θ', i.e.

$$\Lambda^*_N(\theta|\theta') = L(\theta|\mathbf{Z}^*_N) - L(\theta'|\mathbf{Z}^*_N).$$

The bootstrap estimator of θ under P^*_N is represented by $\hat{\theta}^*_N$. Denote by $\psi^*_{N,i}(\theta_N)$, the influence function for $\hat{\theta}^*_N$ under P^*_N, i.e.

$$\psi^*_{N,i}(\theta_N) = X^*_i \frac{W^*_{N,i} - F(X_i^{*\prime}\theta_N)}{F(X_i^{*\prime}\theta_N)(1 - F(X_i^{*\prime}\theta_N))} f(X_i^{*\prime}\theta_N),$$

and let $S^*_N(\theta_N) = N^{-1/2}\sum_{i=1}^N \psi^*_{N,i}(\theta_N)$ denote the corresponding normalized score function.

Suppose that one had access to $e_{1i}(\theta), e_{2i}(W_i;\theta)$ instead of $\hat{e}_{1i}(\theta), \hat{e}_{2i}(W_i;\theta)$. Then denote

$$\tilde{\varepsilon}^*_i(\theta) = e_{1S^*_i}(\theta_N) + W^*_{N,i}\nu_{S^*_i}(1;\theta) - (1 - W^*_{N,i})\nu_{S^*_i}(0;\theta),$$

where

$$\nu_i(w;\theta) = \left(1 + \frac{\tilde{K}(i;w,\theta)}{M}\right) e_{2\mathcal{J}_w(i)}(w;\theta).$$

Additionally set

$$\begin{aligned}\tilde{\Xi}^*(\theta) &\equiv E^*_\theta[\tilde{\varepsilon}^*_i(\theta)] \\ &= \frac{1}{N}\sum_{k=1}^N \left\{e_{1k}(\theta) + F(X_k'\theta)\nu_i(1;\theta) - \left(1 - F(X_k'\theta)\right)\nu_i(1;\theta)\right\}.\end{aligned}$$

Finally define the bootstrap estimator with the 'true' error terms $e_{1i}(\theta), e_{2i}(W_i;\theta)$ as

$$\tilde{T}^*_N(\theta) = \frac{1}{\sqrt{N}}\sum_{i=1}^N \left\{\tilde{\varepsilon}^*_i\left(\hat{\theta}^*\right) - \tilde{\Xi}^*\left(\hat{\theta}^*\right)\right\}. \tag{A.1}$$

A.1. **Proof of Theorem 1.** My proof of the bootstrap consistency builds on the method of proof of Abadie and Imbens (2016, Theorem 1). I aim to show that

$$\mathcal{L}_N^* \begin{pmatrix} T_N^*(\theta_N) \\ \sqrt{N}(\hat{\theta}_N^* - \theta_N) \\ \Lambda_N^*\left(\bar{\hat{\theta}}|\theta_N\right) \end{pmatrix} \xrightarrow{p} \mathcal{L}(\mathbf{V}); \tag{A.2}$$

$$\mathbf{V} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ -h'I_{\theta_0}h/2 \end{pmatrix}, \begin{pmatrix} \sigma^2 & c'I(\theta_0)^{-1} & -c'h \\ I(\theta_0)^{-1}c & I(\theta_0)^{-1} & -h \\ -h'c & -h & h'I(\theta_0)h \end{pmatrix}\right).$$

Given (A.2), the claim follows by similar arguments in Abadie and Imbens (2016) involving the use of Le Cam's third lemma together with a Le Cam skeleton or discretization argument as in Andreou and Werker (2011). Subsequently, I focus on proving (A.2).

To simplify notation I shall assume that $\bar{\hat{\theta}} \to \theta_0$ and $\theta_N \to \theta_0$ almost surely in $\tilde{P}_0$. This is without loss of generality as one can always convert convergence in probability (wrt $\tilde{P}_0$) to almost sure convergence (wrt $\tilde{P}_0$) using a subsequence argument.[11] Henceforth, in all of the proofs it is implicitly assumed that I am working within such a subsequence.

Lemma 1 in Appendix B implies that with probability approaching one under $\tilde{P}_0$,

$$\Lambda_N^*\left(\bar{\hat{\theta}}|\theta_N\right) = -h'S_N^*(\theta_N) - \frac{1}{2}h'I(\theta_0)h + o_{P_N^*}(1);$$
$$\sqrt{N}(\hat{\theta}_N^* - \theta_N) = I(\theta_0)^{-1}S_N^*(\theta_N) + o_{P_N^*}(1).$$

Consequently by the above it suffices for (A.2) to show

$$\mathcal{L}_N^* \begin{pmatrix} T_N^*(\theta_N) \\ S_N^*(\theta_N) \end{pmatrix} \xrightarrow{p} \mathcal{L}(\mathbf{V}_2); \qquad \mathbf{V}_2 \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & c' \\ c & I(\theta_0) \end{pmatrix}\right). \tag{A.3}$$

Now by Lemma 2 in Appendix B, it follows

$$\left\| \; T_N^*(\theta_N) - \tilde{T}_N^*(\theta_N) \; \right\| = o_{P_N^*}(1),$$

with probability approaching one under $\tilde{P}_0$. Hence to prove (A.3), it is enough to show

$$\mathcal{L}_N^* \begin{pmatrix} \tilde{T}_N^*(\theta_N) \\ S_N^*(\theta_N) \end{pmatrix} \xrightarrow{p} \mathcal{L}(\mathbf{V}_2). \tag{A.4}$$

Consider the linear combination $C_N = t_1\tilde{T}_N^*(\theta_N) + t_2'S_N^*(\theta_N)$. For any value of θ let

$$h(x;\theta) \equiv E[X_i|F(X_i'\theta) = x].$$

I can then write $C_N = N^{-1/2}\sum_{i=1}^N \delta_{N,i}^*$, where $\delta_{N,i}^* = t_2'\alpha_{N,i}^* + t_2'\beta_{N,i}^* + t_1\gamma_{N,i}^*$, with

$$\alpha_{N,i}^* = h(X_i^{*\prime}\theta_N;\theta_N)\frac{W_{N,i}^* - F\left(X_i^{*\prime}\theta_N\right)}{F\left(X_i^{*\prime}\theta_N\right)\left(1 - F\left(X_i^{*\prime}\theta_N\right)\right)} f\left(X_i^{*\prime}\theta_N\right),$$

$$\beta_{N.i}^* = \left\{X_i^* - h(X_i^{*\prime}\theta_N;\theta_N)\right\}\frac{W_{N,i}^* - F\left(X_i^{*\prime}\theta_N\right)}{F\left(X_i^{*\prime}\theta_N\right)\left(1 - F\left(X_i^{*\prime}\theta_N\right)\right)} f\left(X_i^{*\prime}\theta_N\right),$$

[11] That $\bar{\hat{\theta}} \xrightarrow{p} \theta_0$ is simply a consequence of $\hat{\theta} \xrightarrow{p} \theta_0$, since the grid size $d/\sqrt{N}$ also goes to 0 as $N \to \infty$.

and

$$\gamma^*_{N,i} = \tilde{\varepsilon}^*_i(\theta_N) - \tilde{\Xi}^*(\theta_N).$$

Observe that under the bootstrap DGP, $E^*_N[\alpha^*_{N,i}|X^*_i] = 0$ and $E^*_N[\beta^*_{N,i}|X^*_i] = 0$ by the construction of $W^*_{N,i}$; and $E^*_N[\gamma^*_{N,i}] = 0$ by the definition of $\tilde{\Xi}^*(\theta_N)$. Hence $\{\delta^*_{N,i}, i = 1 \dots N\}$ are iid zero mean random variables under P^*_N, and by the Lindberg-Feller central limit theorem for triangular arrays with iid sequences I obtain

$$\mathcal{L}^*_N(C_N) \xrightarrow{p} \mathcal{L}(\mathbf{v}); \quad \mathbf{v} \sim N(0, \sigma^2_B),$$

with

$$\sigma^2_B = \text{plim } E^*_N[\delta^{*2}_{N,i}],$$

where the plim is taken over $\tilde{P}_0$.

I characterize the variance by expanding $\delta^{*2}_{N,i} = \left(t_2'\alpha^*_{N,i} + t_2'\beta^*_{N,i} + t_1\gamma^*_{N,i}\right)^2$ and considering the bootstrap expectation of each term in turn. Under the definition of $h(.)$, it follows by the usual algebra involving Assumptions 3 & 4 that

$$E^*_N\left[\left(t_2'\alpha^*_{N,i} + t_2'\beta^*_{N,i}\right)^2\right] \xrightarrow{p} E\left[\frac{f^2(X'\theta_0)}{F(X'\theta_0)(1-F(X'\theta_0))}\left(t_2'X\right)^2\right] = t_2'I(\theta_0)t_2.$$

Thus it only remains to obtain the probability limit under $\tilde{P}_0$ of

$$\begin{aligned} V_1(\theta_N) &\equiv E^*_N\left[\left(t_1\gamma^*_{N,i}\right)\left(t_2'\alpha^*_{N,i}\right)\right], \\ V_2(\theta_N) &\equiv E^*_N\left[\left(t_1\gamma^*_{N,i}\right)^2\right], \qquad \text{and} \\ V_3(\theta_N) &\equiv E^*_N\left[\left(t_1\gamma^*_{N,i}\right)\cdot\left(t_2'\beta^*_{N,i}\right)\right]. \end{aligned}$$

A difficulty with proving the above is that within the matching function, $K_M(i;\theta_N)$, the treatments in the original sample are distributed as $W_i \sim \text{Bernoulli}(F(X_i'\theta_0))$, whereas the matches are evaluated in terms of the proximity with respect to $F(X_i'\theta_N)$. Thus, to obtain the probability limits, I make a second use of the skeleton argument of Le Cam. This exploits the discretization $\bar{\hat{\theta}}$ of $\hat{\theta}$ defined previously, and involves replacing $\bar{\hat{\theta}}$ with the local asymptotic sequence $\breve{\theta}_N = \theta_0 + \breve{h}/\sqrt{N}$, for some $\breve{h} \in \mathbb{R}$. To this end, I employ the following notation:

Parametrize the multinomial random variables $\mathbf{M}$ (Section 3.2) as $\mathbf{M}(\theta)$ for the case when the estimated propensity score is given by θ (rather than $\bar{\hat{\theta}}$). Denote by $\mathbf{U} = (U_1, \dots, U_N)$ a vector of N independent uniform random variables corresponding to each observation, and drawn independently of $\mathbf{W}, \mathbf{X}, \mathbf{Y}$. Then it is possible to couple $\mathbf{M}(\theta) = \mathcal{H}(\mathbf{U}; F(\mathbf{X}'\theta))$, where $\mathcal{H}(.; F(\mathbf{X}'\theta))$ is some transformation indexed by the parameter θ.[12] I represent by $\bar{P}_\theta$ the probability law for $\mathbf{W}, \mathbf{X}, \mathbf{Y}, \mathbf{U}$ with $\mathbf{W} \sim \text{Bernoulli}(F(\mathbf{X}'\theta))$, and let $\bar{E}_\theta[.]$ denote the corresponding expectation over $\bar{P}_\theta$. A convenient feature of $\bar{P}_\theta$ (as compared to $\tilde{P}_\theta$) is that it doesn't depend on the value of $\breve{\theta}_N$; indeed, this is the reason for employing the coupling. Given $\breve{\theta}_N$, I construct

[12] $\mathcal{H}(,;F(\mathbf{X}'\theta))$ can be interpreted as a function that transforms a uniformly distributed random variable into a single-draw multinomial random variable. Note that knowledge of $F(\mathbf{X}'\theta)$ uniquely pins down the quantiles $\left\{\pi_1(\theta), \dots, \pi_{q_N-1}(\theta)\right\}$ and the number of treated and untreated populations denoted by $N_1(l;\theta)$, $N_0(l;\theta)$ in each partition. Thus the uniform random variable can be transformed into the multinomial random variable, $M(i;\theta)$, for each observation i, by partitioning the unit interval into $N_w(l;\theta)$ equi-spaced segments.

a local asymptotic sequence for the bootstrap indexed by $\bar{\theta}_N = \breve{\theta}_N + h/\sqrt{N}$. Let $\bar{P}_N^* \equiv P_{\bar{\theta}_N}^*$ denote the bootstrap probability indexed by $\bar{\theta}_N$, and $\bar{E}_N^*[.]$ the bootstrap expectation under $\bar{P}_N^*$. For convenience set $\bar{P}_N \equiv \bar{P}_{\bar{\theta}_N}$ and $\bar{P}_0 \equiv \bar{P}_{\theta_0}$, with the corresponding expectation operators $\bar{E}_N[.] \equiv \bar{E}_{\bar{\theta}_N}[.]$ and $\bar{E}_0[.] \equiv \bar{E}_{\theta_0}[.]$. Finally, I also introduce the quantities

$$V_1(h,\theta) \equiv E^*_{\theta+h/\sqrt{N}}\left[\left(t_1\gamma_i^*\left(\theta+\frac{h}{\sqrt{N}};\theta\right)\right)\left(t_2'\alpha_i^*\left(\theta+\frac{h}{\sqrt{N}}\right)\right)\right];$$
$$V_2(h,\theta) \equiv E^*_{\theta+h/\sqrt{N}}\left[\left(t_1\gamma_i^*\left(\theta+\frac{h}{\sqrt{N}};\theta\right)\right)^2\right]; \quad \text{and}$$
$$V_3(h,\theta) \equiv E^*_{\theta+h/\sqrt{N}}\left[\left(t_1\gamma_i^*\left(\theta+\frac{h}{\sqrt{N}};\theta\right)\right)\left(t_2'\beta_i^*\left(\theta+\frac{h}{\sqrt{N}}\right)\right)\right],$$

where, for any θ_1, θ_2,

$$\alpha_i^*(\theta_1) = h^*(X_i^{*\prime}\theta_1;\theta_1)\frac{W_i^* - F\left(X_i^{*\prime}\theta_1\right)}{F\left(X_i^{*\prime}\theta_1\right)\left(1-F\left(X_i^{*\prime}\theta_1\right)\right)}f\left(X_i^{*\prime}\theta_1\right),$$

$$\beta_i^*(\theta_1) = \left\{X_i^* - h^*(X_i^{*\prime}\theta_1;\theta_1)\right\}\frac{W_i^* - F\left(X_i^{*\prime}\theta_1\right)}{F\left(X_i^{*\prime}\theta_1\right)\left(1-F\left(X_i^{*\prime}\theta_1\right)\right)}f\left(X_i^{*\prime}\theta_1\right); \text{ and}$$

$$\gamma_i^*(\theta_1;\theta_2) = \tilde{\varepsilon}_i^*\left(\theta_1;\theta_2\right) - E_{\theta_1}^*[\tilde{\varepsilon}_i^*\left(\theta_1;\theta_2\right)].$$

Here $\tilde{\varepsilon}_i^*\left(\theta_1;\theta_2\right)$ is defined analogously to $\tilde{\varepsilon}_i^*\left(\theta_1\right)$ but with $\bar{\hat{\theta}}$ replaced by θ_2; in particular, this involves replacing $\mathbf{M}$ in the definition of $\tilde{\varepsilon}_i^*\left(\theta_1\right)$ with $\mathbf{M}(\theta_2) = \mathcal{H}(\mathbf{U}; F(\mathbf{X}'\theta_2))$. It is useful to observe that $V_k(\theta_N) = V_k\left(h;\bar{\hat{\theta}}\right)$ for $k = 1,2,3$.

In Lemmas 3 - 5 in Appendix B, I show that for any bounded $\breve{h}$ within the definition of $\breve{\theta}_N$,[13]

$$\begin{aligned} V_1(h,\breve{\theta}_N) &= o_{\bar{P}_N}(1); \\ V_2(h,\breve{\theta}_N) &= t_1^2\sigma^2 + o_{\bar{P}_N}(1); \text{ and} \\ V_3(h,\breve{\theta}_N) &= 2t_1c't_2 + o_{\bar{P}_N}(1). \end{aligned} \tag{A.5}$$

Then, employing a version of Le Cam's skeleton argument, I show that

$$\begin{aligned} V_1\left(h,\bar{\hat{\theta}}\right) &= o_{\bar{P}_0}(1); \\ V_2\left(h,\bar{\hat{\theta}}\right) &= t_1^2\sigma^2 + o_{\bar{P}_0}(1); \text{ and} \\ V_3\left(h,\bar{\hat{\theta}}\right) &= 2t_1c't_2 + o_{\bar{P}_0}(1). \end{aligned}$$

I illustrate the reasoning for the case of $V_2\left(h,\bar{\hat{\theta}}\right)$; the others can be argued similarly. Note that $\bar{P}_N$ and $\bar{P}_0$ are mutually contiguous by the usual arguments involving Le Cam's first lemma. Thus by (A.5) and contiguity, I have $V_2(h,\breve{\theta}_N) = t_1^2\sigma^2 + o_{\bar{P}_0}(1)$. Let $\mathbf{v}$ denote the asymptotic normal limit of $\sqrt{N}(\hat{\theta}-\theta_0)$ under $\bar{P}_0$. Then for any $j \in \mathbb{Z}^d$,

$$\mathcal{L}_{\bar{P}_0}\begin{pmatrix} V_2\left(h,\theta_0+dj/\sqrt{N}\right) - t_1^2\sigma^2 \\ \sqrt{N}(\hat{\theta}-\theta_0) - dj \end{pmatrix} \to \mathcal{L}\begin{pmatrix} 0 \\ \mathbf{v} - dj \end{pmatrix}.$$

[13]For equation (A.5), note that the matches are now evaluated in terms of proximity wrt $F(X_i'\bar{\theta}_N)$, which is also the propensity score characterizing the distribution of the treatments since $W_i \sim \text{Bernoulli}(F(X_i'\bar{\theta}_N))$ under $\bar{P}_N$.

Additionally, the following events are equivalent for each $j \in \mathbb{Z}^d$:

$$\left\{\sqrt{N}\left(\bar{\hat{\theta}} - \theta_0\right) = dj\right\} \equiv \left\{-\frac{d}{2}i < \sqrt{N}(\hat{\theta} - \theta_0) - dj \leq -\frac{d}{2}i\right\},$$

where i denotes a vector of ones of dimension d. Combining the above gives that for each $j \in \mathbb{Z}^d$, and any $\epsilon > 0$,

$$\bar{P}_0\left\{\left|V_2\left(h, \theta_0 + dj/\sqrt{N}\right) - t_1^2\sigma^2\right| > \epsilon \cap \sqrt{N}\left(\bar{\hat{\theta}} - \theta_0\right) = dj\right\} \to 0$$

as $N \to \infty$. Hence for each $C < \infty$,

$$\begin{aligned}
&\bar{P}_0\left\{\left|V_2\left(h, \bar{\hat{\theta}}\right) - t_1^2\sigma^2\right| > \epsilon \cap \left|\sqrt{N}\left(\bar{\hat{\theta}} - \theta_0\right)\right| \leq C\right\} \\
&= \sum_{j \in \mathbb{Z}^d : d|j| \leq C} \bar{P}_0\left\{\left|V_2\left(h, \theta_0 + dj/\sqrt{N}\right) - t_1^2\sigma^2\right| > \epsilon \cap \sqrt{N}\left(\bar{\hat{\theta}} - \theta_0\right) = dj\right\} \to 0.
\end{aligned}$$

Since $\sqrt{N}\left(\bar{\hat{\theta}} - \theta_0\right)$ is $O_{\bar{P}_0}(1)$, letting $C \to \infty$ above implies $V_2\left(h, \bar{\hat{\theta}}\right) = t_1^2\sigma^2 + o_{\bar{P}_0}(1)$, as claimed.

By definition, the probability distribution of $V_2(\theta_N)$ under $\tilde{P}_0$ is equivalent to that of $V_2\left(h, \bar{\hat{\theta}}\right)$ under $\bar{P}_0$; and similarly for the distribution of $V_1(\theta_N), V_3(\theta_N)$ under $\tilde{P}_0$. Combining the above results, I have thus shown

$$\sigma_B^2 = t_1^2\sigma^2 + 2t_1 c' t_2 + t_2' I(\theta_0) t_2.$$

This proves (A.3), which completes the proof of the theorem.

A.2. **Proof of Corollary 1.** Let $F(.)$ denote the cdf of $\mathbf{v} \sim N\left(0, \sigma^2 - c' I_{\theta_0}^{-1} c\right)$. By taking L (cf Step 7 in Section 3.3) sufficiently large, the claim follows if I show that

$$E_{\mathbf{M}}\left[F_n^*(t)|\mathbf{Z}\right] \xrightarrow{p} F(t) + O(d) \tag{A.6}$$

uniformly over $t \in \mathbb{R}$ under P_0 (here $E_{\mathbf{M}}[.|\mathbf{Z}]$ denotes the expectation over $\mathbf{M}$ conditional on the data). But by the Glivenko-Cantelli theorem, pointwise convergence implies uniform convergence, hence it suffices to show (A.6) holds for each $t \in \mathbb{R}$ under P_0. So I fix some arbitrary $t \in \mathbb{R}$.

Recall the definitions of $\bar{P}_0$ and $\mathbf{U}$ from the proof of Theorem 1. By Theorem 1, $F_n^*(t) \xrightarrow{p} F(t) + O(d)$ under $\bar{P}_0$. By employing a subsequence argument, the convergence in probability (wrt $\bar{P}_0$) can be converted to almost sure convergence (wrt $\bar{P}_0$). By this construction,

$$F_n^*(t) \to F(t) + O(d), \text{ a.s} - \bar{P}_0. \tag{A.7}$$

Note that by independence (of $\mathbf{Z}$,$\mathbf{U}$), $\bar{P}_0$ is equivalent to the product measure, $P_0 \times P_U$, of the respective marginal measures, P_0, P_U, of $\mathbf{Z}$ and $\mathbf{U}$. Denote by Ω the set of all realizations, z, of $\mathbf{Z}$ for which $F_n^*(t) \to F(t) + O(d)$, a.s $- P_U$. By the independence of $\mathbf{Z}$ and $\mathbf{U}$, it must be that $P_0(\Omega) = 1$ for (A.7) to hold. At the same time, the dominated convergence theorem gives

$$E_{\mathbf{U}}\left[F_n^*(t)\right] \to F(t) + O(d) \tag{A.8}$$

for each $z \in \Omega$; hence (A.8) holds almost surely over P_0. Since $E_{\mathbf{M}}\left[F_n^*(t)|\mathbf{Z}\right] \equiv E_{\mathbf{U}}\left[F_n^*(t)\right]$, this immediately proves (A.6).

Appendix B. Lemmas

Hereafter, I shall use the notation wpa1-$\tilde{P}_0$ as a shorthand for 'with probability approaching one under $\tilde{P}_0$'.

I also introduce the following notation: For $w = 0, 1$ let

$$e_{3i}(w;\theta) = \bar{\mu}(w; X_i) - \mu(w; F(X_i'\theta)).$$

Also let

$$e_{4i}(W_i;\theta) = Y_i - \bar{\mu}(W_i, X_i).$$

Note that it is possible to decompose $e_{2i}(W_i;\theta) = e_{3i}(W_i;\theta) + e_{4i}(W_i;\theta)$.

In Lemmas 3-5, I work with the local asymptotic sequence $\breve{\theta}_N = \theta_0 + \breve{h}/\sqrt{N}$ in place of $\bar{\hat{\theta}}$. To this end, I employ the notation introduced in Appendix A. Represent by $\left\{\pi_1(\breve{\theta}_N), \ldots, \pi_{q_N-1}(\breve{\theta}_N)\right\}$ the sample q_N-quantiles of $F(X'\breve{\theta}_N)$ with $\pi_0(\breve{\theta}_N) = 0$ and $\pi_{q_N}(\breve{\theta}_N) = 1$. I introduce $l(i)$ as the block index of observation i wrt $F(X_i'\breve{\theta}_N)$, i.e $l(i) = k$ if $\pi_{l-1}(\breve{\theta}_N) \leq F(X_i'\breve{\theta}_N) < \pi_l(\breve{\theta}_N)$. Also, denote by $S_w(l;\theta)$ the set of all observations with $W_i = w$ whose propensity scores evaluated at θ - i.e $F(X_i'\theta)$ - lie in the l-th block (even as the blocks themselves are obtained from quantiles of $F(X'\breve{\theta}_N)$):

$$S_w(l;\theta) \equiv \left\{i : \ \pi_{l-1}(\breve{\theta}_N) \leq F(X_i'\theta) < \pi_l(\breve{\theta}_N) \ \cap \ W_i = w\right\}.$$

Based on the above, I set $S(l;\theta) = S_1(l;\theta) \cup S_0(l;\theta)$. Furthermore, I also denote

$$N_0(l;\theta) = \#S_0(l;\theta); \quad N_1(l;\theta) = \#S_1(l;\theta); \quad N(l;\theta) = N_0(l;\theta) + N_1(l;\theta),$$

where $\#A$ denotes the cardinality of any set A.

For $w = 0, 1$, the average matching function, defined as the expectation of $\tilde{K}_M(i; w, \theta)$ over $\mathbf{U}$ given $(\mathbf{X}, \mathbf{W})$, is represented by

$$\bar{K}_M(i; w, \theta) = \begin{cases} K_M(i;\theta) & \text{if } w = W_i \\ \frac{1}{N_w\left(l(i);\breve{\theta}_N\right)} \sum_{j \in S_w\left(l(i);\breve{\theta}_N\right)} K_M(j;\theta) & \text{if } w \neq W_i \ . \end{cases}$$

Slightly abusing notation, I suppress indexing the quantities $\tilde{K}_M(.), \bar{K}_M(.), \nu(.), l(.)$ with the additional label $\breve{\theta}_N$. However it should be understood implicitly that these quantities are now constructed by replacing $\bar{\hat{\theta}}$ with $\breve{\theta}_N$. Finally, I also define (again suppressing the index with respect to $\breve{\theta}_N$),

$$\nu_{(3)i}(w;\theta) = \left(1 + \frac{\tilde{K}_M(i; w, \theta)}{M}\right) e_{3\mathcal{J}_w(i)}(w;\theta);.$$

$$\nu_{(4)i}(w;\theta) = \left(1 + \frac{\tilde{K}_M(i; w, \theta)}{M}\right) e_{4\mathcal{J}_w(i)}(w;\theta).$$

Lemma 1. *Suppose that $\bar{\hat{\theta}} \to \theta_0$ a.s-$\tilde{P}_0$. Then under Assumptions 1-5, wpa1-$\tilde{P}_0$,*

$$\Lambda_N^*\left(\bar{\hat{\theta}}|\theta_N\right) = -h'S_N^*(\theta_N) - \frac{1}{2}h'I(\theta_0)h + o_{P_N^*}(1), \tag{B.1}$$

and

$$\sqrt{N}(\hat{\theta}_N^* - \theta_N) = I(\theta_0)^{-1}S_N^*(\theta_N) + o_{P_N^*}(1). \tag{B.2}$$

Proof. Define

$$\hat{I}_N^*(\theta) = \frac{1}{N}\frac{d^2 L(\theta|\mathbf{Z}_N^*)}{d\theta d\theta'}; \quad \breve{I}_N^*(\theta) = \frac{1}{N}\sum_{i=1}^N \psi_{N,i}^*(\theta_N)\psi_{N,i}^{*\prime}(\theta_N).$$

Under Assumptions 3(i)-(ii), I can show that $\sup_{\theta\in\mathcal{N}} E_N^* \left\|\hat{I}_N^*(\theta) - \breve{I}_N^*(\theta)\right\|^2 \xrightarrow{p} 0$. The same assumptions also suffice to show $\sup_{\theta\in\mathcal{N}} \left\|\hat{I}_N^*(\theta) - I^*(\theta_N)\right\| \xrightarrow{p} 0$, where

$$I^*(\theta_N) \equiv E_N^*\left[\psi_{N,i}^*(\theta_N)\psi_{N,i}^{*\prime}(\theta_N)\right] = \frac{1}{N}\sum_{i=1}^N X_i X_i' \frac{f^2(X_i'\theta_N)}{F(X_i'\theta_N)(1-F(X_i'\theta_N))}.$$

The term inside the summation is non-negative and uniformly bounded for all N sufficiently large (by Assumptions 3(i)-(ii)). Consequently, by Assumption 4 and standard arguments, $\sup_{\theta\in\mathcal{N}} \|I_N^*(\theta) - I(\theta)\| \xrightarrow{p} 0$. Combining the above proves that wpa1-$\tilde{P}_0$,

$$\sup_{\theta\in\mathcal{N}} \left\|\hat{I}_N^*(\theta) - I(\theta)\right\| = o_{P_N^*}(1). \tag{B.3}$$

Under Assumptions 3(i)-(ii), standard second order Taylor expansion arguments assure that for any $\epsilon > 0$,

$$\sup_{\theta\in\mathcal{N}} P_\theta^*\left(\left|\Lambda_N^*\left(\theta + h/\sqrt{N}|\theta\right) - h'S_N^*(\theta) + \frac{1}{2}h'\hat{I}_N^*(\theta)h\right| > \epsilon\right) \xrightarrow{p} 0. \tag{B.4}$$

The above implies

$$P_N^*\left(\left|\Lambda_N^*\left(\bar{\theta}|\theta_N\right) + h'S_N^*(\theta_N) + \frac{1}{2}h'\hat{I}_N^*(\theta_N)h\right| > \epsilon\right) \xrightarrow{p} 0. \tag{B.5}$$

Combined with (B.3), I have thus shown the following: wpa1-$\tilde{P}_0$,

$$\Lambda_N^*\left(\bar{\theta}|\theta_N\right) = -h'S_N^*(\theta_N) - \frac{1}{2}h'I(\theta_0)h + o_{P_N^*}(1). \tag{B.6}$$

This proves the first claim of the lemma.

The limiting distribution of $S_N^*(\theta_N)$ under P_N^* can be ascertained using the Lindberg-Feller central limit theorem for triangular arrays. Indeed, $\psi_{N,i}^*(.)$ is mean zero and uniformly bounded by Assumptions 3(i)-(ii), which implies the Lyapunov condition is trivially satisfied. The bootstrap variance of $\psi_{N,i}^*(.)$ is also simply $I^*(\theta_N)$. Thus by the arguments leading to (B.3), I obtain

$$\mathcal{L}_N^*(S_N^*(\theta_N)) \xrightarrow{p} \mathcal{L}(\mathbf{v_2}) \tag{B.7}$$

with $\mathbf{v_2} \sim N(0, I(\theta_0))$. From (B.6) and (B.7), it follows by an application of Le Cam's first lemma that P_N^* and P^* are mutually contiguous, wpa1-$\tilde{P}_0$.

I shall now prove that wpa1-$\tilde{P}_0$,

$$\left\|\hat{\theta}^* - \bar{\hat{\theta}}\right\| = o_{P^*}(1). \tag{B.8}$$

I shall show $P^*\left(\left\|\hat{\theta}^* - \theta_0\right\| > \epsilon\right) \xrightarrow{p} 0$ for any $\epsilon > 0$. Since $\bar{\hat{\theta}} \to \theta_0$ a.s-$\tilde{P}_0$, this proves (B.8). To this end it suffices to verify the conditions for the consistency result of Newey and McFadden (1994, Theorem 2.7). Note that each summand within $L(\theta|\mathbf{W}^*, \mathbf{X}^*)$ is uniformly bounded wpa1-$\tilde{P}_0$ (due to Assumptions 3(i)-(ii) and 5(ii)); hence standard arguments using Markov's inequality

assure that wpa1-$\tilde{P}_0$,

$$\frac{1}{N}L(\theta|\mathbf{W}^*,\mathbf{X}^*)-\frac{1}{N}E^*\left[L(\theta|\mathbf{W}^*,\mathbf{X}^*)\right]=o_{P^*}(1).$$

Now it is possible to expand

$$\frac{1}{N}E^*\left[L(\theta|\mathbf{W}^*,\mathbf{X}^*)\right]=\frac{1}{N}\sum_{i=1}^{N}A_{1N,i}(\theta),$$

where

$$A_{1N,i}(\theta)=F\left(X_i'\bar{\hat{\theta}}\right)\ln F\left(X_i'\theta\right)+\left(1-F\left(X_i'\bar{\hat{\theta}}\right)\right)\ln\left(1-F\left(X_i'\theta\right)\right).$$

The uniform law of large numbers, together with the fact $\bar{\hat{\theta}}\to\theta_0$ a.s-$\tilde{P}_0$, assures

$$\frac{1}{N}\sum_{i=1}^{N}A_{1N,i}(\theta)\xrightarrow{p}E_0\left[F\left(X_i'\theta_0\right)\ln F\left(X_i'\theta\right)+\left(1-F\left(X_i'\theta_0\right)\right)\ln\left(1-F\left(X_i'\theta\right)\right)\right]\equiv M(\theta).$$

I have thus shown that pointwise for each θ,

$$\frac{1}{N}L(\theta|\mathbf{W}^*,\mathbf{X}^*)=M(\theta)+o_{P^*}(1),$$

wpa1-$\tilde{P}_0$. Clearly $M(\theta)$ is concave. Furthermore, since $E_0[X_iX_i']$ is positive definite, θ_0 is the unique maximiser of $M(\theta)$ (see Newey and McFadden, 1994, Example 2.1 in p.2125). Combining the above, it can be noted that all the conditions for applying Theorem 2.7 of Newey and McFadden (1994) are verified. This proves (B.8).

I can now prove the second claim of the lemma. Using (B.8) and Assumption 3(ii) (finite second derivatives for $F(.)$), the usual linearization arguments can be applied show that wpa1-$\tilde{P}_0$,

$$\sqrt{N}\left(\hat{\theta}^*-\bar{\hat{\theta}}\right)=\hat{I}_N^*\left(\bar{\hat{\theta}}\right)^{-1}S_N^*\left(\bar{\hat{\theta}}\right)+o_{P^*}(1).$$

Contiguity, proven earlier, then gives

$$\sqrt{N}\left(\hat{\theta}_N^*-\theta_N\right)=-h+\hat{I}_N^*\left(\bar{\hat{\theta}}\right)^{-1}S_N^*\left(\bar{\hat{\theta}}\right)+o_{P_N^*}(1), \tag{B.9}$$

wpa1-$\tilde{P}_0$. Using (B.4) and (B.3), together with Assumption 4 (which implies $I(.)$ is continuous on $\mathcal{N}$), I adapt the arguments of Bickel et al (1998, Proposition 2.1.2) to show that wpa1-$\tilde{P}_0$,

$$\left\|S_N^*(\theta_N)-S_N^*\left(\bar{\hat{\theta}}\right)-\hat{I}_N^*\left(\bar{\hat{\theta}}\right)h\right\|=o_{P_N^*}(1).$$

Substituting the above in (B.9), and using (B.3) proves (B.2), the second claim of the lemma. □

Lemma 2. *Under Assumptions 1-5 and $\theta_N\to\theta_0$ a.s-$\tilde{P}_0$, it holds $\left|T_N^*(\theta_N)-\tilde{T}_N^*(\theta_N)\right|=o_{P_N^*}(1)$, wpa1-$\tilde{P}_0$.*

Proof. Define $\varrho_{N,i}^*(\theta_N)=\varepsilon_i^*\left(\theta_N\right)-\tilde{\varepsilon}_i^*\left(\theta_N\right)$, and observe that

$$T_N^*(\theta_N)-\tilde{T}_N^*(\theta_N)=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{\varrho_{N,i}^*(\theta_N)-E^*[\varrho_{N,i}^*(\theta_N)]\right\}.$$

Hence, I obtain

$$E^*\left|T_N^*(\theta_N)-\tilde{T}_N^*(\theta_N)\right|^2\leq E^*\left|\varrho_{N,i}^*(\theta_N)\right|^2\equiv A_N.$$

I can further bound $A_N \leq 2(A_{1N} + A_{2N}^{(0)} + A_{2N}^{(1)})$, where, for $w = 0, 1$

$$A_{1N} = E^* \left| \hat{e}_{1S_i^*}(\theta_N) - e_{1S_i^*}(\theta_N) \right|^2, \quad \text{and}$$
$$A_{2N}^{(w)} = E^* \left| \hat{\nu}_{S_i^*}(w; \theta_N) - \nu_{S_i^*}(w; \theta_N) \right|^2.$$

By Assumption 5, standard arguments assure $A_{1N} \xrightarrow{p} 0$ under $\tilde{P}_0$. Consequently I focus on the term $A_{2N}^{(1)}$. By the definition of $\tilde{K}_M(i; w, \theta)$, there exists some constant $C < \infty$ for which

$$\begin{aligned} A_{2N}^{(1)} &\leq C \left\{ 1 + \sup_{1 \leq i \leq N} K_M^2(i; \theta_N) \right\} \times \frac{1}{N} \sum_{i=1}^{N} \left\{ e_{2\mathcal{J}_1(i)}(1; \theta_N) - \hat{e}_{2\mathcal{J}_1(i)}(1; \theta_N) \right\}^2 \\ &\equiv \Gamma_{1N} \times \Gamma_{2N}. \end{aligned}$$

By Lemma 6 in Appendix C, $\Gamma_{1N} = o_p(N^{\xi/2})$ under $\tilde{P}_0$ for any ξ arbitrarily small. Next consider the term Γ_{2N}: The maximum number of times an observation i is used as a secondary match is bounded by the matching function for the nearest neighbor matching, given by $K_{\text{NN}}(i)$. Consequently,

$$\Gamma_{2N} \leq \left\{ 1 + \sup_{1 \leq i \leq N} K_{\text{NN}}(i) \right\} \sup_{\theta \in \mathcal{N}} \frac{1}{N} \sum_{i=1}^{N} \{ \hat{e}_{2i}(1; \theta) - e_{2i}(1; \theta) \}^2.$$

Now, by Abadie and Imbens (2006, Lemma 3), $\sup_{1 \leq i \leq N} K_{\text{NN}}(i) = o_p(N^{\xi/2})$ under $\tilde{P}_0$ for any ξ arbitrarily small. Combined with Assumption 5, this assures $\Gamma_{2N} = O_p(N^{-\xi/2})$ under $\tilde{P}_0$. Taken together, the above imply $A_{2N}^{(1)} \xrightarrow{p} 0$. Analogous arguments for $w = 0$ similarly imply $A_{2N}^{(0)} \xrightarrow{p} 0$. This completes the proof of the lemma. □

Lemma 3. *Under Assumptions 1-5 and $\bar{\theta}_N \to \theta_0$, it holds $V_1(h, \breve{\theta}_N) = o_{\bar{P}_N}(1)$.*

Proof. I first note that

$$\bar{E}_N^* \left[\left(t_1 \gamma_i^*(\bar{\theta}_N; \breve{\theta}_N \right) \left(t_2' \alpha_i^*(\theta_N) \right) \right] = \bar{E}_N^* \left[\left(t_1 \tilde{\varepsilon}_{N,i}^*(\bar{\theta}_N; \breve{\theta}_N) \right) \left(t_2' \alpha_i^*(\bar{\theta}_N) \right) \right]$$

since α_i^* is mean zero under $\bar{E}_N^*[.]$. Decompose

$$\tilde{\varepsilon}_{N,i}^*(\bar{\theta}_N; \breve{\theta}_N) = e_{1S_i^*}(\bar{\theta}_N) + W_i^* \nu_{S_i^*}(1; \bar{\theta}_N) - (1 - W_i^*) \nu_{S_i^*}(0; \bar{\theta}_N).$$

Now based on the bootstrap DGP it is straightforward to verify $\bar{E}_N^* \left[e_{1S_i^*}(\bar{\theta}_N) \left(t_2' \alpha_i^*(\bar{\theta}_N) \right) \right] = 0$. Hence the claim follows if I show that

$$\begin{aligned} Q_N^{(1)}(\bar{\theta}_N) &\equiv \bar{E}_N^* \left[\left(W_i^* \nu_{S_i^*}(1; \bar{\theta}_N) \right) \left(t_2' \alpha_i^*(\bar{\theta}_N) \right) \right] = o_{\bar{P}_N}(1); \\ Q_N^{(0)}(\bar{\theta}_N) &\equiv \bar{E}_N^* \left[\left((1 - W_i^*) \nu_{S_i^*}(0; \bar{\theta}_N) \right) \left(t_2' \alpha_i^*(\bar{\theta}_N) \right) \right] = o_{\bar{P}_N}(1). \end{aligned}$$

I show $Q_N^{(1)}(\bar{\theta}_N) \xrightarrow{p} 0$ under $\bar{P}_N$; that $Q_N^{(0)}(\bar{\theta}_N) \xrightarrow{p} 0$ follows by similar reasoning. To this end, first define the quantity

$$\tau(X_i' \bar{\theta}_N) = t_1 t_2' h \left(X_i \bar{\theta}_N; \bar{\theta}_N \right) f \left(X_i' \bar{\theta}_N \right).$$

Due to Assumption 3(i), which implies X_i^* is bounded, it follows $h(.; \bar{\theta}_N)$ is uniformly bounded over its domain for all $\bar{\theta}_N$. Combined with Assumption 3(ii) (boundedness of $f(.)$), this implies $\tau(X_i' \bar{\theta}_N) \leq C < \infty$ uniformly in both i and N .

Taking the bootstrap expectations, I obtain after some algebra

$$Q_N^{(1)}(\bar{\theta}_N) = \frac{1}{N}\sum_{i=1}^{N} \tau(X_i'\bar{\theta}_N)\left(1 + \frac{\tilde{K}_M(i;1,\bar{\theta}_N)}{M}\right) e_{2\mathcal{J}_1(i)}\left(1;\bar{\theta}_N\right).$$

I can decompose $Q_N^{(1)}(\bar{\theta}_N)$ further as

$$Q_N^{(1)}(\bar{\theta}_N) = \frac{1}{N}\sum_{i=1}^{N} \vartheta_{(3)N,i} + \frac{1}{N}\sum_{i=1}^{N} \vartheta_{(4)N,i} \equiv Q_{3N}^{(1)}(\bar{\theta}_N) + Q_{4N}^{(1)}(\bar{\theta}_N),$$

where

$$\vartheta_{(3)N,i} = \tau(X_i'\bar{\theta}_N)\left(1 + \frac{\tilde{K}_M(i;1,\bar{\theta}_N)}{M}\right) e_{3\mathcal{J}_1(i)}\left(1;\bar{\theta}_N\right);$$

$$\vartheta_{(4)N,i} = \tau(X_i'\bar{\theta}_N)\left(1 + \frac{\tilde{K}_M(i;1,\bar{\theta}_N)}{M}\right) e_{4\mathcal{J}_1(i)}\left(1;\bar{\theta}_N\right).$$

First consider the term $Q_{4N}^{(1)}(\bar{\theta}_N)$: For each i, $\bar{E}_N[\vartheta_{(4)N,i}|\mathbf{W},\mathbf{X},\mathbf{U}] = 0$ due to the definition of $e_{4i}(.)$. Furthermore, I also have $\bar{E}_N[\vartheta_{(4)N,i}\vartheta_{(4)N,,j}|\mathbf{W},\mathbf{X},\mathbf{U}] = 0$ for all i,j for which $\mathcal{J}_1(i) \neq \mathcal{J}_1(j)$. Denoting $\mathcal{S}_k = \{i \in \{1,\ldots,N\} : \mathcal{J}_1(i) = k\}$, I note that the cardinality of $\mathcal{S}_k$ is bounded by $K_{\mathrm{NN}}(k)$. Hence it follows that the number of pairs (i,j) for which $\bar{E}_N[\vartheta_{(4)N,i}\vartheta_{(4)N,,j}|\mathbf{W},\mathbf{X},\mathbf{U}] \neq 0$ is bounded above by $N\sup_{1\leq k\leq N} K_{\mathrm{NN}}(k)$. Now by Assumption 3(v) (which assures $\sup_x E_0[Y^4|X=x] \leq C < \infty$), it follows $\sup_{1\leq i\leq N}\bar{E}_N\left[|e_{4i}(1;\theta)|^2|\mathbf{W},\mathbf{X},\mathbf{U}\right] \leq C < \infty$ uniformly over $\theta \in \mathcal{N}$ (note that $\mathbf{U}$ is independent of $\mathbf{Y}_1, \mathbf{Y}_0$ by definition). Thus, by the Markov inequality and the boundedness of $\tau(.)$, there exists some $C_1 < \infty$for which

$$\begin{aligned}\bar{E}_N\left[\left\{Q_{4N}^{(1)}(\bar{\theta}_N)\right\}^2|\mathbf{W},\mathbf{X},\mathbf{U}\right] &\leq C_1 N^{-1}\left\{1 + \sup_{1\leq i\leq N} K_{\mathrm{NN}}(i)\right\}\left\{1 + \sup_{1\leq i\leq N}\tilde{K}_M^2(i;1,\bar{\theta}_N)\right\} \\ &= C_1 N^{-1}\left\{1 + \sup_{1\leq i\leq N} K_{\mathrm{NN}}(i)\right\}\left\{1 + \sup_{1\leq i\leq N} K_M^2(i;\bar{\theta}_N)\right\}.\end{aligned}$$

Using the result of Abadie and Imbens (2006, Lemma 3), $\bar{E}_N[\sup_{1\leq i\leq N} K_{\mathrm{NN}}^r(i)] = O(N^{\xi})$ for any finite r, and some $\xi > 0$ arbitrarily small. Taking a further expectation on both sides of the above equation and employing Lemma 6, together with Holder's inequality, gives $\bar{E}_N\left[\left\{Q_{4N}^{(1)}(\bar{\theta}_N)\right\}^2\right] = O(N^{-(1-\xi)})$ for some $\xi > 0$ arbitrarily small. This proves $Q_{4N}^{(1)}(\bar{\theta}_N) = o_{\bar{P}_N}(1)$.

Next consider the term $Q_{3N}^{(1)}(\bar{\theta}_N)$. First I successively approximate this term by the quantities $Q_{31N}^{(1)}(\bar{\theta}_N)$, $Q_{32N}^{(1)}(\bar{\theta}_N)$, where

$$Q_{31N}^{(1)}(\bar{\theta}_N) = \frac{1}{N}\sum_{i=1}^{N}\vartheta_{(31)N,i}; \quad \vartheta_{(31)N,i} = \tau(X_i'\bar{\theta}_N)\left(1 + \frac{\tilde{K}_M(i;1,\bar{\theta}_N)}{M}\right) e_{3i}\left(1;\bar{\theta}_N\right);$$

$$Q_{32N}^{(1)}(\bar{\theta}_N) = \frac{1}{N}\sum_{i=1}^{N}\vartheta_{(32)N,i}; \quad \vartheta_{(31)N,i} = \tau(X_i'\bar{\theta}_N)\left(1 + \frac{\bar{K}_M(i;1,\bar{\theta}_N)}{M}\right) e_{3i}\left(1;\bar{\theta}_N\right);$$

In the first case, by Lemma 6,

$$\begin{aligned}\left|Q_{3N}^{(1)}(\bar{\theta}_N) - Q_{31N}^{(1)}(\bar{\theta}_N)\right| &\leq C\left\{1 + \sup_{1\leq i\leq N} K_M(i;\bar{\theta}_N)\right\} \max_{1\leq i\leq N}\left|e_{3\mathcal{J}_1(i)}(1;\bar{\theta}_N) - e_{3i}(1;\bar{\theta}_N)\right| \\ &= O_{\bar{P}_N}(N^{\xi}) \cdot \max_{1\leq i\leq N}\left|e_{3\mathcal{J}_1(i)}(1;\bar{\theta}_N) - e_{3i}(1;\bar{\theta}_N)\right|.\end{aligned}$$

The last term can in turn be bounded as

$$\begin{aligned}&\max_{1\leq i\leq N}\left|e_{3\mathcal{J}_1(i)}(1;\bar{\theta}_N) - e_{3i}(1;\bar{\theta}_N)\right| \\ &\leq \max_{1\leq i\leq N}\left|\bar{\mu}(1;X_{\mathcal{J}_1(i)}) - \bar{\mu}(1;X_i)\right| + \max_{1\leq i\leq N}\left|\mu\left(1;F(X'_{\mathcal{J}_1(i)}\bar{\theta}_N);\bar{\theta}_N\right) - \mu\left(1;F(X_i'\bar{\theta}_N);\bar{\theta}_N\right)\right| \\ &\leq \max_{1\leq i\leq N}\left\|X_{\mathcal{J}_1(i)} - X_i\right\| = O_{\bar{P}_N}(N^{-1/k}),\end{aligned}$$

where the first inequality follows by Assumption 3(i)-(ii); the third by Assumption 3(v) (which implies Lipschitz continuity of $\bar{\mu}(1;.)$ and $\mu(1;.;\bar{\theta}_N)$ uniformly over $\bar{\theta}_N \in \mathcal{N}$); and the final step follows by the results of Abadie and Imbens (2006, Lemma 2) on the bias of nearest neighbor matching. This proves $\left|Q_{31N}^{(1)}(\bar{\theta}_N) - Q_{32N}^{(1)}(\bar{\theta}_N)\right| \xrightarrow{p} 0$ under $\bar{P}_N$. I now argue $\left|Q_{31N}^{(1)}(\bar{\theta}_N) - Q_{32N}^{(1)}(\bar{\theta}_N)\right| \xrightarrow{p} 0$ under $\bar{P}_N$: Observe that $Q_{32N}^{(1)}(\bar{\theta}_N) = \bar{E}_N[Q_{31N}^{(1)}(\bar{\theta}_N)|\mathbf{W},\mathbf{X}]$ (the expectation being taken over $\mathbf{U}$, conditional on $\mathbf{W},\mathbf{X}$). But conditional on $\mathbf{W},\mathbf{X}$, the random variables $\{U_i : 1 \leq i \leq N\}$ are all independent of each other. Hence, by standard arguments involving the Markov inequality, together with Lemma 6 (i.e, $\bar{E}_N\left[\sup_{1\leq i\leq N} K_M^2(i;\bar{\theta}_N)\right] = o(N^{\delta})$) and Assumption 3, (which implies $\tau(X_i'\bar{\theta}_N) < \infty$ and $|e_{3i}(1;\bar{\theta}_N)| < \infty$ uniformly in i and N), it follows $\left|Q_{31N}^{(1)}(\bar{\theta}_N) - Q_{32N}^{(1)}(\bar{\theta}_N)\right| = o_{\bar{P}_N}(1)$.

It now remains to obtain the probability limit wrt $\bar{P}_N$ of $Q_{32N}^{(1)}(\bar{\theta}_N)$. Exploiting the definition of $\bar{K}_M(i;1,\bar{\theta}_N)$ and reordering the variables in the summation gives

$$\begin{aligned}Q_{32N}^{(1)}(\bar{\theta}_N) &= \frac{1}{N}\sum_{W_j=1}\tau(X_j'\bar{\theta}_N)\left(1 + \frac{K_M(j;\bar{\theta}_N)}{M}\right)e_{3j}\left(1;\bar{\theta}_N\right) \\ &\quad + \frac{1}{N}\sum_{W_j=1}\left(1 + \frac{K_M(j;\bar{\theta}_N)}{M}\right)\frac{1}{N_1(l(j))}\left\{\sum_{i\in S_0\left(l(j);\breve{\theta}_N\right)}\tau(X_i'\bar{\theta}_N)e_{3i}\left(1;\bar{\theta}_N\right)\right\} \\ &\equiv A_N^{(1)}(\bar{\theta}_N) + B_N^{(1)}(\bar{\theta}_N).\end{aligned}$$

Conditional on $\mathbf{W},\mathbf{X}'\bar{\theta}_N$, the summands within $A_N^{(1)}(\bar{\theta}_N)$ are mean zero and uncorrelated. Hence using Assumption 3 and Lemma 6, standard arguments assure $A_N^{(1)}(\bar{\theta}_N) = o_{\bar{P}_N}(1)$. Next, consider the term $B_N^{(1)}(\bar{\theta}_N)$: Suppose for simplicity that N/q_N is integer valued so that $N(l) = N/q_N$ for all l. I shall successively approximate $B_N^{(1)}(\bar{\theta}_N)$ by $B_{1N}^{(1)}(\bar{\theta}_N)$ and $B_{2N}^{(1)}(\bar{\theta}_N)$,[14] where

$$B_{1N}^{(1)}(\bar{\theta}_N) = \frac{1}{N}\sum_{W_j=1}\left(1 + \frac{K_M(j;\bar{\theta}_N)}{M}\right)\frac{1}{F(X_j'\bar{\theta}_N)}\left\{\frac{q_N}{N}\sum_{i\in S_0\left(l(j);\breve{\theta}_N\right)}\tau(X_i'\bar{\theta}_N)e_{3i}\left(1;\bar{\theta}_N\right)\right\};$$

[14]Note the difference in summation between the two terms.

and

$$B_{2N}^{(1)}(\bar{\theta}_N) = \frac{1}{N}\sum_{W_j=1}\left(1+\frac{K_M(j;\bar{\theta}_N)}{M}\right)\frac{1}{F(X_j'\bar{\theta}_N)}\left\{\frac{q_N}{N}\sum_{i\in S_0\left(l(j);\bar{\theta}_N\right)}\tau(X_i'\bar{\theta}_N)e_{3i}\left(1;\bar{\theta}_N\right)\right\}.$$

I first show that

$$\left|B_N^{(1)}(\bar{\theta}_N) - B_{1N}^{(1)}(\bar{\theta}_N)\right| \xrightarrow{p} 0. \tag{B.10}$$

Indeed a straightforward consequence of Lemma 9 and Assumption 3 (which implies $F^{-1}(.)$ is Lipschitz continuous and X_i is uniformly bounded) is that

$$\begin{aligned}&\sup_{1\le j\le N}\left|\frac{N(l(j))}{N_1(l(j))} - F^{-1}(X_j'\bar{\theta}_N)\right| \\ &\le \sup_{1\le j\le N}\left|\frac{N(l(j))}{N_1(l(j))} - F^{-1}\left(X_j'\breve{\theta}_N\right)\right| + \sup_{1\le j\le N}\left|F^{-1}(X_j'\bar{\theta}_N) - F^{-1}\left(X_j'\breve{\theta}_N\right)\right| = o_{\bar{P}_N}(1).\end{aligned}$$

Combining the above result with Lemma 6, and the fact $\tau(X_i'\bar{\theta}_N), |e_{3i}(1;\bar{\theta}_N)|$ are uniformly bounded, proves (B.10). Next, I show that

$$\left|B_{1N}^{(1)}(\bar{\theta}_N) - B_{2N}^{(1)}(\bar{\theta}_N)\right| = o_{\bar{P}_N}(1). \tag{B.11}$$

Let

$$\Delta(l;\bar{\theta}_N) \equiv S_0(l;\breve{\theta}_N) \triangle S_0(l;\bar{\theta}_N),$$

where $C \triangle D$ denotes the symmetric difference between any two sets C, D. Lemma 10 assures

$$\bar{P}_N\left(\max_{1\le l\le q_N}\#\Delta(l;\bar{\theta}_N) \ge N^{(1+\delta)/2}\right) \le q_N\exp(-N^\delta) \to 0 \tag{B.12}$$

for any $\delta > 0$ arbitrarily small (where $\#C$ denotes the cardinality of a set C). Combined with the boundedness property of $\tau(X_i'\bar{\theta}_N)$ and $|e_{3i}(1;\bar{\theta}_N)|$, (B.12) implies

$$\begin{aligned}\left|B_{1N}^{(1)}(\bar{\theta}_N) - B_{2N}^{(1)}(\bar{\theta}_N)\right| &= O_{\bar{P}_N}\left(\frac{q_N}{N^{(1-\delta)/2}}\right)\times\frac{1}{N}\sum_{W_j=1}\left(1+\frac{K_M(j;\bar{\theta}_N)}{M}\right)\frac{1}{F(X_j'\bar{\theta}_N)} \\ &= O_{\bar{P}_N}\left(\frac{q_N}{N^{(1-\delta)/2}}\right)\times O_{\bar{P}_N}(1) = o_{\bar{P}_N}(1),\end{aligned}$$

where the first equality follows from Lemma 6 and Assumption 3; and the final equality follows by Assumption 6. I have thus shown (B.11).

To complete the proof of the Lemma it remains to show

$$B_{2N}^{(1)}(\bar{\theta}_N) = o_{\bar{P}_N}(1). \tag{B.13}$$

Let $\rho_{N,i} = \tau(X_i'\bar{\theta}_N)e_{3i}(1;\bar{\theta}_N)$. For each l, the collection of random variables $\{\rho_{N,i} : i \in S_0(l;\bar{\theta}_N)\}$ are mean zero and uncorrelated conditional on $\mathbf{X}'\bar{\theta}_N$. Furthermore, wpa1-$\bar{P}_N$,

$$\#S_0(l(j);\bar{\theta}_N) \le \frac{N}{q_N} + \max_{1\le l\le q_N}\#\Delta(l;\bar{\theta}_N) \le \frac{N}{q_N} + N^{(1+\delta)/2}.$$

Hence for each $\epsilon > 0$, by the Markov inequality

$$\bar{P}_N\left(\max_{1\le l\le q_N}\left|\frac{q_N}{N}\sum_{i\in S_0(l;\bar{\theta}_N)}\rho_{N,i}\right|\ge\epsilon\right)\le\sum_{l=1}^{q_N}\bar{P}_N\left(\left|\frac{q_N}{N}\sum_{i\in S_0(l;\bar{\theta}_N)}\rho_{N,i}\right|\ge\epsilon\right)$$
$$=O\left(\frac{q_N^2}{N}+\frac{q_N^2}{N^{(3-\delta)/2}}\right)=o(1).$$

This, combined with the fact

$$\frac{1}{N}\sum_{W_j=1}\left(1+\frac{K_M(j;\bar{\theta}_N)}{M}\right)\frac{1}{F(X_j'\bar{\theta}_N)}=O_{\bar{P}_N}(1),$$

immediately proves (B.13). □

Lemma 4. *Under Assumptions 1-5 and $\bar{\theta}_N\to\theta_0$, it holds $V_2(h,\breve{\theta}_N)=t_1^2\sigma^2+o_{\bar{P}_N}(1)$.*

Proof. For the remainder of this proof I shall denote $p_{i,N}=F(X_i'\bar{\theta}_N)$. Additionally, for $a=3,4$, I set

$$\phi_{(a)i}(w;\theta)=(2w-1)\nu_{(a)i}(w;\theta).$$

First, note that $\bar{E}_N^*\left[\tilde{\varepsilon}_i^*(\bar{\theta}_N;\breve{\theta}_N)\right]=o_{\bar{P}_N}(1)$. Indeed this follows by a similar argument as in the proof of Lemma 3. Hence it suffices to show that $\bar{E}_N^*\left[\tilde{\varepsilon}_i^{*2}\left(\bar{\theta}_N;\breve{\theta}_N\right)\right]=\sigma^2+o_{\bar{P}_N}(1)$. To this end, I decompose

$$\tilde{\varepsilon}_i^*(\bar{\theta}_N;\breve{\theta}_N)=e_{1S_i^*}(\bar{\theta}_N)+\phi_{(3)S_i^*}(W_i^*;\bar{\theta}_N)+\phi_{(4)S_i^*}(W_i^*;\bar{\theta}_N),\tag{B.14}$$

and determine the probability limits of all the squared and cross product terms in (B.14), after taking the bootstrap expectation.

I begin with the probability limit of $\bar{E}_N^*\left[\phi_{(4)S_i^*}^2\left(W_i^*;\bar{\theta}_N\right)\right]$ under $\bar{P}_N$. Note that

$$\bar{E}_N^*\left[\phi_{(4)S_i^*}^2\left(W_i^*;\bar{\theta}_N\right)\right]=\frac{1}{N}\sum_{i=1}^N p_{i,N}\left(1+\frac{\tilde{K}_M(i;1,\bar{\theta}_N)}{M}\right)^2 e_{4\mathcal{J}_1(i)}^2\left(1;\bar{\theta}_N\right)$$
$$+\frac{1}{N}\sum_{i=1}^N(1-p_{i,N})\left(1+\frac{\tilde{K}_M(i;0,\bar{\theta}_N)}{M}\right)^2 e_{4\mathcal{J}_0(i)}^2\left(0;\bar{\theta}_N\right)$$
$$\equiv\Gamma_N^{(1)}+\Gamma_N^{(0)},$$

I shall characterize probability limit of $\Gamma_N^{(1)}$. That for $\Gamma_N^{(0)}$ follows by a similar argument.

Recall the definition $\bar{\sigma}^2(w,X)=E[Y^2|W=w,X]$. I shall first successively approximate $\Gamma_N^{(1)}$ by $\Gamma_{1N}^{(1)}$, $\Gamma_{2N}^{(1)}$, where

$$\Gamma_{1N}^{(1)}=\frac{1}{N}\sum_{i=1}^N p_{i,N}\left(1+\frac{\tilde{K}_M(i;1,\bar{\theta}_N)}{M}\right)^2\bar{\sigma}^2(1;X_{\mathcal{J}_w(i)});$$
$$\Gamma_{2N}^{(1)}=\frac{1}{N}\sum_{i=1}^N p_{i,N}\left(1+\frac{\tilde{K}_M(i;1,\bar{\theta}_N)}{M}\right)^2\bar{\sigma}^2(1;X_i).$$

In the first instance, I can expand

$$\Gamma_N^{(1)} - \Gamma_{1N}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\zeta_{N,i},$$

where

$$\zeta_{N,i} = p_{i,N}\bar{\Psi}_i(1,\bar{\theta}_N)\left\{e^2_{4\mathcal{J}_w(i)}(W_i,\bar{\theta}_N) - \bar{\sigma}^2(1;X_{\mathcal{J}_w(i)})\right\}.$$

Clearly $\bar{E}_N[\zeta_{N,i}|\mathbf{W},\mathbf{X},\mathbf{U}] = 0$ and $\bar{E}_N[\zeta_{N,i}\zeta_{N,j}|\mathbf{W},\mathbf{X},\mathbf{U}] = 0$ for all i,j such that $\mathcal{J}_1(i) \neq \mathcal{J}_1(j)$. Consequently by similar arguments[15] as in the proof of Lemma 3, it follows $\left|\Gamma_N^{(1)} - \Gamma_{1N}^{(1)}\right| = o_{\bar{P}_N}(1)$. Additionally, I can also show $\left|\Gamma_{1N}^{(1)} - \Gamma_{2N}^{(1)}\right| = o_{\bar{P}_N}(1)$ by similar arguments[16] as that used in the proof of Lemma 3 (note that by Assumption 3(v), $\bar{\sigma}^2(1;.)$ is Lipschitz continuous and uniformly bounded).

It now remains to obtain the probability limit wrt $\bar{P}_N$ of $\Gamma_{1N}^{(1)}$. By paralleling some of the steps[17] in the proof of Lemma 3, it follows

$$\left|\Gamma_{2N}^{(1)} - \Gamma_{3N}^{(1)}\right| = o_{\bar{P}_N}(1),$$

where

$$\Gamma_{3N}^{(1)} = \frac{1}{N}\sum_{W_j=1} p_{j,N}\left(1 + \frac{K_M(j;\bar{\theta}_N)}{M}\right)^2 \bar{\sigma}^2(1;X_j)$$
$$+ \frac{1}{N}\sum_{W_j=1}\left(1 + \frac{K_M(j;\bar{\theta}_N)}{M}\right)^2 \frac{1}{p_{j,N}}\left\{\frac{q_N}{N}\sum_{i\in S_0(l(j);\bar{\theta}_N)} p_{i,N}\bar{\sigma}^2(1;X_i)\right\}.$$

Define

$$\Gamma_{4N}^{(1)} = \frac{1}{N}\sum_{W_j=1}\left(1 + \frac{K_M(j;\bar{\theta}_N)}{M}\right)^2 m_1\left(p_{j,N};\bar{\theta}_N\right),$$

where

$$m_1\left(p;\bar{\theta}_N\right) = \bar{E}_N\left[\bar{\sigma}^2(1;X)|F(X_i'\bar{\theta}_N) = p\right].$$

I now show

$$\left|\Gamma_{3N}^{(1)} - \Gamma_{4N}^{(1)}\right| = o_{\bar{P}_N}(1). \tag{B.15}$$

To this end, I define an intermediate variable:

$$\Gamma_{31N}^{(1)} = \frac{1}{N}\sum_{W_j=1}\left(1 + \frac{K_M(j;\bar{\theta}_N)}{M}\right)^2 p_{j,N}\cdot m_1\left(p_{j,N};\bar{\theta}_N\right)$$
$$+ \frac{1}{N}\sum_{W_j=1}\left(1 + \frac{K_M(j;\bar{\theta}_N)}{M}\right)^2 \frac{1-p_{j,N}}{p_{j,N}}\left\{\frac{1}{N_0\left(l(j);\bar{\theta}_N\right)}\sum_{i\in S_0(l(j);\bar{\theta}_N)} p_{i,N}\cdot m_1\left(p_{i,N};\bar{\theta}_N\right)\right\}.$$

[15] Specifically, the ones used to prove $Q_{4N}^{(1)}(\bar{\theta}_N) \xrightarrow{p} 0$.

[16] Specifically, the ones used to prove $\left|Q_{3N}^{(1)}(\bar{\theta}_N) - Q_{31N}^{(1)}(\bar{\theta}_N)\right| = o_{\bar{P}_N}(1)$.

[17] Precisely, the steps leading to $\left|Q_{31N}^{(1)}(\bar{\theta}_N) - Q_{32N}^{(1)}(\bar{\theta}_N)\right| = o_{\bar{P}_N}(1)$, followed by reordering of the terms in $Q_{32N}^{(1)}(\bar{\theta}_N)$, and finally succesive approximations of $B_N^{(1)}(\bar{\theta}_N)$ with $B_{1N}^{(1)}(\bar{\theta}_N)$ and $B_{2N}^{(1)}(\bar{\theta}_N)$.

By Lemmas 9 and 10, and Assumption 6, there exists some $c > 0$ for which it holds

$$\min_{1\leq l\leq q_N} N_0\left(l(j);\bar{\theta}_N\right) \geq \min_{1\leq l\leq q_N} N_0(l) - N^{(1+\delta)/2} \geq cN/q_N, \tag{B.16}$$

with probability approaching one under $\bar{P}_N$. The same lemmas together with Assumptions 3,6 also assure

$$\begin{aligned}\sup_{1\leq j\leq N}\left|\frac{q_N N_0\left(l(j);\bar{\theta}_N\right)}{N} - (1-p_{j,N})\right| &\leq \sup_{1\leq j\leq N}\left|\frac{q_N N_0\left(l(j);\bar{\theta}_N\right)}{N} - \left(1-F(X_j'\breve{\theta})\right)\right| + o_{\bar{P}_N}(N^{-\frac{1}{2}}) \\ &\leq \sup_{1\leq j\leq N}\left|\frac{q_N N_0\left(l(j);\breve{\theta}_N\right)}{N} - \left(1-F(X_j'\breve{\theta})\right)\right| + o_{\bar{P}_N}\left(\frac{q_N}{N^{(1-\delta)/2}} + N^{-\frac{1}{2}}\right) = o_{\bar{P}_N}(1).\end{aligned}$$

Additionally, by the usual arguments based on the Markov inequality, and employing (B.16) together with Assumption 6, it follows

$$\bar{P}_N\left(\max_{1\leq l\leq q_N}\left|\frac{1}{N_0\left(l;\bar{\theta}_N\right)}\sum_{i\in S_0(l;\bar{\theta}_N)} p_{i,N}\bar{\sigma}^2(1;X_i) - \frac{1}{N_0\left(l;\bar{\theta}_N\right)}\sum_{i\in S_0(l;\bar{\theta}_N)} p_{i,N}m_1\left(p_{i,N};\bar{\theta}_N\right)\right| \geq \epsilon\right) \to 0.$$

Combining the above results with the fact

$$\frac{1}{N}\sum_{W_j=1}\left(1+\frac{K_M(j;\bar{\theta}_N)}{M}\right)^2\frac{1}{p_{j,N}} = O_{\bar{P}_N}(1),$$

proves that $\left|\Gamma_{3N}^{(1)} - \Gamma_{31N}^{(1)}\right| = o_{\bar{P}_N}(1)$. Now, define $w_1\left(p;\bar{\theta}_N\right) \equiv p\cdot m_1(p;\bar{\theta}_N)$. I can bound

$$\begin{aligned}&\left|\Gamma_{31N}^{(1)} - \Gamma_{4N}^{(1)}\right| \\ &\leq \left\{\frac{1}{N}\sum_{W_j=1}\frac{1-p_{j,N}}{p_{j,N}}\left(1+\frac{K_M(j;\bar{\theta}_N)}{M}\right)^2\right\}\max_{1\leq l\leq q_N}\max_{i,j\in S_0(l;\bar{\theta}_N)}\left|w_1\left(p_{i,N};\bar{\theta}_N\right) - w_1\left(p_{j,N};\bar{\theta}_N\right)\right| \\ &= O_{\bar{P}_N}(1)\cdot\max_{1\leq l\leq q_N}\max_{i,j\in S_0(l;\bar{\theta}_N)}\left|w_1\left(p_{i,N};\bar{\theta}_N\right) - w_1\left(p_{j,N};\bar{\theta}_N\right)\right| \\ &\leq O_{\bar{P}_N}(1)\cdot\max_{1\leq l\leq q_N}\max_{i,j\in S_0(l;\bar{\theta}_N)}|p_{i,N}-p_{j,N}| \\ &\leq O_{\bar{P}_N}(1)\cdot\max_{1\leq l\leq q_N}\left|\pi_{l-1}(\breve{\theta}_N) - \pi_l(\breve{\theta}_N)\right| = o_{\bar{P}_N}(1),\end{aligned}$$

where the first equality follows by Assumption 3(i)-(iii) together with Lemma 6; the second inequality follows by the uniform Lipschitz continuity of $m_1\left(.;\bar{\theta}_N\right)$ (Assumption 3(v)); the third inequality follows by the definition of $S_0(l;\bar{\theta}_N)$; and the final equality follows by Lemma 8. I have thus shown (B.15).

Now, the probability limit of $\Gamma_{4N}^{(1)}$ under $\bar{P}_N$ can be obtained by the techniques of Abadie and Imbens (2016) (See also Lemmas (14)-(16) in appendix D). The probability limit of $\Gamma_{4N}^{(0)}$ under $\bar{P}_N$ is obtained analogously. Combining the expressions gives the probability limit of $\bar{E}_N^*\left[\phi^2_{(4)S_i^*}\left(W_i^*;\bar{\theta}_N\right)\right]$, which is equivalent to that obtained in Abadie and Imbens (2016).

Next consider the term $\bar{E}_N^*\left[\phi^2_{(3)S_i^*}\left(W_i^*;\bar{\theta}_N\right)\right]$. As before I can decompose

$$\begin{aligned}\bar{E}_N^*\left[\phi^2_{(3)S_i^*}\left(W_i^*;\bar{\theta}_N\right)\right] &= \frac{1}{N}\sum_{i=1}^N p_{i,N}\left(1+\frac{\tilde{K}_M(i;1,\bar{\theta}_N)}{M}\right)^2 e^2_{3\mathcal{J}_1(i)}\left(1;\bar{\theta}_N\right)\\ &\quad+\frac{1}{N}\sum_{i=1}^N (1-p_{i,N})\left(1+\frac{\tilde{K}_M(i;0,\bar{\theta}_N)}{M}\right)^2 e^2_{3\mathcal{J}_0(i)}\left(0;\bar{\theta}_N\right)\\ &\equiv \Delta_N^{(1)}+\Delta_N^{(0)}.\end{aligned}$$

Consider the term $\Delta_N^{(1)}$: By similar arguments as in Lemma 3,

$$\max_{1\le i\le N}\left|e_{3\mathcal{J}_1(i)}(1;\bar{\theta}_N)-e_{3i}(1;\bar{\theta}_N)\right| = O_{\bar{P}_N}(N^{-1/k}).$$

Together with Lemma 6, the above assures $\left|\Delta_N^{(1)}-\Delta_{1N}^{(1)}\right| = o_{\bar{P}_N}(1)$, where

$$\Delta_{1N}^{(1)} = \frac{1}{N}\sum_{i=1}^N p_{i,N}\left(1+\frac{\tilde{K}_M(i;1,\bar{\theta}_N)}{M}\right)^2 e^2_{3i}\left(1;\bar{\theta}_N\right).$$

Now the probability limit of $\Delta_{1N}^{(1)}$ can be analyzed the same arguments as that employed for $\Gamma_{1N}^{(1)}$. Doing so gives the probability limit for $\bar{E}_N^*\left[\phi^2_{(3)S_i^*}\left(W_i^*;\bar{\theta}_N\right)\right]$ under $\bar{P}_N$, which is again equivalent to the corresponding expression in Abadie and Imbens (2016).

Finally, it is straightforward to obtain the probability limit of $\bar{E}_N^*[e^2_{1S_i^*}(\bar{\theta}_N)]$ under $\bar{P}_N$ using standard methods. Taken together I can show

$$\bar{E}_N^*[e^2_{1S_i^*}(\bar{\theta}_N)]+\bar{E}_N^*\left[\phi^2_{(4)S_i^*}\left(W_i^*;\bar{\theta}_N\right)\right]+\bar{E}_N^*\left[\phi^2_{(3)S_i^*}\left(W_i^*;\bar{\theta}_N\right)\right] = \sigma^2+o_{\bar{P}_N}(1).$$

It only remains to verify that the bootstrap expectation of the cross product terms in (B.14) converge in probability to 0 under $\bar{P}_N$. Consider, for instance,

$$\Phi_N \equiv \bar{E}_N^*\left[\phi_{(3)i}(W_i^*;\bar{\theta}_N)\cdot\phi_{(4)i}(W_i^*;\bar{\theta}_N)\right].$$

Taking the bootstrap expectations, I observe $\Phi_N = \Phi_N^{(1)}+\Phi_N^{(0)}$, where

$$\Phi_N^{(1)} = \frac{1}{N}\sum_{i=1}^N p_{i,N}\left(1+\frac{\tilde{K}_M(i;1,\bar{\theta}_N)}{M}\right)^2 e_{3\mathcal{J}_1(i)}(1;\bar{\theta}_N)e_{4\mathcal{J}_1(i)}(1;\bar{\theta}_N),$$

and a similar expression holds for $\Phi_N^{(0)}$. Denoting

$$\varrho_{N,i} = p_{i,N}\left(1+\frac{\tilde{K}_M(i;1,\bar{\theta}_N)}{M}\right)^2 e_{3\mathcal{J}_1(i)}(1;\bar{\theta}_N)e_{4\mathcal{J}_1(i)}(1;\bar{\theta}_N),$$

I note that $\bar{E}_N[\varrho_{N,i}|\mathbf{W},\mathbf{X},\mathbf{U}]=0$ and $\bar{E}_N[\varrho_{N,i}\varrho_{N,j}|\mathbf{W},\mathbf{X},\mathbf{U}]=0$ for all i,j such that $\mathcal{J}_1(i)\neq\mathcal{J}_1(j)$. Consequently, by similar arguments as in the proof of Lemma 3, it follows $\Phi_N^{(1)}\xrightarrow{p}0$ under $\bar{P}_N$. By symmetry, I also have $\Phi_N^{(0)}\xrightarrow{p}0$ under $\bar{P}_N$, implying $\Phi_N = o_{\bar{P}_N}(1)$. Some more algebra on the usual lines shows that the remaining cross product terms also converge in probability to 0 under $\bar{P}_N$. This completes the proof of the lemma. □

Lemma 5. *Under Assumptions 1-5 and $\bar{\theta}_N\to\theta_0$, it holds $V_3(h,\breve{\theta}_N) = 2t_1c't_2+o_{\bar{P}_N}(1)$.*

Proof. For the course of this proof set $h(.;\bar{\theta}_N)$ as $h_N(.)$. Furthermore, to simplify the algebra I again employ the notation (first introduced in the proof of Lemma 4)

$$\phi_{(a)i}(w;\theta) = (2w-1)\nu_{(a)i}(w;\theta).$$

By the construction of the bootstrap DGP, it follows $V_3(h,\breve{\theta}_N) = \bar{E}_N^*\left[\left(t_1\varepsilon_i^*(\bar{\theta}_N;\breve{\theta}_N)\right)\left(t_2'\beta_i^*(\bar{\theta}_N)\right)\right]$ since $\bar{E}_N^*[\beta_i^*(\bar{\theta}_N)] = 0$. I then decompose the term $\varepsilon_i^*(\bar{\theta}_N;\breve{\theta}_N)$ as in equation (B.14) and determine the probability limits of the bootstrap expectations of the resulting terms.

First, taking the bootstrap expectations it can be verified $\bar{E}_N^*\left[e_{1S_i^*}(\bar{\theta}_N)\cdot t_2'\beta_i^*(\bar{\theta}_N)\right] = 0$. At the end of the proof I show that

$$\bar{E}_N^*\left[\phi_{(4)i}(W_i^*;\bar{\theta}_N)\cdot t_2'\beta_i^*(\bar{\theta}_N)\right] = o_{\bar{P}_N}(1). \tag{B.17}$$

Hence it suffices for the claim to prove

$$\bar{E}_N^*\left[\phi_{(3)i}(W_i^*;\bar{\theta}_N)\cdot t_2'\beta_i^*(\bar{\theta}_N)\right] = t_2'c + o_{\bar{P}_N}(1).$$

Taking the bootstrap expectations, I obtain

$$\bar{E}_N^*\left[\phi_{(3)i}(W_i^*;\bar{\theta}_N)\cdot t_2'\beta_i^*(\bar{\theta}_N)\right] = T_N^{(1)} + T_N^{(0)},$$

where for $w = 0, 1$,

$$T_N^{(w)} = \frac{1}{N}\sum_{i=1}^N f\left(X_i'\bar{\theta}_N\right) t_2'\left\{X_i - h_N\left(X_i'\bar{\theta}_N\right)\right\}\left(1 + \frac{\tilde{K}_M(i;w,\theta)}{M}\right) e_{3\mathcal{J}_w(i)}(w;\theta).$$

Let me now denote

$$T_{1N}^{(w)} = \frac{1}{N}\sum_{i=1}^N f\left(X_i'\bar{\theta}_N\right) t_2'\left\{X_i - h_N\left(X_i'\bar{\theta}_N\right)\right\}\left(1 + \frac{\tilde{K}_M(i;w,\bar{\theta}_N)}{M}\right) e_{3i}(w;\bar{\theta}_N).$$

Using the properties of nearest neighbor matching, I can employ similar arguments as in the proof of Lemma (3) to show that for $w = 0, 1$,

$$\left|T_N^{(w)} - T_{1N}^{(w)}\right| = o_{\bar{P}_N}(1).$$

Thus the probability limit under $\bar{P}_N$ of $\bar{E}_N^*\left[\phi_{(3)i}(W_i^*;\bar{\theta}_N)\cdot t_2'\beta_i^*(\bar{\theta}_N)\right]$ is equivalent to that of $T_{1N}^{(1)} + T_{1N}^{(0)}$. The latter in turn can be obtained by following similar arguments as in the proof of Lemma 4. Hence, after some algebra I obtain $T_{1N}^{(1)} + T_{1N}^{(0)} = t_2'c + o_{\bar{P}_N}(1)$.

It only remains now to show (B.17). Taking the bootstrap expectation gives

$$\bar{E}_N^*\left[\phi_{(4)i}(W_i^*;\bar{\theta}_N)\cdot t_2'\beta_i^*(\bar{\theta}_N)\right] = V_N^{(1)} + V_N^{(0)},$$

where for $w = 0, 1$,

$$V_N^{(w)} = \frac{1}{N}\sum_{i=1}^N f\left(X_i'\bar{\theta}_N\right) t_2'\left\{X_i - h_N\left(X_i'\bar{\theta}_N\right)\right\}\nu_{(4)i}(w;\bar{\theta}_N) \equiv \frac{1}{N}\sum_{i=1}^N \sigma_{N,i}.$$

By law of iterated expectations $\bar{E}_N[\sigma_{N,i}|\mathbf{W},\mathbf{X},\mathbf{U}] = 0$ and $\bar{E}_N[\sigma_{N,i}\sigma_{N,j}|\mathbf{W},\mathbf{X},\mathbf{U}] = 0$ for all i, j such that $\mathcal{J}_w(i) \neq \mathcal{J}_w(j)$. Consequently by similar arguments as in the proof of Lemma 3, it follows $V_N^{(w)} = o_{\bar{P}_N}(1)$ for $w = 0, 1$. This concludes the proof of the lemma. □

Appendix C. Additional Lemmas

I use the same notation as in Appendices A and B.

Lemma 6. *Suppose that Assumptions 1-3 hold . Then for any $q < \infty$ and δ arbitrarily small, it holds that uniformly in N,*

$$\sup_{\theta \in \mathcal{N}} \bar{E}_\theta \left[|K_M(i;\theta)|^q \right] < \infty,$$

and

$$\sup_{\theta \in \mathcal{N}} \bar{E}_\theta \left[\sup_{1 \leq i \leq N} |K_M(i;\theta)|^q \right] = o(N^\delta).$$

Proof. The first claim follows by similar arguments as in Abadie and Imbens (2016, Lemma S.8), after employing Lemma 11 (in particular the second statement) and Lemma 12. The second claim follows by paralleling the arguments of Abadie and Imbens (2006, Additional proofs p.23). □

Let $\mathcal{N}$ denote some neighborhood of θ_0 such that Assumptions 1-5 hold for each $\theta \in \mathcal{N}$. Additionally let $G_{w,\theta}(.)$ denote the CDF of the sample propensity score $F(X'\theta)$ conditional on $W = w$; and $g_{w,\theta}(.)$ the corresponding density function (where it exists). At the same time $G_\theta(.)$ denotes the unconditional CDF of the propensity score $F(X'\theta)$, $Q_\theta(.) \equiv G_\theta^{-1}(.)$ its corresponding quantile function, and $g_\theta(.)$ its density function. The empirical CDF of $F(X'\theta)$ is denoted as

$$\hat{G}_\theta(t) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\{F(X_i'\theta) \leq t),$$

and the corresponding empirical quantile function as

$$\hat{Q}_\theta(p) = \inf\{t : \hat{G}_\theta(t) \geq p\}.$$

Note that by construction $\hat{G}_\theta(\hat{Q}_\theta(p)) = p$ for any $p \in (0,1)$. To simplify notation I shall employ the convention $G_{w,N}(.) \equiv G_{w,\breve{\theta}_N}(.)$, $G_N(.) \equiv G_{\breve{\theta}_N}(.)$, $G_w(.) \equiv G_{w,\theta_0}(.)$ and $G(.) \equiv G_{\theta_0}(.)$ The other terms $g_{w,N}(.), g_w(.), g_N(.), g(.)$ and $\hat{G}_N(.), \hat{Q}_N(.)$ for $w = 0, 1$ are defined analogously. As in Appendix B, in what follows I suppress indexing the quantities with the additional label $\breve{\theta}_N$. However it should be implicitly understood that I have replaced $\bar{\hat{\theta}}$ with $\breve{\theta}_N$.

Lemma 7. *Suppose that Assumptions 3 hold. Then for any sequence $\breve{\theta}_N$ such that $\breve{\theta}_N \to \theta_0$,*

$$\sup_{p \in (0,1)} \left| \hat{Q}_N(p) - Q(p) \right| = o_{\bar{P}_N}(1).$$

Proof. I first show that

$$\sup_{t \in [0,1]} \left| \hat{G}_\theta(t) - G(t) \right| = o_{\bar{P}_N}(1). \tag{C.1}$$

Consider the class of functions $\mathcal{G} \equiv \{x'\theta; \theta \in \mathcal{N}\}$ (here x denotes the functional argument). Observe that $\mathcal{G}$ is finite dimensional, being a subset of the space of all linear combinations of $e_1(x), \ldots, e_k(x)$: the (linear) functions corresponding to each axis in the Euclidean $\mathbb{R}^k$ space. By the results of Pollard (2012), this implies that the class of all sets of the form $\{x : x'\theta \leq t\}$ for $\theta \in \mathcal{N}$ and $t \in \mathbb{R}$ is a VC class; equivalently, so is the class of sets $\{x : F(x'\theta) \leq t\}$ a VC

class for $\theta \in \mathcal{N}$ and $t \in \mathbb{R}$, since $F(.)$ is strictly monotone. Hence, by the uniform law of large numbers for VC class sets (see Pollard, 2012; also Vapnik and Chervonenkis, 1971), I obtain

$$\sup_{\theta \in \mathcal{N};\ t \in [0,1]} \left| \hat{G}_\theta(t) - G_\theta(t) \right| = o_{\bar{P}_N}(1).$$

By the fact $\breve{\theta}_N \to \theta_0$, together with Assumption 3(i)-(ii),

$$\sup_{t \in [0,1]} |G_N(t) - G(t)| \to 0.$$

Combining the above immediately proves (C.1).

Using (C.1), and recalling that $\hat{G}_N\left(\hat{Q}_N(q)\right) = q$, I have

$$\sup_{q \in (0,1)} \left| q - G\left(\hat{Q}_N(q)\right) \right| = \sup_{q \in (0,1)} \left| \hat{G}_N\left(\hat{Q}_N(q)\right) - G\left(\hat{Q}_N(q)\right) \right| \to 0.$$

Now $Q(.) \equiv G^{-1}(.)$ is uniformly continuous on $(0,1)$ by virtue of the fact - implied by Assumption 3(iii) - that $G(.)$ is strictly increasing and continuous on its interval valued support. Hence it follows from the previous display equation that

$$\sup_{q \in (0,1)} \left| Q(q) - \hat{Q}_N(q) \right| = o_{\bar{P}_N}(1),$$

as claimed in the Lemma. □

Lemma 8. *Suppose that Assumptions 3,7 hold. Then for any sequence $\breve{\theta}_N$ such that $\breve{\theta}_N \to \theta_0$,*

$$\max_{1 \le l \le q_N} \left| \pi_{l-1}(\breve{\theta}_N) - \pi_l(\breve{\theta}_N) \right| = o_{\bar{P}_N}(1).$$

Proof. Note that $\pi_1(\breve{\theta}_N), \ldots, \pi_{q_N}(\breve{\theta}_N)$ are obtained by evaluating the quantile function $\hat{Q}_N(.)$ at the values $\{1/q_N, 2/q_N, \ldots, q_N - 1/q_N\}$. The claim is thus a straightforward consequence of the previous lemma together with uniform continuity of $Q(.)$ and $q_N \to \infty$. □

Lemma 9. *Suppose that Assumptions 3,6 hold. Then for any sequence $\breve{\theta}_N$ such that $\breve{\theta}_N \to \theta_0$, there exists some universal constant $c > 0$ for which*

$$\bar{P}_N\left(\min_{1 \le l \le q_N} N_w(l) \ge c \frac{N}{q_N} \right) \ge 1 - o\left(\frac{q_N^2}{N} \right)$$

for $w = 0, 1$. Furthermore,

$$\max_{1 \le j \le N} \left| \frac{N_1(l(j))}{N(l(j))} - F(X_j' \bar{\theta}_N) \right| = o_{\bar{P}_N}(1),$$

and

$$\max_{1 \le j \le N} \left| \frac{N(l(j))}{N_1(l(j))} - \frac{1}{F(X_j' \bar{\theta}_N)} \right| = o_{\bar{P}_N}(1).$$

Proof. Assume for simplicity that N/q_N is an integer. Then $N(l) = N/q_N$ for all l. Now,

$$\begin{aligned}
&\bar{P}_N\left(\min_{1\le l\le q_N} N_w(l) \ge c\frac{N}{q_N}\right) = \bar{P}_N\left(N_w(l) \ge c\frac{N}{q_N} \text{ for } l = 1,\dots q_N\right) \\
&= \prod_{l=1}^{q_N} \bar{P}_N\left(N_w(l) \ge c\frac{N}{q_N}\right) = \prod_{l=1}^{q_N} \bar{P}_N\left(\frac{q_N}{N}\sum_{i\in S_w(l)} W_i \ge c\right) \\
&\ge \left(1 - \underline{\eta}\frac{q_N}{N}\right)^{q_N} = 1 - o\left(\frac{q_N^2}{N}\right),
\end{aligned}$$

where the second equality follows by the iid property of the observations; and the inequality is based on an application of the Markov inequality after noting $\bar{E}_N[W_i] = F(X_i'\bar{\theta}_N)$ with $\min_{1\le i\le N} F(X_i'\bar{\theta}_N) \ge \underline{\eta}$ for some $\underline{\eta} > 0$ by Assumption 3(i). This proves the first claim of the lemma.

For each l, let $\dot{p}_{l,N} \equiv \bar{E}_N[q_N N_1(l)/N]$. Since both $\dot{p}_{l(j),N}$ and $F(X_j'\bar{\theta}_N)$ lie within $\left[\pi_{l-1}(\breve{\theta}_N) - \pi_l(\breve{\theta}_N)\right]$ for some l, by Lemma 8 it suffices for the second claim to show that

$$\max_{1\le l\le q_N}\left|\frac{q_N N_1(l)}{N} - \dot{p}_{l,N}\right| = o_{\bar{P}_N}(1). \tag{C.2}$$

Fix some $\epsilon > 0$. By the Markov inequality, for each $1 \le l \le q_N$,

$$\bar{P}_N\left(\left|\frac{q_N N_1(l)}{N} - \dot{p}_{l,N}\right| > \epsilon\right) \le \frac{q_N}{N\epsilon}.$$

Hence, by Assumption 6 it follows

$$\bar{P}_N\left(\max_{1\le l\le q_N}\left|\frac{q_N N_1(l)}{N} - \dot{p}_{l,N}\right| > \epsilon\right) \le \frac{q_N^2}{N\epsilon} \to 0.$$

This proves (C.2), which completes the proof of the second claim of the lemma. The third claim follows immediately from the second, since by the previous arguments in this proof the events

$$\min_{1\le l\le q_N}\frac{N(l)}{N_1(l)} \ge c > 0; \quad \text{and} \quad \min_{1\le j\le N} F(X_j'\bar{\theta}_N) \ge \underline{\eta} > 0$$

occur with probability greater than or equal to $1 - o\left(q_N^2/N\right)$ under $\bar{P}_N$. □

For $w = 0, 1$ let $\Delta_w(l;\bar{\theta}_N) \equiv S_w(l;\breve{\theta}_N) \triangle S_w(l;\bar{\theta}_N)$. Also for any set A, let $\#A$ denote the cardinality of that set.

Lemma 10. *Suppose that Assumptions 3,7 hold. Then for any sequence $\breve{\theta}_N$ such that $\breve{\theta}_N \to \theta_0$, it holds, for $w = 0, 1$ and some $\delta > 0$ arbitrarily small, that*

$$\bar{P}_N\left(\max_{1\le l\le q_N} \#\Delta_w(l;\bar{\theta}_N) \ge N^{(1+\delta)/2}\right) \le q_N \exp(-N^\delta).$$

Proof. Without loss of generality I consider the case when $w = 1$. Define

$$\delta_N = \max_{1\le i\le N}\left|F(X_i'\bar{\theta}_N) - F(X_i'\breve{\theta}_N)\right|.$$

By Assumption 3(i)-(iii), $\delta_N \leq C/\sqrt{N}$ for some $C < \infty$. Also, let $\mathcal{C}_{l,N}$ denote the set

$$\mathcal{C}_{l,N} \equiv \Big\{ i : \pi_{l-1}(\breve{\theta}_N) - \delta_N \leq F(X_i'\breve{\theta}_N) \leq \pi_{l-1}(\breve{\theta}_N) + \delta_N \\ \cup\, \pi_l(\breve{\theta}_N) - \delta_N \leq F(X_i'\breve{\theta}_N) \leq \pi_l(\breve{\theta}_N) + \delta_N \Big\}.$$

Clearly $\#\Delta_w(l;\bar{\theta}_N) \leq \#\mathcal{C}_{l,N}$. Represent by $\varpi_{i,l,N}$ the random variable $\mathbb{I}\{i \in \mathcal{C}_{l,N}\}$. By the bound on δ_N and the fact $g_{1,N} \leq C_2 < \infty$ uniformly in N (in turn due to Assumption 3(iii), see Lemma 11), it follows $\bar{E}_N[\varpi_{i,l,N}] \leq C_3/\sqrt{N}$ for some $C_3 < \infty$ independent of l, N. Hence for each l, and some sequence $M_N \asymp N^{\delta}$ independent of l, I obtain

$$\begin{aligned} &\bar{P}_N\left(\#\Delta_w(l;\bar{\theta}_N) \geq N^{(1+\delta)/2}\right) \leq \bar{P}_N\left(\#\mathcal{C}_{l,N} \geq N^{(1+\delta)/2}\right) \\ &= \bar{P}_N\left(\frac{1}{N}\sum_{i=1}^{N} \varpi_{i,l,N} \geq \sqrt{N^{\delta-1}}\right) \\ &\leq \bar{P}_N\left(\left|\frac{1}{N}\sum_{i=1}^{N} \varpi_{i,l,N} - \bar{E}_N\left[\varpi_{i,l,N}\right]\right| \geq \sqrt{\frac{M_N}{N}}\right) \leq \exp(-M_N), \end{aligned}$$

where the final step follows by Hoeffding's inequality. But

$$\bar{P}_N\left(\max_{1\leq l\leq q_N} \#\Delta_w(l;\bar{\theta}_N) \geq N^{(1+\delta)/2}\right) \leq \sum_{l=1}^{q_N} \bar{P}_N\left(\#\Delta_w(l;\bar{\theta}_N) \geq N^{(1+\delta)/2}\right);$$

hence the claim follows immediately through the above arguments. □

Appendix D. Uniform statements of the results in Abadie and Imbens (2016)

The lemmas in this appendix are based on Abadie and Imbens (2016), which are extended to apply uniformly over all θ in a neighborhood of θ_0. I thus modify the proofs of Abadie and Imbens (2016) accordingly.

I use the same notation as in Appendices A,B and C. In addition, I employ the following: Let $p_i(\theta)$ denote $p(X;\theta) \equiv F(X'\theta)$ and $p_{i,N} = p(X_i;\bar{\theta}_N)$. Also let $q_{0,\theta} = E_0[1 - F(X_i'\theta)]$ and $q_{1,\theta} = E_0[F(X_i'\theta)]$ denote the unconditional probabilities that $W_i = 0$ and $W_i = 1$ respectively when the propensity score is $F(X'\theta)$. To simplify notation I shall employ the convention $q_{w,N}(.) \equiv q_{w,\bar{\theta}_N}$ and $q_w \equiv q_{w,\theta_0}$ for $w = 0, 1$. Finally for $w = 0, 1$, let $N_{w,\theta} = \left\{\sum_{i=1}^N \mathbb{I}_{W_i=w};\ W_i \sim \text{Bernoulli}(F(X_i'\theta)\right\}$. Per convention, let $N_w \equiv N_{w,\theta_0}$ and $N_{w,N} \equiv N_{w,\bar{\theta}_N}$.

Lemma 11. *Suppose that Assumptions 3 hold. Then: (i) the support of $g_\theta(.)$ and $g(.)$ lies within the interval $[\underline{p}, \bar{p}]$ for some $0 < \underline{p} < \bar{p} < 1$; (ii) there exist universal constants $\underline{c}$ and $\bar{C}$ such that $\underline{c} < \sup_{\theta\in\mathcal{N}} (g_{1,\theta}(p)/g_{0,\theta}(p)) < \bar{C}$ uniformly over all p such that $g_\theta(p) \neq 0$; and (iii) there exist universal constants $1 > \bar{\eta} \geq \underline{\eta} > 0$ such that $q_{w,\theta} \in [\underline{\eta}, \bar{\eta}]$ uniformly in $\theta \in \mathcal{N}$.*

Proof. That the support of $g_\theta(.)$ and $g(.)$ is within some interval $[\underline{p}, \bar{p}]$ follows from the bounded support assumption for X (Assumption 3(i)), and the fact $f(.)$ is strictly positive and bounded (Assumption 3(ii)). Additionally, the support condition on X together with Assumption 3(ii) also ensures existence of universal constants $1 > \bar{\eta} \geq \underline{\eta} > 0$ such that $q_{w,\theta} \in [\underline{\eta}, \bar{\eta}]$ uniformly in $\theta \in \mathcal{N}$. By the Bayes theorem, $g_{0,\theta}(p) = (1-p)g_\theta(p)/q_{0,\theta}$ and $g_{1,\theta}(p) = pg_\theta(p)/q_{1,\theta}$. This proves the existence of $g_{w,\theta}(.)$ for $w = 0, 1$. Given the support condition for $g_\theta(.)$ proved already, the claim $\underline{c} < \sup_{\theta\in\mathcal{N}} (g_{1,\theta}(p)/g_{0,\theta}(p)) < \bar{C}$ follows by similar arguments as in the proof of Abadie and Imbens (2016, Lemma S.2). □

Lemma 12. *Suppose that for $w = 0, 1$, $N_{w,\theta}$ are truncated for values smaller than M and greater than $N - M$ where $N > 2M$. Then for any $q < \infty$ and $w = 0, 1$ there exists $M_q < \infty$ such that,*

$$\sup_{\theta\in\mathcal{N}} E_\theta \left[\left|\frac{N}{N_{w,\theta}}\right|^q\right] \leq M_q.$$

Proof. Observe that $N_{w,\theta}$ is a binomial variable with parameters $(N, q_{w,\theta})$ where $q_{w,\theta} \in [\underline{\eta}, \bar{\eta}]$ uniformly in $\theta \in \mathcal{N}$ by Lemma 11. Hence the claim follows by similar arguments as in the proof of Abadie and Imbens (2016, Lemma S.3). □

Let $\xi_{1:N_w}, \ldots, \xi_{N_w:N_w}$ denote the order statistics for a set of N_w random variables drawn from the uniform distribution. Denote the interval support of $F(X'\bar{\theta}_N)$ by $[a_N, b_N]$.

Lemma 13. *Suppose that Assumptions 1-4 hold. Then for any sequence $\{\bar{\theta}_N\}$ satisfying $\bar{\theta}_N \to \theta_0$ it holds that under $\bar{P}_N$*

$$\max_{i=1,\ldots,N} \left|G_{w,N}^{-1}(\xi_{i:N_w}) - G_{w,N}^{-1}(i/N_w)\right| = o_p(1). \tag{D.1}$$

Proof. I first prove (D.1). By the fact $\bar{\theta}_N \to \theta_0$ and Assumptions 3(i),(ii), it follows that $G_{w,N}(.)$ is compactly supported for all N sufficiently large. Furthermore, under the same assumptions,

it follows

$$\sup_{p\in\mathbb{R}} |G_{w,N}(p) - G_w(p)| \to 0. \tag{D.2}$$

By Assumption 3(iii), $G^{-1}_{w,N}(.)$ exists for N sufficiently large (since $g_N(.)$ and consequently $g_{w,N}(.)$ are strictly positive within an interval support for $F(X'\theta_N)$)[18]. Then

$$\sup_{q\in(0,1)} \left|G_w\left(G^{-1}_{w,N}(q)\right) - q\right| = \sup_{q\in(0,1)} \left|G_w\left(G^{-1}_{w,N}(q)\right) - G_{w,N}\left(G^{-1}_{w,N}(q)\right)\right| \to 0.$$

Now $G^{-1}_w(.)$ is uniformly continuous on $[0,1]$ by virtue of the fact $G_w(.)$ is strictly increasing and, therefore, continuous on a compact set. Hence, it follows from the above that

$$\sup_{q\in(0,1)} \left|G^{-1}_{w,N}(q) - G^{-1}_w(q)\right| \to 0. \tag{D.3}$$

I thus obtain

$$\max_{i=1,\ldots,N} \left|G^{-1}_{w,N}(\xi_{i:N_w}) - G^{-1}_{w,N}(i/N_w)\right| = \max_{i=1,\ldots,N_w} \left|G^{-1}_w(\xi_{i:N_w}) - G^{-1}_w(i/N_w)\right| + o(1) = o_{\bar{P}_N}(1),$$

where the second equality follows by similar arguments as in Abadie and Imbens (2016, Lemma S.4). This proves (D.1). □

Let $S_{N,k}$ denote the probability that observation k (with W_i equal to w say) will be used as a match for an arbitrary observation from the opposite treatment arm under the propensity score $F(X'\bar{\theta}_N)$, conditional on both $\mathbf{W}$ and all the observations from its own treatment status, denoted by $\mathbf{X}_w$.

Lemma 14. *Suppose that Assumptions 1-4 hold. Further suppose that for all $\theta \in \mathcal{N}$, the function $l_w(p;\theta) \leq C$ uniformly in both $p \in \mathbb{R}$ and $\theta \in \mathcal{N}$. Then under $\bar{P}_N$,*

$$\frac{1}{N}\sum_{i=1}^{N} l_w\left(p_{i,N};\bar{\theta}_N\right) K_M(i;\bar{\theta}_N) - \frac{N_{1-w,N}}{N}\sum_{i=1}^{N} l_w\left(p_{i,N};\bar{\theta}_N\right) S_{N,i} = o_p(1),$$

and

$$\frac{1}{N}\sum_{i=1}^{N} l_w\left(p_{i,N};\bar{\theta}_N\right) K^2_M(i;\bar{\theta}_N)$$
$$- \frac{1}{N}\sum_{i=1}^{N} l_w\left(p_{i,N};\bar{\theta}_N\right)\left(N^2_{1-w,N}S^2_{N,i} + N_{1-w,N}S_{N,i}(1-S_{N,i})\right) = o_p(1).$$

Proof. The proof of this result is a straightforward extension of Abadie and Imbens (2016, Lemma S.10) and therefore omitted. □

For the next Lemma, let $p_{w,j:N}$ denote j-th order statistic of $\{p_{i,N} : W_{N,i} = w\}$. I set $p_{w,j:N} = a_N$ if $j < 1$ and $p_{w,j:N} = b_N$ if $j > N$, where $[a_N, b_N]$ denotes the interval support of $F(X'\bar{\theta}_N)$. Also let V_i denote the rank of observation i, in terms of $F(X_i'\bar{\theta}_N)$, within the sample of observations that have the same treatment status as itself.

Additionally, define $\chi_{0,\theta}(p) = \frac{p}{1-p}\frac{q_{0,N}}{q_{1,N}}$ for $p \in [a_\theta, b_\theta]$ and $\chi_{1,\theta}(.) = \chi^{-1}_{0,\theta}(.)$, where $[a_\theta, b_\theta]$ denotes the interval support of $F(X'\theta)$. I also set $\chi_{0,N} \equiv \chi_{0,\bar{\theta}_N}$ and $\chi_{1,N} \equiv \chi_{1,\bar{\theta}_N}$. Note that

[18]For the end points I set $G^{-1}_{w,N}(0) = a_N$ and $G^{-1}_{w,N}(1) = b_N$.

$\chi_{w,N}(p) = (g_{1-w,N}/g_{w,N})(p)$ except on the set $\{p \in [a_N, b_N] : g_N(p) = 0\}$, which has Lebesgue measure zero by Assumption 3(iii).

Lemma 15. *Suppose that Assumptions 1-4 hold and that $\bar{\theta}_N \to \theta_0$. Further suppose that for all $\theta \in \mathcal{N}$, the function $l_w(p;\theta)$ is uniformly bounded in both $p \in [a_\theta, b_\theta]$, and $\theta \in \mathcal{N}$. Then for each $w = 0, 1$, under $\bar{P}_N$, (i)*

$$\sum_{i=1}^{N} l_w\left(p_{i,N}; \bar{\theta}_N\right) \times \left(S_{N,i} - \chi_{w,N}(p_{i,N}) \frac{G_{w,N}(p_{w,V_i+M:N_{w,N}}) - G_{w,N}(p_{w,V_i-M:N_{w,N}})}{2}\right) = o_p(1).$$

and (ii)

$$\sum_{i=1}^{N} l_w\left(p_{i,N}; \bar{\theta}_N\right) N_{w,N} \times \left(S_{N,i}^2 - \left(\chi_{w,N}(p_{i,N}) \frac{G_{w,N}(p_{w,V_i+M:N_{w,N}}) - G_{w,N}(p_{w,V_i-M:N_{w,N}})}{2}\right)^2\right) = o_p(1).$$

Proof. In terms of the method for the proof, I adapt the arguments of Abadie and Imbens (2016, Lemma S.7) to allow for triangular arrays. Without loss of generality I prove the above for the case $w = 0$. Also to simplify notation, I set $N_{0,N} = N_0$ for the duration of this proof.

I first show that for any fixed $K < \infty$,

$$\max_{1 \le i \le N_0} |p_{0,V_i+K:N_0} - p_{0,V_i:N_0}| = o_{\bar{P}_N}(1). \tag{D.4}$$

By equation (D.3) in Lemma 13, and the fact $G_0^{-1}(.)$ is uniformly continuous on $[0,1]$, it follows that the sequence $G_{0,N}^{-1}(.)$ is uniformly equicontinuous. Hence for each $\epsilon > 0$, there exists $\delta > 0$ such that

$$\begin{aligned}
&\bar{P}_N\left(\max_{1 \le i \le N_0} |p_{0,V_i+K:N_0} - p_{0,V_i:N_0}| > \epsilon\right) \\
&\le \bar{P}_N\left(\max_{1 \le i \le N_0} |G_{0,N}\left(p_{0,V_i+K:N_0}\right) - G_{0,N}\left(p_{0,V_i:N_0}\right)| > \delta\right) \\
&\le \Pr\left(\max_{1 \le i \le N_0} |\xi_{V_i+K:N_0} - \xi_{V_i:N_0}| > \delta\right) \to 0,
\end{aligned}$$

where the limit follows by standard properties of uniform spacings. This proves (D.4).

Define

$$\Omega_{Ni} \equiv \left[\frac{p_{0,V_i:N_0} + p_{0,V_i+M:N_0}}{2}, \frac{p_{0,V_i:N_0} + p_{0,V_i-M:N_0}}{2}\right].$$

Let

$$Z_{N,i} = l_0\left(p_{i,N}; \bar{\theta}_N\right) N_0 \left(S_{N,i} - h_{0,N}(p_{i,N}) \frac{G_{0,N}\left(p_{0,V_i+M:N_0}\right) - G_{0,N}\left(p_{0,V_i-M:N_0}\right)}{2}\right).$$

As in the proof of Abadie and Imbens (2016, Lemma S.7), note that

$$S_{N,i} = \int_{a_N}^{(p_{0,V_i:N_0}+p_{0,V_i+M:N_0})/2} g_{1,N}(p)dp\ \mathbb{I}_{\{V_i\leq M\}} + \int_{\Omega_{Ni}} g_{1,N}(p)dp\ \mathbb{I}_{\{M<V_i\leq N-M\}} + \int_{(p_{0,V_i:N_0}+p_{0,V_i-M:N_0})/2}^{b_N} g_{1,N}(p)dp\ \mathbb{I}_{\{V_i>N-M\}}. \tag{D.5}$$

Then by the properties of uniform spacings I obtain

$$N_0 S_{N,i} - N_0 \int_{\Omega_{Ni}} g_{1,N}(p)dp = o_{\bar{P}_N}(1). \tag{D.6}$$

Now by the proof of Lemma 11, for each $p \in [a_\theta, b_\theta]$,

$$\left(\frac{g_{1,N}}{g_{0,N}}\right)(p) = \frac{p}{1-p}\frac{q_{0,N}}{q_{1,N}}\mathbb{I}\left\{g_N(p)\neq 0\right\} \equiv \chi_{0,N}(p)\mathbb{I}\left\{g_N(p)\neq 0\right\},$$

where $\{\chi_{0,N}\}$ is uniformly equicontinuous on $p \in [a_N, b_N]$ by Lemma 11. Since $g_{1,N}(p) = g_{0,N}(p) = 0$ whenever $g_N(p) = 0$, the mean value theorem for Lebesgue-Steltjes integrals ensures

$$\int_{\Omega_{Ni}} g_{1,N}(p)dp = \int_{\Omega_{Ni}} \chi_{0,N}(p)g_{0,N}(p)dp = \int_{\Omega_{Ni}} \chi_{0,N}(p)dG_{0,N}(p)$$
$$= \chi_{0,N}\left(\bar{p}_{i,N,M}\right)\left(G_{0,N}\left(\frac{p_{0,V_i:N_0}+p_{0,V_i+M:N_0}}{2}\right) - G_{0,N}\left(\frac{p_{0,V_i:N_0}+p_{0,V_i-M:N_0}}{2}\right)\right).$$

for some $\bar{p}_{i,N,M} \in \Omega_{Nl}$. Substituting in (D.6), a second application of the mean value theorem then implies

$$N_0 S_{N,i} - N_0 \chi_{0,N}\left(\bar{p}_{i,N,M}\right) g_{0,N}\left(\tilde{p}_{i,N,M}\right)\left(p_{0,V_i+M:N_0} - p_{0,V_i-M:N_0}\right)/2 = o_{\bar{P}_N}(1),$$

for some $\tilde{p}_{i,N,M} \in \Omega_{Ni}$. Substituting the above in the expression for Z_{Ni}, and applying the mean value theorem again on $G_{0,N}\left(p_{0,V_i+M:N_0}\right) - G_{0,N}\left(p_{0,V_i-M:N_0}\right)$, I obtain for some $\check{p}_{i,N,M} \in [p_{0,V_i+M:N_0}, p_{0,V_i-M:N_0}]$,

$$Z_{N,i} = o_{\bar{P}_N}(1) + l_0\left(p_{i,N};\bar{\theta}_N\right) N_0 \left\{\chi_{0,N}\left(\bar{p}_{i,N,M}\right) g_{0,N}\left(\tilde{p}_{i,N,M}\right) - \chi_{0,N}\left(p_{i,N}\right) g_{0,N}\left(\check{p}_{i,N,M}\right)\right\}$$
$$\times \left(p_{0,V_i+M:N_0} - p_{0,V_i-M:N_0}\right)/2.$$

Now using (D.4) together with the facts $\{\chi_{0,N}\}$ and $\{g_{0,N}\}$ are uniformly equicontinuous (the latter by Assumption 3-(iii)), it follows $Z_{N,i} = o_p(1)$ under $\bar{P}_N$ for each i.

I now show that for any $r < \infty$, there exists some constant $M_r < \infty$ such that,

$$\bar{E}_N\left|Z_{N,i}\right|^r < M_r \quad \text{for all } 1 \leq i \leq N, \tag{D.7}$$

where the expectation here, and in the rest of the proof, is taken under $\bar{P}_N$. By standard properties of uniform spacings,

$$\bar{E}_N\left|N_{0,N}\left\{G_{0,N}\left(p_{0,V_i+M:N_0}\right) - G_{0,N}\left(p_{0,V_i-M:N_0}\right)\right\}\right|^r$$
$$= \bar{E}_N\left|N_{0,N}\left(\xi_{V_i+M:N_0} - \xi_{V_i-M:N_0}\right)\right|^r < M_{1r}$$

for some constant $M_{1r} < \infty$. Hence, by part (ii) of Lemma 11, and the assumption $l_0(p;\theta)$ is uniformly bounded, it suffices for (D.7) to show $\bar{E}_N |N_0 S_{N,i}|^r$ is uniformly bounded. Let $S_{Ni,(a)}$, $S_{Ni,(b)}$ and $S_{Ni,(c)}$ denote the three terms in that order from the expression for $S_{N,i}$ in equation (D.5). By part (ii) of Lemma 11, and the properties of uniform spacings (see Abadie and Imbens, Lemma S.5; or as applied in their Lemma S.7),

$$\begin{aligned}\bar{E}_N \left|N_0 S_{Ni,(a)}\right|^r &\leq \bar{C}^r \bar{E}_N |N_0 G_{0,N}(p_{0,2M:N_0})|^r \\ &\leq \bar{C}^r \bar{E}_N |N_0 \xi_{2M:N_0}|^r < M_{r,(a)}\end{aligned}$$

for some $M_{r,(a)} < \infty$. A similar argument also shows that $\bar{E}_N \left|N_0 S_{Ni,(c)}\right|^r < M_{r,(c)} < \infty$. Finally, consider

$$\begin{aligned}\left|S_{Ni,(b)}\right|^r &= \left|\int_{\Omega_{Ni}} \chi_{0,N}(p) g_{0,N}(p) dp\right|^r \leq \bar{C}^r \left|\int_{V_i - M:N_0}^{V_i + M:N_0} g_{0,N}(p) dp\right|^r \\ &= \bar{C}^r |\xi_{V_i+M:N_0} - \xi_{V_i-M:N_0}|^r,\end{aligned}$$

where the inequality follows from $\sup_p |\chi_{0,N}(p)| < \bar{C}$ due to Lemma 11. Hence by the properties of uniform spacings, $\bar{E}_N \left|N_0 S_{Ni,(b)}\right|^r < M_{r,(b)} < \infty$. By the above I have thus shown (D.7).

Equation (D.7), together with $Z_{N,i} = o_p(1)$ under $\bar{P}_N$, implies $\bar{E}_N |Z_{Ni}| \to 0$. Since the choice of i was arbitrary, the above holds true for all $1 \leq i \leq N_0$. Hence application of the Markov inequality assures $N^{-1} \sum_{i=1}^N Z_{N,i} = o_p(1)$ under $\bar{P}_N$. This proves the first part of the Lemma. Part (ii) follows by analogous arguments. □

Lemma 16. *Suppose that Assumptions 1-4 hold. Further suppose that for all $\theta \in \mathcal{N}$, the function $m_w(.;\theta) : [\underline{p}, \bar{p}] \to \mathbb{R}$ is non-negative, uniformly equicontinuous in $\mathcal{N}$ i.e*

$$\lim_{\delta \to 0} \sup_{p \in \mathbb{R}, \theta \in \mathcal{N}} |m_w(p;\theta) - m_w(p+\delta;\theta)| = 0,$$

and also satisfies $m_w(p;\dot{\theta}_N) \to m_w(p;\theta_0)$ point-wise in each p for any sequence $\dot{\theta}_N \to \theta_0$. Then for any non-negative integer M, and sequence $\{\bar{\theta}_N\}$ satisfying $\bar{\theta}_N \to \theta_0$ a.s-$\bar{P}_N$, it holds that under $\bar{P}_N$,

$$\begin{aligned}&\sum_{i=1}^N m_w\left(G_{w,N}^{-1}(\xi_{i:N_{w,N}}); \bar{\theta}_N\right)\left(\xi_{i+M:N_{w,N}} - \xi_{i-M:N_{w,N}}\right) \\ &= \frac{2M}{N_{w,N}} \sum_{i=1}^N m_w\left(G_{w,N}^{-1}(\xi_{i:N_{w,N}}); \bar{\theta}_N\right) + o_p(1),\end{aligned}$$

and

$$\begin{aligned}&N_{w,N} \sum_{i=1}^N m_w\left(G_{w,N}^{-1}(\xi_{i:N_{w,N}}); \bar{\theta}_N\right)\left(\xi_{i+M:N_{w,N}} - \xi_{i-M:N_{w,N}}\right)^2 \\ &= \frac{2M(2M+1)}{N_{w,N}} \sum_{i=1}^N m_w\left(G_{w,N}^{-1}(\xi_{i:N_{w,N}}); \bar{\theta}_N\right) + o_p(1).\end{aligned}$$

Proof. Using Lemma 6 and Lemma 13, the proof of this result is a straightforward extension of Abadie and Imbens (2016, Lemma S.6), and therefore omitted. □

References

1. Alberto Abadie and Guido W Imbens, *Large sample properties of matching estimators for average treatment effects*, Econometrica **74** (2006), no. 1, 235–267.
2. ———, *On the failure of the bootstrap for matching estimators*, Econometrica **76** (2008), no. 6, 1537–1557.
3. ———, *Bias-corrected matching estimators for average treatment effects*, Journal of Business & Economic Statistics **29** (2011), no. 1, 1–11.
4. ———, *A martingale representation for matching estimators*, Journal of the American Statistical Association **107** (2012), no. 498, 833–843.
5. ———, *Matching on the estimated propensity score*, Econometrica **84** (2016), no. 2, 781–807.
6. Elena Andreou and Bas JM Werker, *An alternative asymptotic analysis of residual-based statistics*, Review of Economics and Statistics **94** (2012), no. 1, 88–99.
7. Peter J Bickel, Friedrich Götze, and Willem R van Zwet, *Resampling fewer than n observations: gains, losses, and remedies for losses*, Selected Works of Willem van Zwet, Springer, 2012, pp. 267–297.
8. P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J.A. Wellner, *Efficient and adaptive estimation for semiparametric models*, Johns Hopkins series in the mathematical sciences, Springer New York, 1998.
9. Matias Busso, John DiNardo, and Justin McCrary, *New evidence on the finite sample properties of propensity score reweighting and matching estimators*, Review of Economics and Statistics **96** (2014), no. 5, 885–897.
10. Matias D Cattaneo, Michael Jansson, and Whitney K Newey, *Alternative asymptotics and the partially linear model with many regressors*, Econometric Theory (2016), 1–25.
11. William G Cochran, *The effectiveness of adjustment by subclassification in removing bias in observational studies*, Biometrics (1968), 295–313.
12. Rajeev H Dehejia and Sadek Wahba, *Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs*, Journal of the American statistical Association **94** (1999), no. 448, 1053–1062.
13. Bradley Efron and Charles Stein, *The jackknife estimate of variance*, The Annals of Statistics (1981), 586–596.
14. James Heckman, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd, *Characterizing selection bias using experimental data*, Econometrica **66** (1998), no. 5, 1017–1098.
15. James J Heckman and V Joseph Hotz, *Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training*, Journal of the American statistical Association **84** (1989), no. 408, 862–874.
16. James J Heckman, Hidehiko Ichimura, and Petra E Todd, *Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme*, The review of economic studies **64** (1997), no. 4, 605–654.
17. Guido W Imbens and Donald B Rubin, *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press, 2015.
18. Joseph DY Kang and Joseph L Schafer, *Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data*, Statistical science (2007), 523–539.
19. Shakeeb Khan and Elie Tamer, *Irregular identification, support conditions, and inverse weight estimation*, Econometrica **78** (2010), no. 6, 2021–2042.
20. Robert J LaLonde, *Evaluating the econometric evaluations of training programs with experimental data*, The American economic review (1986), 604–620.
21. Michael Lechner, *Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies*, Review of Economics and Statistics **84** (2002), no. 2, 205–220.
22. Whitney K Newey and Daniel McFadden, *Large sample estimation and hypothesis testing*, Handbook of econometrics **4** (1994), 2111–2245.
23. Taisuke Otsu and Yoshiyasu Rai, *Bootstrap inference of matching estimators for average treatment effects*, Journal of the American Statistical Association (2016), no. just-accepted.
24. Dimitris N Politis and Joseph P Romano, *Large sample confidence regions based on subsamples under minimal assumptions*, The Annals of Statistics (1994), 2031–2050.

25. David Pollard, *Convergence of stochastic processes*, Springer Science & Business Media, 2012.
26. Paul R Rosenbaum, *Optimal matching for observational studies*, Journal of the American Statistical Association **84** (1989), no. 408, 1024–1032.
27. _____, *Design of observational studies*, Springer Science & Business Media, 2009.
28. Paul R Rosenbaum and Donald B Rubin, *The central role of the propensity score in observational studies for causal effects*, Biometrika (1983), 41–55.
29. _____, *Reducing bias in observational studies using subclassification on the propensity score*, Journal of the American statistical Association **79** (1984), no. 387, 516–524.
30. Donald B Rubin, *Estimating causal effects of treatments in randomized and nonrandomized studies.*, Journal of educational Psychology **66** (1974), no. 5, 688.
31. Jeffrey A Smith and Petra E Todd, *Reconciling conflicting evidence on the performance of propensity-score matching methods*, The American Economic Review **91** (2001), no. 2, 112–118.
32. VN Vapnik and A Ya Chervonenkis, *On the uniform convergence of relative frequencies of events to their probabilities*, Theory of Probability and its Applications **16** (1971), no. 2, 264.

Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK.

E-mail address: k.adusumilli@lse.ac.uk