A Martingale Representation for Matching Estimators

Alberto Abadie – Harvard University and NBER
Guido W. Imbens – Harvard University and NBER

September 2011

Abstract

Matching estimators (Rubin, 1973a, 1977; Rosenbaum, 2002) are widely used in statistical data analysis. However, the large sample distribution of matching estimators has been derived only for particular cases (Abadie and Imbens, 2006). This article establishes a martingale representation for matching estimators. This representation allows the use of martingale limit theorems to derive the large sample distribution of matching estimators. As an illustration of the applicability of the theory, we derive the asymptotic distribution of a matching estimator when matching is carried out without replacement, a result previously unavailable in the literature. In addition, we apply the techniques proposed in this article to derive a correction to the standard error of a sample mean when missing data are imputed using the "hot deck", a matching imputation method widely used in the Current Population Survey (CPS) and other large surveys in the social sciences. We demonstrate the empirical relevance of our methods using two Monte Carlo designs based on actual data sets. In these realistic Monte Carlo exercises the large sample distribution of matching estimators derived in this article provides an accurate approximation to the small sample behavior of these estimators. In addition, our simulations show that standard errors that do not take into account hot deck imputation of missing data may be severely downward biased, while standard errors that incorporate the correction proposed in this article for hot deck imputation perform extremely well. This result demonstrates the practical relevance of the standard error correction for the hot deck proposed in this article.

Keywords: *Matching, Martingales, Treatment Effects, Hot Deck Imputation*

Alberto Abadie, John F. Kennedy School of Government, 79 John F. Kennedy Street, Cambridge MA 02138, USA. E-mail: alberto_abadie@harvard.edu. Guido W. Imbens, Department of Economics, 1805 Cambridge Street, Cambridge, MA 02138. E-mail: imbens@harvard.edu. We thank Rustam Ibragimov, Don Rubin, and seminar participants at Harvard/MIT, Brown, Georgetown, UPenn, Montreal, and the 2009 Conference on Causal Inference in Statistics and the Quantitative Sciences at the Banff International Research Station for useful comments and suggestions. We are also grateful to Greg Weyland, from the US Census Bureau, for patiently explaining to us the intricacies of the hot deck imputation algorithms employed in the Current Population Survey.

I. Introduction

Matching methods provide simple and intuitive tools for adjusting the distribution of covariates among samples from different populations. Probably because of their transparency and intuitive appeal, matching methods are widely used in evaluation research to estimate treatment effects when all treatment confounders are observed (Rubin, 1977; Dehejia and Wahba, 1999; Rosenbaum, 2002, Hansen, 2004). Matching is also used for the analysis of missing data, where it is often referred to as "hot deck imputation" (Little and Rubin, 2002). As a notorious example, missing weekly earnings are currently imputed using hot deck methods for more than 30 percent of the records with weekly earnings data in the monthly U.S. Current Population Survey (CPS) files (Bollinger and Hirsch, 2009).

In spite of the pervasiveness of matching methods, the asymptotic distribution of matching estimators has been derived only for special cases (Abadie and Imbens, 2006). In the absence of large sample approximation results to the distribution of matching estimators, empirical researchers employing matching methods have sometimes used the bootstrap as a basis for inference. However, recent results have shown that, in general, the bootstrap does not provide valid large sample inference for matching estimators (Abadie and Imbens, 2008). Similarly, the properties of statistics based on data imputed using sequential hot deck methods, like those employed in the CPS and other large surveys, are not well-understood, and empirical researchers using these surveys typically ignore missing data imputation issues when they construct standard errors. Andridge and Little (2010) provide a recent survey on hot deck imputation methods.

The main contribution of this article is to establish a martingale representation for matching estimators. This representation allows the use of martingale limit theorems (Hall and Heyde, 1980; Billingsley, 1995; Shorack, 2000) to derive the asymptotic distribution of matching estimators. Because the martingale representation applies to a large class of matching estimators, the applicability of the methods presented in this article is very broad. Despite its simplicity and immediate implications, the martingale representation of matching estimators described in this article seems to have been previously unnoticed

in the literature. The use of martingale methods is attractive because the limit behavior of martingale sequences has been extensively studied in the statistics literature (see, for example, Hall and Heyde, 1980).

As an illustration of the usefulness of the theory, we apply the martingale methods proposed in this paper to derive the asymptotic distribution of a matching estimator when matching is carried out without replacement, a result previously unavailable in the literature. In addition, we apply the techniques proposed in this article to derive a correction to the standard error of a sample mean when missing data are imputed using the hot deck.

Finally, we demonstrate the empirical relevance of our methods using two Monte Carlo designs based on actual data sets. In these realistic Monte Carlo exercises the large sample distribution of matching estimators derived in this article provides an accurate approximation to the small sample behavior of these estimators. In addition, our simulations show that standard errors that do not take into account hot deck imputation of missing data may be severely downward biased while standard errors that incorporate the correction proposed in this article for hot deck imputation perform extremely well. This result demonstrates the practical relevance of the standard error correction for the hot deck proposed in this article.

In this article we reserve the term "matching" for procedures that use a small number of matches per unit. Heckman, Ichimura, and Todd (1998) have proposed estimators that treat the number of matches as an increasing function of the sample size. Under certain conditions, these estimators have asymptotically linear representations, so their large sample distributions can be derived using the standard machinery for asymptotically linear estimators. In contrast, despite the pervasiveness of matching estimators that use a small number of matches (e.g., hot deck imputation in the CPS), the previous literature does not provide a general framework for establishing their large sample properties.

The rest of the article is organized as follows. Section II describes matching estimators. Section III presents the main result of the article, which establishes a martingale representation for matching estimators. In section IV, we apply martingale techniques to

analyze the large sample properties of a matching estimator when matching is carried out without replacement. In section V, we apply martingale techniques to study hot deck imputation. Section VI describes of the Monte Carlo simulation exercises and reports the results. Section VII concludes.

II. Matching Estimators

Let W be a binary variable that indicates membership to a particular population of interest. Empirical researchers often compare the distributions of some variable, Y, between units with $W = 1$ and units with $W = 0$ after adjusting for the differences in a $(k \times 1)$ vector of observed covariates, X. For example, in discrimination litigation research, W may represent membership in a certain demographic group, Y may represent labor wages, and X may represent a vector of variables describing job and/or worker characteristics. In evaluation research, W typically indicates exposure to an active treatment or intervention, Y is an outcome of interest, and X is a vector of observed confounders. As in that literature, we will say that units with $W = 1$ are "treated" and units with $W = 0$ are "untreated". Let

$$\tau = E[Y|W = 1] - E\Big[E[Y|X, W = 0]\Big|W = 1\Big]. \tag{1}$$

In evaluation research, τ is given a causal interpretation as the "average treatment effect on the treated" under unconfoundedness assumptions (Rubin, 1977). Applied researchers often use matching methods to estimate τ. Other parameters of interest that can be estimated by matching methods include: (i) the "average treatment effect", which is of widespread interest in evaluation studies, (ii) parameters that focus on features of the distribution of Y other than the mean, (iii) parameters estimated by hot deck imputation methods in the presence of missing data. Rosenbaum (2002), Imbens (2004), and Rubin (2006) provide detailed surveys of the literature. For concreteness, and to avoid tedious repetition or unnecessary abstraction, in this section we discuss matching estimation of τ only. However, the techniques proposed in this paper are of immediate application to the estimation of parameters other than τ via matching (see, for example, section V).

Also, to avoid notational clutter, we consider only estimators with a fixed number of

matches, M, per unit. However, as it will be explained later, our techniques can also be applied to estimators for which the number of matches may differ across units (see, e.g., Hansen, 2004).

Consider two random samples of sizes N_0 and N_1 of untreated and treated units, respectively. Pooling these two samples, we obtain a sample of size $N = N_0 + N_1$ containing both treated and untreated units. For each unit in the pooled sample we observe the triple (Y, X, W). For each treated unit i, let $\mathcal{J}_M(i)$ be the indices of M untreated units with values in the covariates similar to X_i (where M is some small positive integer). In other words, $\mathcal{J}_M(i)$ is a set of M matches for observation i. To simplify notation, we will assume that at least one of the variables in the vector X has a continuous distribution, so perfect matches happen with probability zero. Let $\|\cdot\|$ be some norm in $\mathbb{R}^k$ (typically the Euclidean norm). Let 1_A be the indicator function for the event A. For matching with replacement $\mathcal{J}_M(i)$ consists of the indices of the M untreated observations with the closest value covariate values to X_i:

$$\mathcal{J}_M(i) = \left\{ j \in \{1, \ldots, N\} \text{ s.t. } W_j = 0, \ \left(\sum_{k=1}^{N} (1 - W_k)\, 1_{\{\|X_i - X_j\| \leq \|X_i - X_k\|\}} \right) \leq M \right\}.$$

For matching without replacement, the elements of $\{\mathcal{J}_M(i) \text{ s.t. } W_i = 1\}$ are non-overlapping subsets of $\{j \in \{1, \ldots, N\} \text{ s.t. } W_j = 0\}$, chosen to minimize the sum of the matching discrepancies:

$$\sum_{i=1}^{N} W_i \sum_{j \in \mathcal{J}_M(i)} \|X_i - X_j\|.$$

In both cases, the matching estimator of τ is defined as:

$$\widehat{\tau} = \frac{1}{N_1} \sum_{i=1}^{N} W_i \Big(Y_i - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} Y_j \Big). \tag{2}$$

Many other matching schemes are possible (see, e.g., Gu and Rosenbaum, 1993; Rosenbaum, 2002; Hansen, 2004; Diamond and Sekhon, 2008; Iacus, King, and Porro, 2009), and the results in this article are of broad generality. However, as discussed above, our results pertain to matching estimators that employ a small number, M, of matches per unit.

Heckman, Ichimura, and Todd (1998) have proposed "kernel matching" estimators, which require that the number of matches increase with the sample size (with $M \to \infty$ as $N \to \infty$) in order to consistently estimate the conditional expectation function $E[Y|X, W = 0]$ in equation (1). In addition, the results of this article apply to estimators that match directly on the covariates, X, and do not directly apply to matching on the estimated propensity score (Rosenbaum and Rubin, 1983). Abadie and Imbens (2010) derive an adjustment to the distribution of the propensity score matching estimators for the case when the propensity score is not known, so matching is done on a first step estimator of the propensity score.

III. A Martingale Representation for Matching Estimators

This section derives a martingale representation for matching estimators. For $w \in \{0, 1\}$, let $\mu_w(x) = E[Y|X = x, W = w]$ and $\sigma_w^2(x) = \text{var}(Y|X = x, W = w)$. Given equation (2), we can write $\widehat{\tau} - \tau = D_N + R_N$, where

$$D_N = \frac{1}{N_1} \sum_{i=1}^{N} W_i \big(\mu_1(X_i) - \mu_0(X_i) - \tau\big) + \frac{1}{N_1} \sum_{i=1}^{N} W_i \Big(\big(Y_i - \mu_1(X_i)\big) - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} \big(Y_j - \mu_0(X_j)\big)\Big),$$

and

$$R_N = \frac{1}{N_1} \sum_{i=1}^{N} W_i \Big(\mu_0(X_i) - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} \mu_0(X_j)\Big).$$

The term R_N is the conditional bias of matching estimator described in Abadie and Imbens (2006). This term is zero if all matches are perfect (that is, if all matching discrepancies, $X_i - X_j$ for $j \in \mathcal{J}_M(i)$, are zero), or if the regression μ_0 is a constant function. In most cases of interest, however, this term is different from zero, as perfect matches happen with probability zero for continuous covariates. The order of magnitude of R_N depends on the number of continuous covariates, as well as the magnitude of N_0 relative to N_1. Under appropriate conditions $\sqrt{N}_1 R_N$ converges in probability to zero (see section IV for the case of matching without replacement, or Abadie and Imbens, 2006, for the case of matching

with replacement).

Next, it will be shown that the term D_N is a martingale array with respect to a certain filtration. First notice that:

$$D_N = \frac{1}{N_1}\sum_{i=1}^{N} W_i\big(\mu_1(X_i) - \mu_0(X_i) - \tau\big) + \frac{1}{N_1}\sum_{i=1}^{N}\left(W_i - (1-W_i)\frac{K_{N,i}}{M}\right)\big(Y_i - \mu_{W_i}(X_i)\big),$$

where $K_{N,i}$ is the number of times that observation i (with $W_i = 0$) is used as a match:

$$K_{N,i} = \sum_{j=1}^{N} 1_{\{i\in\mathcal{J}_M(j)\}}.$$

Therefore, we can write:

$$\sqrt{N_1}D_N = \sum_{k=1}^{2N}\xi_{N,k},$$

where

$$\xi_{N,k} = \begin{cases} \dfrac{1}{\sqrt{N_1}}\,W_k\big(\mu_1(X_k) - \mu_0(X_k) - \tau\big) & \text{if } 1 \le k \le N, \\ \dfrac{1}{\sqrt{N_1}}\left(W_{k-N} - (1-W_{k-N})\dfrac{K_{N,k-N}}{M}\right)\big(Y_{k-N} - \mu_{W_{k-N}}(X_{k-N})\big) & \text{if } N+1 \le k \le 2N. \end{cases}$$

Let $\mathbf{X}_N = \{X_1, \ldots, X_N\}$ and $\mathbf{W}_N = \{W_1, \ldots, W_N\}$. Consider the σ-fields $\mathcal{F}_{N,k} = \sigma\{\mathbf{W}_N, X_1, \ldots, X_k\}$ for $1 \le k \le N$ and $\mathcal{F}_{N,k} = \sigma\{\mathbf{W}_N, \mathbf{X}_N, Y_1, \ldots, Y_{k-N}\}$ for $N+1 \le k \le 2N$. Then, and this is the key insight in this article,

$$\left\{\sum_{j=1}^{i}\xi_{N,j}, \mathcal{F}_{N,i}, 1 \le i \le 2N\right\}$$

is a martingale for each $N \ge 1$. As a result, the asymptotic behavior of $\sqrt{N_1}D_N$ can be analyzed using martingale methods. Analogous martingale representations hold for alternative matching estimators. Regardless of the choice of matching scheme, if matches depend only on the covariates X, a martingale representation holds for $\sqrt{N_1}D_N$. The reason is that no matter how matching is implemented, (i) the number of times that unit k is used as a match, $K_{N,k}$, is a deterministic function of $\mathbf{X}_N$ and $\mathbf{W}_N$, and (ii) $E[Y_k - \mu_{W_k}(X_k) \,|\, \mathbf{X}_N, \mathbf{W}_N, Y_1, \ldots, Y_{k-1}] = 0$.

So far, we have considered the case where $K_{N,i}$ is fixed given $\mathbf{X}_N$ and $\mathbf{W}_N$, for all $1 \leq i \leq N$. This assumption does not hold for certain matching schemes that break matching ties using randomization. Notice, however, that any sequence of randomized tie-breaks can be included in the set of variables that span $\mathcal{F}_{N,k}$ for $N+1 \leq k \leq 2N$ to preserve the martingale representation of D_N. As a result, our derivations extend easily to randomized matching methods.

IV. Application: Matching without Replacement

In this section, we illustrate the usefulness of the martingale representation of matching estimators by deriving the asymptotic distribution of a matching estimator when matching is done without replacement, so $K_{N,i} \in \{0,1\}$ for every unit i with $W_i = 0$. To simplify the exposition we obviate some regularity conditions in the derivations. A precise statement of the result, including all regularity conditions, is provided at the end of the section.

For $1 \leq k \leq N$, the conditional variances of the martingale differences are given by:

$$\begin{aligned} E[\xi_{N,k}^2|\mathcal{F}_{N,k-1}] &= \frac{1}{N_1} W_k E[(\mu_1(X_k) - \mu_0(X_k) - \tau)^2 | \mathcal{F}_{N,k-1}] \\ &= \frac{1}{N_1} W_k E[(\mu_1(X_k) - \mu_0(X_k) - \tau)^2 | W_k = 1]. \end{aligned}$$

For $N+1 \leq k \leq 2N$, the conditional variances of the martingale differences are given by:

$$\begin{aligned} E[\xi_{N,k}^2|\mathcal{F}_{N,k-1}] &= \frac{1}{N_1} E\left[\left(W_{k-N} - (1-W_{k-N}) \frac{K_{N,k-N}}{M} \right)^2 \left(Y_{k-N} - \mu_{W_{k-N}}(X_{k-N}) \right)^2 \middle| \mathcal{F}_{N,k-1} \right] \\ &= \frac{1}{N_1} \left(W_{k-N} \sigma_1^2(X_{k-N}) + (1-W_{k-N}) \frac{K_{N,k-N}}{M^2} \sigma_0^2(X_{k-N}) \right) \\ &= \frac{1}{N_1} W_{k-N} \left(\sigma_1^2(X_{k-N}) + \frac{\sigma_0^2(X_{k-N})}{M} \right) + r_{N,k-N}, \end{aligned}$$

where

$$r_{N,k-N} = \frac{1}{N_1} \left((1-W_{k-N}) \frac{K_{N,k-N}}{M^2} \sigma_0^2(X_{k-N}) - W_{k-N} \frac{\sigma_0^2(X_{k-N})}{M} \right).$$

Assume that the conditional variance function $\sigma_0^2(x)$ is Lipschitz-continuous, with Lipschitz constant equal to c_1. For $1 \leq i \leq N$ such that $W_i = 1$, let $\|U_{N_0,N_1,i}^{(M,m)}\|$ be the m-th matching discrepancy for treated unit i when untreated units are matched without replacement to

treated units in such a way that the sum of the matching discrepancies is minimized. That is, if unit i is a treated observation, and unit j is the m-th match for unit i, then $\|U^{(M,m)}_{N_0,N_1,i}\| = \|X_i - X_j\|$. Lipschitz-continuity of $\sigma_0^2(x)$ implies:

$$\left| \sum_{k=N+1}^{2N} r_{N,k-N} \right| \leq \frac{c_1}{M^2} \frac{1}{N_1} \sum_{i=1}^{N} \sum_{m=1}^{M} W_i \|U^{(M,m)}_{N_0,N_1,i}\|.$$

Because the average matching discrepancy converges to zero in probability (see Proposition 1 in the appendix for a stronger result), the Weak Law of Large Numbers implies

$$\sum_{k=1}^{2N} E[\xi^2_{N,k} | \mathcal{F}_{N,k-1}] \xrightarrow{p} \sigma^2,$$

where

$$\sigma^2 = E[(\mu_1(X) - \mu_0(X) - \tau)^2 | W = 1] + E\left[\sigma_1^2(X) + \frac{\sigma_0^2(X)}{M} \Big| W = 1 \right]. \tag{3}$$

In view of this result, to apply a Martingale Central Limit Theorem to D_N, it is sufficient to check the Lindeberg condition,

$$\sum_{k=1}^{2N} E[\xi^2_{N,k} 1_{\{|\xi_{N,k}| \geq \varepsilon\}}] \to 0 \qquad \text{for all } \varepsilon > 0$$

(Billingsley, 1995, see Hall and Heyde, 1980, and Shorack, 2000, for alternative conditions). Because for all $\delta > 0$, $|\xi_{N,k}|^2 1_{\{|\xi_{N,k}| \geq \varepsilon\}} \varepsilon^\delta \leq |\xi_{N,k}|^{2+\delta}$, it follows that Lindeberg's condition is implied by Lyapounov's condition:

$$\sum_{k=1}^{2N} E[|\xi_{N,k}|^{2+\delta}] \to 0 \qquad \text{for some } \delta > 0,$$

For the matching estimators considered in this section, Lyapounov's condition can be established imposing regularity conditions on the existence of moments (like condition *(iii)* in the statement of Theorem 1 below). Then, the Central Limit Theorem for Triangular Martingale Arrays implies:

$$\sqrt{N_1} D_N \xrightarrow{d} N(0, \sigma^2).$$

The proof concludes by showing that $\sqrt{N_1}R_N \xrightarrow{p} 0$. If μ_0 is Lipschitz-continuous, then there exists a constant c_2 such that

$$\sqrt{N_1}R_N \leq c_2 \frac{1}{\sqrt{N_1}} \frac{1}{M} \sum_{i=1}^{N} \sum_{m=1}^{M} W_i \| U_{N_0,N_1,i}^{(M,m)} \|.$$

Proposition 1 in the appendix shows that under some conditions, and if there exists $c > 0$ and $r > k$ where k is the number of (continuous) covariates, such that $N_1^r/N_0 \leq c$, then,

$$\frac{1}{\sqrt{N_1}} \sum_{i=1}^{N} \sum_{m=1}^{M} W_i \| U_{N_0,N_1,i}^{(M,m)} \| \xrightarrow{p} 0,$$

so $\sqrt{N_1}R_N$ vanishes asymptotically.

We now collect in a Theorem the result of this section along with precise regularity conditions.

THEOREM 1: *Suppose that (i)* $\{Y_i, X_i, W_i\}_{i=1}^N$ *is a pooled sample of* N_1 *treated and* N_0 *untreated observations obtained by random sampling from their respective population counterparts, (ii) the support of* X *given* $W = 1$ *is a subset of the support of* X *given* $W = 0$*, (iii) for some* $\delta > 0$*, and* $w = 0, 1$*,* $E[|Y|^{2+\delta}|X = x, W = w]$ *is bounded on the support of* X *given* $W = w$*, (iv) the functions* $\mu_0(\cdot)$ *and* $\sigma_0^2(\cdot)$ *are Lipschitz-continuous, and (v)* $(1/\sqrt{N_1}) \sum_{i=1}^{N} \sum_{m=1}^{M} W_i \| U_{N_0,N_1,i}^{(M,m)} \| \xrightarrow{p} 0$ *as* $N_1 \to \infty$*. Then,* $\sqrt{N_1}(\widehat{\tau} - \tau) \xrightarrow{d} N(0, \sigma^2)$ *as* $N_1 \to \infty$*.*

Assumption *(v)* in Theorem 1 is not primitive and Proposition 1 in the appendix provides a set of primitive regularity conditions under which assumption *(v)* holds. The conditions of Proposition 1 assume that all covariates have continuous distributions. This is done without loss of generality for large enough samples. As sample sizes increase discrete covariates with a finite number of support points are perfectly matched, so they can be easily dealt with by conditioning on their values, in which case k is equal to the number of continuous covariates in X. In practice, however, discrete covariates may not be perfectly matched, and may therefore contribute to the bias of the matching estimator.

The proof of Proposition 1 indicates that the support conditions in this proposition can also be relaxed. However, the requirement that the size of the untreated group is of

larger order of magnitude than the size of the treated group is crucial to the result in the proposition. To see that $r = 1$ is not sufficient (even in the one-dimensional case), consider the case with $M = 1$ and $N_0 = N_1$. Then, because matching is done without replacement and all treated units are matched, the matching estimator is equal to the difference in sample means of Y between treated and nontreated, regardless of the total sample size N.

Proposition 1 provides conditions under which matching discrepancies are negligible in large samples. In practical terms, Proposition 1 demonstrates the benefits of having a large "donor pool" of control units for matching estimators. However, for any given sample matching discrepancies are observed, and researchers can assess the quality of the matches directly from the data.

When matching discrepancies are large the resulting bias can be eliminated or reduced using the bias correction techniques in Rubin (1973b), Quade (1982), and Abadie and Imbens (2009). These authors propose a bias-corrected matching estimator that adjusts each matched pair for its contribution to the conditional bias term:

$$\widehat{\tau}_{\text{bc}} = \frac{1}{N_1} \sum_{i=1}^{N} W_i \Big((Y_i - \widehat{\mu}_0(X_i)) - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} (Y_j - \widehat{\mu}_0(X_{j(i)})) \Big), \tag{4}$$

where $\widehat{\mu}_0(\cdot)$ is an estimator of $\mu_0(\cdot)$. Under certain conditions, Abadie and Imbens (2009) show that this bias-correction technique eliminates the asymptotic bias of a matching with replacement estimator without affecting its asymptotic variance.

Straightforward calculations show that the variance estimator

$$\widehat{\sigma}^2 = \frac{1}{N_1 - 1} \sum_{i=1}^{N} W_i \Big(Y_i - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} Y_j - \widehat{\tau} \Big)^2 \tag{5}$$

is consistent for σ^2. Despite the simplicity of this result, to our knowledge the validity of $\widehat{\sigma}^2/N_1$ as an estimator of the variance of $\widehat{\tau}$ when matching is done without replacement has not been established previously. Conversely, it is known that $\widehat{\sigma}^2/N_1$ is not a valid estimator of the variance of $\widehat{\tau}$ when matching is done with replacement (Abadie and Imbens, 2006).

V. Application: Hot Deck Imputation

In this section, we consider a "cell hot deck" imputation scheme where incomplete records of Y are imputed using complete observations within the same "cell" of the covariates, X. That is, the support of the covariates is partitioned into T cells, $\mathcal{C}_1, \ldots \mathcal{C}_T$, and each incomplete record of Y is filled using a complete record from the same cell. Other hot deck imputation procedures are possible (see, for example, Little and Rubin, 2002). However, the cell hot deck methods is probably the most widely used in practice, as it is the one used by the US Census Bureau to impute missing data in the Current Population Survey (CPS), the decennial census, the Survey of Income and Program Participation (SIPP), and other large surveys. Derivations similar to the ones presented in this section can be applied to alternative hot deck imputation schemes.

Cell hot deck imputation methods like the one employed in the CPS can be justified by a "Missing and Coarsening at Random" assumption. Let W be an indicator for complete record, that is $W = 1$ indicates that Y is observed. A missing and coarsening at random assumption states that Y is independent of (X, W) conditional on $X \in \mathcal{C}_t$, for $1 \leq t \leq T$. Missing and coarsening at random may be a strong assumption in many contexts where data are imputed using the cell hot deck. However, without this assumption, or a similar one, the cell hot deck will produce inconsistent estimators in general. Therefore, in our analysis we assume missing and coarsening at random. Also, we restrict our derivations to the case of simple random sampling. In practice, Let $\mu = E[Y]$, $\mu(x) = E[Y|X = x]$, $\mu_t = E[Y|X \in \mathcal{C}_t]$ and $\sigma_t^2 = \text{var}(Y|X \in \mathcal{C}_t)$. Let $j(i)$ be the index of the observation used to impute Y for observation i (if $W_i = 1$, then $j(i) = i$). Let

$$\begin{aligned} \bar{Y} &= \frac{1}{N} \sum_{i=1}^{N} Y_{j(i)} \\ &= \frac{1}{N} \sum_{i=1}^{N} W_i (1 + K_{N,i}) Y_i, \end{aligned} \tag{6}$$

where now $K_{N,i}$ is the number of times that observation i is used to impute an incomplete record. The variables $K_{N,i}$ depend on how imputations are chosen from the complete records within a cell. One possibility is the *random cell hot deck*, which imputes missing

records using a record chosen at random among the complete observation in the same cell. The CPS and other large surveys use a more complicated procedure called the *sequential cell hot deck*. The sequential cell hot deck imputes missing records using the last complete record in the same cell. That is, unlike the random cell hot deck, the sequential cell hot deck uses information about the order of the observations in the sample.

Notice that

$$\begin{aligned}\bar{Y} - \mu = & \frac{1}{N}\sum_{i=1}^{N}(\mu(X_i) - \mu) \\ & + \frac{1}{N}\sum_{i=1}^{N} W_i(1 + K_{N,i})(Y_i - \mu(X_i)) \\ & + \frac{1}{N}\sum_{i=1}^{N}(\mu(X_{j(i)}) - \mu(X_i)).\end{aligned}$$

By the Missing and Coarsening at Random assumption, $\mu(X_{j(i)}) - \mu(X_i) = 0$ for all i. Assume that the second moment of $K_{N,i}$ exists, and that for each cell, t, we have:

$$\left| \frac{1}{N_t}\sum_{i=1}^{N} 1_{\{X_i \in \mathcal{C}_t\}} W_i(1 + K_{N,i})^2 - E\left[\frac{1}{N_t}\sum_{i=1}^{N} 1_{\{X_i \in \mathcal{C}_t\}} W_i(1 + K_{N,i})^2\right] \right| \xrightarrow{p} 0, \qquad (7)$$

which can be usually established using negative association properties of $\{K_{N,i}$ s.t. $W_i = 1, X_i \in \mathcal{C}_t\}$ (Joag-Dev and Proschan, 1983). We can write:

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{N}} = \sum_{k=1}^{2N} \xi_{N,k},$$

where

$$\sigma^2 = E\left[\sum_{t=1}^{T}\left(\frac{N_t}{N}\right)(\mu_t - \mu)^2\right] + E\left[\sum_{t=1}^{T}\left(\frac{N_t}{N}\right)\sigma_t^2 \frac{1}{N_t}\sum_{i=1}^{N} 1_{\{X_i \in \mathcal{C}_t\}} W_i(1 + K_{N,i})^2\right],$$

and

$$\xi_{N,k} = \begin{cases} \dfrac{1}{\sigma\sqrt{N}}(\mu(X_k) - \mu) & \text{if } 1 \le k \le N, \\ \dfrac{1}{\sigma\sqrt{N}} W_{k-N}(1 + K_{N,k-N})\big(Y_{k-N} - \mu(X_{k-N})\big) & \text{if } N+1 \le k \le 2N. \end{cases}$$

Let $\mathbf{X}_N = \{X_1, \ldots, X_N\}$, $\mathbf{W}_N = \{W_1, \ldots, W_N\}$ and $\mathbf{J}_N = \{j(1), \ldots, j(N)\}$. Consider the σ-fields $\mathcal{F}_{N,k} = \sigma\{W_1, \ldots, W_k, X_1, \ldots, X_k\}$ for $1 \le k \le N$ and $\mathcal{F}_{N,k} = \sigma\{\mathbf{W}_N, \mathbf{X}_N,$

$\mathbf{J}_N, Y_1, \ldots, Y_{k-N}\}$ for $N+1 \leq k \leq 2N$. Then,

$$\left\{ \sum_{j=1}^{i} \xi_{N,j}, \mathcal{F}_{N,i}, 1 \leq i \leq 2N \right\}$$

is a martingale for each $N \geq 1$. Equation (7) along with the Central Limit Theorem for martingale arrays (e.g., Theorem 3.2 in Hall and Heyde, 1980) imply:

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{N}} \xrightarrow{d} N(0,1). \tag{8}$$

We now present the result of this section in the form of a Theorem, along with precise regularity condition.

THEOREM 2: *Suppose that (i)* $\{X_1, \ldots, X_N\}_{i=1}^N$ *are sampled at random from the population of interest, (ii)* $\Pr(W = 1 | X \in \mathcal{C}_t) > 0$*, for* $t = 1, \ldots, T$*, (iii)* Y *is independent of* (W, X) *given* $X \in \mathcal{C}_t$*, for* $t = 1, \ldots, T$*, (iv)* $\text{var}(Y) > 0$*, and (v) for some* $\delta > 0$*,* $E[|Y|^{2+\delta}] < \infty$*. Then, equation (8) holds.*

Consider now the usual variance estimator that ignores missing data imputation:

$$\widehat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_{j(i)} - \bar{Y})^2. \tag{9}$$

Notice that

$$\left| \widehat{\sigma}^2 - \sum_{t=1}^{T} \left(\frac{N_t}{N} \right) (\mu_t - \mu)^2 - \sum_{t=1}^{T} \left(\frac{N_t}{N} \right) \sigma_t^2 \right| \xrightarrow{p} 0.$$

In addition, because $\sum_{i=1}^{N} 1_{\{X_i \in \mathcal{C}_t\}} W_i (1 + K_{N,i}) = N_t$, then

$$\frac{1}{N_t} \sum_{i=1}^{N} 1_{\{X_i \in \mathcal{C}_t\}} W_i (1 + K_{N,i})^2 = 1 + \frac{1}{N_t} \sum_{i=1}^{N} 1_{\{X_i \in \mathcal{C}_t\}} W_i (K_{N,i}^2 + K_{N,i}).$$

This suggests using the following estimator of the variance of the re-scaled estimator:

$$\begin{aligned} \widehat{\sigma}^2_{\text{adj}} &= \widehat{\sigma}^2 + \frac{1}{N} \sum_{t=1}^{T} \left(\sum_{i=1}^{N} 1_{\{X_i \in \mathcal{C}_t\}} W_i (K_{N,i}^2 + K_{N,i}) \right) \widehat{\sigma}_t^2 \\ &= \widehat{\sigma}^2 + \sum_{t=1}^{T} \left(\frac{N_t}{N} \right) \left(\frac{1}{N_t} \sum_{i=1}^{N} 1_{\{X_i \in \mathcal{C}_t\}} W_i (K_{N,i}^2 + K_{N,i}) \right) \widehat{\sigma}_t^2. \end{aligned} \tag{10}$$

where $\widehat{\sigma}_t^2$ is the sample variance of Y calculated from the complete observations in cell $\mathcal{C}_t$. Notice that this formula applies no matter how imputation is done within the cells (for example, randomized or based on the order of the observations in the sample) as long as equation (7) holds.

VI. Monte Carlo Analysis

This section reports the results of two Monte Carlo simulations based on actual data. Section VI.A uses the Boston HMDA data set, a data set collected by the Federal Reserve Bank of Boston to investigate racial discrimination in mortgage credit markets, to assess the quality of the large sample approximation to the distribution of matching estimators derived in section IV. Section VI.B uses CPS data to investigate the performance of the standard error correction for missing data imputation derived in section V.

A. Matching without Replacement in the Boston HMDA Dataset

In order to detect potential discriminatory practices of mortgage credit lenders against minority applicants, the U.S. Home Mortgage Disclosure Act (HMDA) of 1975 requires lenders to routinely disclose information on mortgage applications, including the race and ethnicity of the applicants. The information collected under the HMDA does not include, however, data on the credit histories of the applicants, and other loan and applicant characteristics that are considered to be important factors in determining the approval or denial of mortgage loans. The absence of such information has generated some skepticism about whether the HMDA data can effectively be used to detect discrimination in the mortgage credit market. To overcome this criticism, the Federal Reserve Bank of Boston collected an additional set of 38 variables included in mortgage applications for a sample of applications in the Boston metropolitan area in 1990. The Boston HMDA data set includes all mortgage applications by black and Hispanic applicants in the Boston metropolitan area in 1990, as well as a random sample of mortgage applications by white applicants in the same year and geographical area. Regression analysis of the Boston HMDA data indicated that minority applicants were more likely to be denied mortgage than white applicants with the same

characteristics (Munnell et al., 1996).

In this section, we use the Boston HMDA data set to evaluate the empirical performance of the large sample approximation to the distribution of matching estimators derived in section IV. The HMDA data provides a relevant context for this evaluation because the Federal Reserve System employs matching in the HMDA data as an screening device for fair lending regulation compliance (Avery, Beeson, and Calem, 1997, Avery, Canner, and Cook, 2005). We restrict our sample to single-family residences and male applicants who are white non-Hispanic or black non-Hispanic, not self-employed, who were approved for private mortgage insurance, and who do not have a public record of default or bankruptcy at the time of the application. This leaves us with a sample of 148 black applicants and 1336 white applicants, for a total of 1484 applicants.

In the context of this application, the outcome variable, Y, is an indicator variable that takes value one if the mortgage application was denied, and zero if the mortgage application was approved, W is a binary indicator that takes value one for black applicants, and X is a vector of six applicant and loan characteristics used in Munnell et al. (1996): housing expense to income ratio, total debt payments to income ratio, consumer credit history, mortgage credit history, regional unemployment rate in the applicant's industry, and loan amount to appraised value ratio (see Munnell et al., 1996, for a precise definition of these variables).

To run our simulations for samples sizes of N_1 black observations and N_0 white observations we proceed in five steps. First, for the entire sample, we estimate a logistic model of the mortgage denial indicator on the black indicator and the covariates in X. Second, we draw (with replacement) N_1 observations from the empirical distribution of X for black applicants and N_0 observations from the empirical distribution of X for white applicants. Third, for each individual in the simulated sample, we generate the mortgage denial indicator, Y, using the logistic model estimated in the first step. Fourth, for the simulated sample, we compute $\widehat{\tau}$, the matching estimator in equation (2), matching without replacement, the bias-corrected version of this estimator, $\widehat{\tau}_{\text{bc}}$, in equation (4), and the variance estimator, $\widehat{\sigma}^2$,

in equation (5). All covariates are normalized to have unit variance prior to matching, and a logistic model is employed to calculate the bias correction. Finally, we repeat steps two to four for a total number of 10000 simulations. That is, in this simulation we sample from a population distribution of the covariates that is equal to the distribution of the covariates in the HMDA sample of 1484 applicants. The distribution of Y conditional W and X in our simulation is given by a logistic model with parameters equal to those estimated in the HMDA sample of 1484 applicants. In this Monte Carlo design, the parameter τ in equation (1) is equal to 0.099, which represents the difference in the probability of denial between black applicants and white applicants of the same characteristics in our simulation.

Table I reports the results of the simulation, for different sample sizes, N_1 and N_0. Column (1) reports the bias of $\widehat{\tau}$ relative to τ. As suggested by the results in section IV, our simulation results indicate that for a fixed N_1 the bias of $\widehat{\tau}$ decreases when N_0 increases. For small samples, however, the bias of $\widehat{\tau}$ may be substantial, reflecting the high dimensionality of the vector of matching variables. The bias-corrected estimator in column (2) generates much smaller biases. Columns (3) and (4) report the variance of $\widehat{\tau}$ across simulations and the average, also across simulations, of the variance estimator of $\widehat{\tau}$ in equation (5). Even in fairly small samples ($N_1 = 25$ and $N_0 = 250$), $\widehat{\sigma}^2/N_1$ provides a very precise approximation to the variance of $\widehat{\tau}$. Finally, columns (5) and (6) report coverage rates of nominal 95% confidence intervals constructed with $(\widehat{\tau}, \widehat{\sigma}^2)$ and $(\widehat{\tau}_{\text{bc}}, \widehat{\sigma}^2)$, respectively. The results indicate that, in this simulation, the Normal approximation to the distribution of matching estimators derived in section IV is very accurate, especially when the bias of the matching estimator is corrected using the bias correction techniques in Rubin (1973b), Quade (1982), and Abadie and Imbens (2009).

B. Hot Deck Imputation in the Current Population Survey

Hot deck methods have long been used to impute missing data in large surveys (see, for example, Andridge and Little, 2010). However, the sampling properties of complex hot deck imputation methods, like the sequential hot deck used by the Census Bureau in the CPS, are largely unknown. This void in the literature has become an object of serious concern

in recent years, because the proportion of observations in the CPS with imputed values of weekly earnings has increased steadily: from around 16 percent in 1979, when weekly earnings were included in the monthly survey questionnaire, to more than 30 percent in recent years (Hirsch and Schumacher, 2004; Bollinger and Hirsch, 2009).

In this section we investigate the performance of the approximation to the distribution of a sample mean proposed in section V, when data are imputed using a sequential hot deck like in the CPS. In order to make our exercise as realistic as possible we base our Monte Carlo design on actual CPS data.

Hot deck imputation in the CPS Outgoing Rotation Groups is done through a series of steps, each one imputing a specific survey item. Here, we focus on imputation of missing earnings, because earnings are affected by imputation rates that are much higher than for other survey items. As for other missing survey items, imputation of weekly earnings for non-hourly workers is implemented through a cell hot deck procedure. Observations are assigned to cells defined by age, race, gender, education, occupation, hours worked, and receipt of overtime wages, tips, or commissions, for a total of 11,520 cells (see Bollinger and Hirsch, 2006, for details). Then each missing record is imputed using the value of weekly earnings of last complete record in the same cell.

The imputation of weekly earnings in the CPS Outgoing Rotation Groups cannot be perfectly reproduced with the CPS public use data files. The main reason is that the race variable used by the imputation algorithm is different from the one included in the public use data release. Nevertheless, the Monte Carlo exercise carried out in this section is designed to reproduce as closely as possible the imputation algorithm used by the Census Bureau for weekly earnings. In our simulation we use data from the CPS monthly file of August 2009. In order to simplify the analysis, we first restrict our sample to male individuals working for a pay, who are white, aged 25 to 64, have a high school diploma or equivalent, hold one job only, have a tertiary occupation, do not receive overtime wages, tips, or commissions, and work 40 hours/week. In addition, we discard four observations with zero recorded weekly earnings. This leaves us with 856 observations in 30 of the 11,520

original hot deck cells. The 30 hot deck cells are defined by three categories of age, two of education, and five of occupation. The average number of observations per cell is 28.53, the minimum is 2, and the maximum is 149. In this sample the percentage of observations with missing weekly earnings is 32.83, and each cell has at least two complete observations.

For a fixed number of observations, N, the simulation proceeds as follows. First, for each cell t we simulate two observations of log weekly earnings, $Y^*_{t,1}$ and $Y^*_{t,2}$, from a normal distribution with the same mean and variance as in the distribution of log weekly earnings for complete the CPS observations in the same cell. In our simulation, $Y^*_{t,1}$ and $Y^*_{t,2}$ represent the last two complete observations in cell t in previous CPS waves. Second, we sample N observation from the multinomial distribution of cell frequencies in the CPS sample. For each of these N observations, we simulate log weekly earnings using a normal distribution with the same mean and variance as log weekly earnings for complete CPS observations in the same cell. Then, for each observation we mark weekly earnings as unrecorded with probability equal to the proportion of missing weekly earnings in the same cell of the CPS sample. Third, in our simulated sample of N observations, we impute missing log weekly earnings using the last complete observation in the cell (which may possibly be $Y^*_{t,2}$). This creates a partially imputed sample with N values of log weekly earnings. Four, we calculate the sample average, $\bar{Y}$ in equation (6), as well as the usual and adjusted variance estimators: $\widehat{\sigma}^2$ and $\widehat{\sigma}^2_{\text{adj}}$ in equations (9) and (10), respectively. To compute the intra-cell variances, $\widehat{\sigma}^2_t$ of equation (10), we use all the complete simulated observations in the cell plus $Y^*_{t,1}$ and $Y^*_{t,2}$. Simulating two complete observations per cell, $Y^*_{t,1}$ and $Y^*_{t,2}$, that correspond to the last two complete observations in the cell in previous CPS waves allows us to compute $\widehat{\sigma}^2_t$ even for cells with no other complete observations in the simulation. Finally, we repeat steps one to four for a total number of 50000 simulations.

The results are reported on Table II for sample sizes 50, 100, 200, and 856, the actual number of observations in the CPS sample. The average of our adjusted variance estimator across simulations, in column (2), closely approximates the variance of $\bar{Y}$, in column (1), even for fairly small sample sizes. In contrast, columns (3) and (4) show that the usual

variance estimator is severely downward biased, and that the bias of this estimator (as a percentage of the true variance) increases with the sample size. For 856 observations, that is the actual size of the CPS data sample used in the simulation, the usual variance estimator is only 58 percent of the true variance of $\bar{Y}$. Large sample sizes make possible that some observations are repeatedly used for imputation, increasing the difference between the adjusted and unadjusted variances in equation (10). This happens when missing observations arrive consecutively to a cell, without the observation used for imputation being "refreshed" by another complete observation. Columns (5) and (6) report coverage rates of nominal 95% confidence intervals constructed with $\widehat{\sigma}^2_{\text{adj}}$ and $\widehat{\sigma}^2$, respectively. The results show coverage rates close to nominal coverage in column (5), when the adjusted variance estimator is used to construct confidence interval. In contrast, confidence intervals calculated with the usual variance estimator suffer from severe under-coverage, as reported in column (6).

VII. Conclusion

This article establishes a martingale array representation for matching estimators. This representation allows the use of well-known martingale limit theorems to determine the large sample distribution of matching estimators. Because the martingale representation applies to a large class of matching estimators, the applicability of the methods presented in this article is very broad. Specific applications include matching estimators of average treatment effects as well as "hot deck" imputation methods for missing data. Two realistic simulations demonstrate the empirical relevance of the results of this article.

Appendix

Proposition 1: *Let F_0 and F_1 be the distributions of X given $W = 0$ and X given $W = 1$, respectively. Assume that F_0 and F_1 have a common support that is a Cartesian product of intervals, and that the densities $f_0(x)$ and $f_1(x)$ are bounded and bounded away from zero: $\underline{f} \leq f_0 \leq \bar{f}$ and $\underline{f} \leq f_1 \leq \bar{f}$. Assume that there exists $c > 0$ and $r > k$ where k is the number of (continuous) covariates, such that $N_1^r/N_0 \leq c$. Then,*

$$\frac{1}{\sqrt{N_1}} \sum_{i=1}^{N} \sum_{m=1}^{M} W_i \| U_{N_0,N_1,i}^{(M,m)} \| \xrightarrow{p} 0.$$

Proof of Proposition 1: By changing units of measurement, we can always make the support of the covariates equal to the unit k-cube. (This only adds a multiplicative constant to our bounds.) Notice that we can always divide a unit k-cube into N_1^k identical cubes, for $N_1 = 1, 2, 3, \ldots$.

Divide the support of F_0 and F_1 into N_1^k identical cubes. Let Z_{M,N_0,N_1} be the number of such cells where the number of untreated observation is less than M times the number of observations from the treated sample. Let M_{N_1} be the maximum number of observations from the treated sample in a single cell. Let m_{N_0,N_1} be the minimum number of untreated observations in a single cell. Notice that for any series, $f(N_1)$, such that $1 \leq f(N_1) < N_1$, we have:

$$\begin{aligned}
\Pr(Z_{M,N_0,N_1} > 0) &\leq \sum_{n=1}^{N_1} \Pr(m_{N_0,N_1} < Mn) \Pr(M_{N_1} = n) \\
&\leq \sum_{n=1}^{\lfloor f(N_1) \rfloor} \Pr(m_{N_0,N_1} < Mn) \Pr(M_{N_1} = n) \\
&+ \sum_{n=\lfloor f(N_1) \rfloor + 1}^{N_1} \Pr(m_{N_0,N_1} < Mn) \Pr(M_{N_1} = n) \\
&\leq f(N_1) \Pr(m_{N_0,N_1} < M f(N_1)) \\
&+ (N_1 - f(N_1)) \Pr(M_{N_1} > f(N_1)).
\end{aligned}$$

Let $D_{N_1,n}$ be the number of cells where the number of treated observations is larger than n. Let $0 < \alpha < \min\{r - k, 1\}$. Consider $f(N_1) = N_1^\alpha$. For N_1 large enough, $\bar{f}/N_1^k < 1$. Using Bonferroni Inequality we obtain for N_1 large enough:

$$\begin{aligned}
\Pr(M_{N_1} > f(N_1)) &= \Pr(D_{N_1,N_1^\alpha} \geq 1) \\
&\leq N_1^k \Pr\left(B(N_1, \bar{f}/N_1^k) > N_1^\alpha\right),
\end{aligned}$$

where $B(N, p)$ denotes a Binomial random variable with parameters (N, p). Using Bennett's bound for binomial tails (e.g., Shorack and Wellner, 1996, p. 440), we obtain:

$$\Pr\left(B(N_1, \bar{f}/N_1^k) > N_1^\alpha\right) = \Pr\left(\frac{B(N_1, \bar{f}/N_1^k) - \bar{f}/N_1^{k-1}}{\sqrt{N_1}} > \frac{N_1^\alpha - \bar{f}/N_1^{k-1}}{\sqrt{N_1}}\right)$$

$$\leq \exp\left\{-\frac{\bar{f}/N_1^{k-1}}{1-\bar{f}/N_1^k}\left[\frac{N_1^{\alpha+k-1}}{\bar{f}}\left(\log\left(\frac{N_1^{\alpha+k-1}}{\bar{f}}\right)-1\right)+1\right]\right\}$$
$$= \exp\left\{-\frac{1}{1-\bar{f}/N_1^k}\left[N_1^{\alpha}\left(\log\left(\frac{N_1^{\alpha+k-1}}{\bar{f}}\right)-1\right)+\frac{\bar{f}}{N_1^{k-1}}\right]\right\}.$$

Similarly, let $C_{N_0,N_1,m}$ be the number of cells with less than m untreated observations. Then, using Bonferroni Inequality:

$$\begin{aligned}\Pr(m_{N_0,N_1} < m) &= \Pr(C_{N_0,N_1,m} \geq 1) \\ &\leq \sum_{n=1}^{N_1^k} \Pr\big(B(N_0,p_n) < m\big),\end{aligned}$$

where p_n is the probability that an untreated observation falls in cell n. Then, because for all n, $p_n \geq \underline{f}/N_1^k$, we obtain:

$$\Pr(m_{N_0,N_1} < m) \leq N_1^k \Pr\big(B(N_0,\underline{f}/N_1^k) < m\big).$$

Also, for large enough N_1, there exists δ such that $(Mc/\underline{f})/N_1^{r-\alpha-k} < \delta < 1$. Using Chernoff's bound for the lower tail of a sum of independent Poisson trials (e.g., Motwani and Raghavan, 1995, p. 70), we obtain that for large enough N_1:

$$\begin{aligned}\Pr\big(B(N_0,\underline{f}/N_1^k) < MN_1^{\alpha}\big) &= \Pr\left(B(N_0,\underline{f}/N_1^k) < \underline{f}\frac{N_0}{N_1^k}\frac{MN_1^{\alpha+k}}{\underline{f}N_0}\right) \\ &\leq \Pr\left(B(N_0,\underline{f}/N_1^k) < \underline{f}\frac{N_0}{N_1^k}\frac{Mc/\underline{f}}{N_1^{r-\alpha-k}}\right) \\ &\leq \exp\Big(-(\underline{f}N_0/N_1^k)\big(1-(Mc/\underline{f})/N_1^{r-\alpha-k}\big)^2/2\Big) \\ &\leq \exp\Big(-\underline{f}N_1^{r-k}(1-\delta)^2/2c\Big).\end{aligned}$$

This proves an exponential bound for $\Pr(Z_{M,N_0,N_1} > 0)$.

Rearrange the observations so the first N_1 observations in the sample are the treated observations. For $1 \leq i \leq N_1$, let $\|U_{N_0,N_1,i}^{(M,m)}\|$ be the m-th matching discrepancy for treated unit i when untreated units are matched without replacement to treated units in such a way that the sum of the matching discrepancies is minimized. For $1 \leq i \leq N_1$, let $\|V_{N_0,N_1,i}^{(M,m)}\|$ be the m-th matching discrepancy for treated unit i when untreated units are matched without replacement to treated units in such a way that the matches are first done within cells and, after all possible within-cell matches are exhausted, untreated units that were not previously used as a match are matched without replacement to previously unmatched treated units in other cells. Notice that:

$$\sum_{i=1}^{N_1}\sum_{m=1}^{M}\|U_{N_0,N_1,i}^{(M,m)}\| \leq \sum_{i=1}^{N_1}\sum_{m=1}^{M}\|V_{N_0,N_1,i}^{(M,m)}\|.$$

Let $d_{N_1,k}$ be the diameter of the cells. Let C_k be the diameter of the unit k-cube. Notice that if the unit k-cube is divided in N_1^k identical cells, then $C_k = N_1 d_{N_1,k}$. For $1 \leq n \leq N_1^k$, let $A_{N_1,n}$

be the n-th cell. Then,

$$\begin{aligned} E\left[\|V^{(M,m)}_{N_0,N_1,i}\| \,\middle|\, Z_{M,N_0,N_1}=0\right] &\leq \sum_{n=1}^{N_1^k} d_{N_1,k}\Pr(X_{1,i}\in A_{N_1,n}|Z_{N_0,N_1}=0) \\ &\leq d_{N_1,k} \\ &= \frac{C_k}{N_1}. \end{aligned}$$

Now,

$$\begin{aligned} E\left[\frac{1}{\sqrt{N_1}}\sum_{i=1}^{N_1}\sum_{m=1}^{M}\|U^{(M,m)}_{N_0,N_1,i}\|\right] &\leq E\left[\frac{1}{\sqrt{N_1}}\sum_{i=1}^{N_1}\sum_{m=1}^{M}\|V^{(M,m)}_{N_0,N_1,i}\|\right] \\ &= E\left[\frac{1}{\sqrt{N_1}}\sum_{i=1}^{N_1}\sum_{m=1}^{M}\|V^{(M,m)}_{N_0,N_1,i}\|\Big| Z_{M,N_0,N_1}=0\right]\Pr(Z_{M,N_0,N_1}=0) \\ &+ E\left[\frac{1}{\sqrt{N_1}}\sum_{i=1}^{N_1}\sum_{m=1}^{M}\|V^{(M,m)}_{N_0,N_1,i}\|\Big| Z_{M,N_0,N_1}>0\right]\Pr(Z_{M,N_0,N_1}>0) \\ &\leq M\frac{C_k}{\sqrt{N_1}}+\sqrt{N_1}MC_k\Pr(Z_{M,N_0,N_1}>0)\longrightarrow 0. \end{aligned}$$

Markov's Inequality produces the desired result. □

PROOF OF THEOREM 1: Notice that condition *(iii)* in Theorem 1 implies that for $w=0,1$, $\mu_w(x)$ and $\sigma^2_w(x)$ are bounded on the support of X given $W=w$. Then, the result of the theorem follows easily from the derivations in section IV. □

Before proving Theorem 2 it is useful to prove the following proposition.

PROPOSITION 2: *Let*

$$A_{N,t}=\frac{1}{N}\sum_{i=1}^{N}1_{\{X_i\in\mathcal{C}_t\}}W_i(1+K_{N,i})^2-E\left[\frac{1}{N}\sum_{i=1}^{N}1_{\{X_i\in\mathcal{C}_t\}}W_i(1+K_{N,i})^2\right].$$

Under the conditions of Theorem 2, we have $A_{N,t}\xrightarrow{p}0$, *for all* $t=1,2,\ldots,T$.

PROOF OF PROPOSITION 2: Given the nature of the sequential hot-deck, it is easy to check that for any N and i the positive moments of $K_{N,i}$ conditional on $X_i\in\mathcal{C}_t$ are bounded by the corresponding moments of a Geometric distribution with parameter $\Pr(W=1|X\in\mathcal{C}_t)$. Therefore, we obtain that for any $r>0$ there exists a constant c_r such that $E[K^r_{N,i}]\leq c_r$ for all N and i.

Because $E[A_{N,t}]=0$, Markov's inequality implies that if $\text{var}(A_{N,t})\to 0$, then $A_{N,t}\xrightarrow{p}0$.

$$\begin{aligned} \text{var}(A_{N,t}) &= \text{var}\left(\frac{1}{N}\sum_{i=1}^{N}1_{\{X_i\in\mathcal{C}_t\}}W_i(1+K_{N,i})^2\right) \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\text{var}\left(1_{\{X_i\in\mathcal{C}_t\}}W_i(1+K_{N,i})^2\right) \end{aligned}$$

$$+ \frac{2}{N^2} \sum_{i=1}^{N} \sum_{j>i} \text{cov}\left(1_{\{X_i \in \mathcal{C}_t\}} W_i (1 + K_{N,i})^2, 1_{\{X_j \in \mathcal{C}_t\}} W_j (1 + K_{N,j})^2\right).$$

To show that $\text{var}(A_{N,t})$ converges to zero, we will first prove the following intermediate result: for all $i = 1, \ldots, N-1$, all $j = i+1, \ldots, N$, and all $p \geq 0$, $\Pr(1_{\{X_j \in \mathcal{C}_t\}} W_j = 1 \,|\, 1_{\{X_i \in \mathcal{C}_t\}} W_i (1 + K_{N,i}) \leq p) \geq \Pr(1_{\{X_j \in \mathcal{C}_t\}} W_j = 1)$. To prove this result notice that

$$\Pr((1 + K_{N,i}) > p | W_i = 1, X_i \in \mathcal{C}_t) = \Pr(W = 0 | X \in \mathcal{C}_t)^p \Pr\left(\sum_{k=i+1}^{N} 1_{\{X_k \in \mathcal{C}_t\}} \geq p\right).$$

Therefore,

$$\begin{aligned}\Pr(1_{\{X_i \in \mathcal{C}_t\}} W_i (1 + K_{N,i}) > p) = \Pr(W = 0 | X \in \mathcal{C}_t)^p \Pr\left(\sum_{k=i+1}^{N} 1_{\{X_k \in \mathcal{C}_t\}} \geq p\right) \\ \times \Pr(W_i = 1 | X_i \in \mathcal{C}_t) \Pr(X_i \in \mathcal{C}_t).\end{aligned}$$

Similarly,

$$\begin{aligned}\Pr(1_{\{X_i \in \mathcal{C}_t\}} W_i (1 + K_{N,i}) > p \,|\, 1_{\{X_j \in \mathcal{C}_t\}} W_j = 1) = \Pr(W = 0 | X \in \mathcal{C}_t)^p \Pr\left(\sum_{k=i+1}^{j-1} 1_{\{X_k \in \mathcal{C}_t\}} \geq p\right) \\ \times \Pr(W_i = 1 | X_i \in \mathcal{C}_t) \Pr(X_i \in \mathcal{C}_t).\end{aligned}$$

Now, because

$$\Pr\left(\sum_{k=i+1}^{j-1} 1_{\{X_k \in \mathcal{C}_t\}} \geq p\right) \leq \Pr\left(\sum_{k=i+1}^{N} 1_{\{X_k \in \mathcal{C}_t\}} \geq p\right),$$

we obtain that

$$\Pr(1_{\{X_i \in \mathcal{C}_t\}} W_i (1 + K_{N,i}) > p \,|\, 1_{\{X_j \in \mathcal{C}_t\}} W_j = 1) \leq \Pr(1_{\{X_i \in \mathcal{C}_t\}} W_i (1 + K_{N,i}) > p),$$

or equivalently,

$$\Pr(1_{\{X_i \in \mathcal{C}_t\}} W_i (1 + K_{N,i}) \leq p \,|\, 1_{\{X_j \in \mathcal{C}_t\}} W_j = 1) \geq \Pr(1_{\{X_i \in \mathcal{C}_t\}} W_i (1 + K_{N,i}) \leq p).$$

By Bayes' theorem,

$$\frac{\Pr(1_{\{X_j \in \mathcal{C}_t\}} W_j = 1 \,|\, 1_{\{X_i \in \mathcal{C}_t\}} W_i (1 + K_{N,i}) \leq p)}{\Pr(1_{\{X_j \in \mathcal{C}_t\}} W_j = 1)} = \frac{\Pr(1_{\{X_i \in \mathcal{C}_t\}} W_i (1 + K_{N,i}) \leq p \,|\, 1_{\{X_j \in \mathcal{C}_t\}} W_j = 1)}{\Pr(1_{\{X_i \in \mathcal{C}_t\}} W_i (1 + K_{N,i}) \leq p)}$$

and we therefore obtain the desired result,

$$\Pr(1_{\{X_j \in \mathcal{C}_t\}} W_j = 1 \,|\, 1_{\{X_i \in \mathcal{C}_t\}} W_i (1 + K_{N,i}) \leq p) \geq \Pr(1_{\{X_j \in \mathcal{C}_t\}} W_j = 1). \tag{11}$$

We will now show that, for all $i = 1, \ldots, N-1$ and all $j = i+1, \ldots, N$, $\text{cov}(1_{\{X_i \in \mathcal{C}_t\}} W_i (1 + K_{N,i})^2, 1_{\{X_j \in \mathcal{C}_t\}} W_j (1 + K_{N,j})^2) \leq 0$. Consider two units i and j, with $j > i$. Notice that because

of the sequential nature of hot-deck imputation, $K_{N,j}$ is independent of $(W_i, K_{N,i})$ conditional on W_j. Therefore:

$$\begin{aligned}
&\Pr(1_{\{X_j\in\mathcal{C}_t\}}W_j(1+K_{N,j}) \le q \,|\, 1_{\{X_i\in\mathcal{C}_t\}}W_i(1+K_{N,i}) \le p) \\
&\quad = \Pr(1_{\{X_j\in\mathcal{C}_t\}}W_j(1+K_{N,j}) \le q \,|\, 1_{\{X_j\in\mathcal{C}_t\}}W_j = 1, 1_{\{X_i\in\mathcal{C}_t\}}W_i(1+K_{N,i}) \le p) \\
&\qquad \times \Pr(1_{\{X_j\in\mathcal{C}_t\}}W_j = 1 \,|\, 1_{\{X_i\in\mathcal{C}_t\}}W_i(1+K_{N,i}) \le p) \\
&\qquad + \Pr(1_{\{X_j\in\mathcal{C}_t\}}W_j(1+K_{N,j}) \le q \,|\, 1_{\{X_j\in\mathcal{C}_t\}}W_j = 0, 1_{\{X_i\in\mathcal{C}_t\}}W_i(1+K_{N,i}) \le p) \\
&\qquad \times \Pr(1_{\{X_j\in\mathcal{C}_t\}}W_j = 0 \,|\, 1_{\{X_i\in\mathcal{C}_t\}}W_i(1+K_{N,i}) \le p) \\
&\quad = \Pr(1_{\{X_j\in\mathcal{C}_t\}}W_j(1+K_{N,j}) \le q \,|\, 1_{\{X_j\in\mathcal{C}_t\}}W_j = 1) \\
&\qquad \times \Pr(1_{\{X_j\in\mathcal{C}_t\}}W_j = 1 \,|\, 1_{\{X_i\in\mathcal{C}_t\}}W_i(1+K_{N,i}) \le p) \\
&\qquad + \Pr(1_{\{X_j\in\mathcal{C}_t\}}W_j = 0 \,|\, 1_{\{X_i\in\mathcal{C}_t\}}W_i(1+K_{N,i}) \le p) \\
&\quad = 1 - \Big(1 - \Pr(1_{\{X_j\in\mathcal{C}_t\}}W_j(1+K_{N,j}) \le q \,|\, 1_{\{X_j\in\mathcal{C}_t\}}W_j = 1)\Big) \\
&\qquad \times \Pr(1_{\{X_j\in\mathcal{C}_t\}}W_j = 1 \,|\, 1_{\{X_i\in\mathcal{C}_t\}}W_i(1+K_{N,i}) \le p).
\end{aligned}$$

Now, because $\Pr(1_{\{X_j\in\mathcal{C}_t\}}W_j = 1 \,|\, 1_{\{X_i\in\mathcal{C}_t\}}W_i(1+K_{N,i}) \le p) \ge \Pr(1_{\{X_j\in\mathcal{C}_t\}}W_j = 1)$ (equation (11)), we obtain:

$$\begin{aligned}
&\Pr(1_{\{X_j\in\mathcal{C}_t\}}W_j(1+K_{N,j}) \le q \,|\, 1_{\{X_i\in\mathcal{C}_t\}}W_i(1+K_{N,i}) \le p) \\
&\quad \le 1 - (1 - \Pr(1_{\{X_j\in\mathcal{C}_t\}}W_j(1+K_{N,j}) \le q \,|\, 1_{\{X_j\in\mathcal{C}_t\}}W_j = 1)\Pr(1_{\{X_j\in\mathcal{C}_t\}}W_j = 1) \\
&\quad = \Pr(1_{\{X_j\in\mathcal{C}_t\}}W_j(1+K_{N,j}) \le q).
\end{aligned}$$

As a result, the varibles $1_{\{X_j\in\mathcal{C}_t\}}W_j(1+K_{N,j})$ and $1_{\{X_j\in\mathcal{C}_t\}}W_j(1+K_{N,j})$ are negative quadrant dependent and, therefore, negatively associated (Joag-Dev and Proschan, 1983). Furthermore, because increasing transformations of negatively associated random variables are also negatively associated (Joag-Dev and Proschan, 1983), we obtain:

$$\operatorname{cov}(1_{\{X_i\in\mathcal{C}_t\}}W_i(1+K_{N,i})^2, 1_{\{X_j\in\mathcal{C}_t\}}W_j(1+K_{N,j})^2) \le 0,$$

for all $i = 1, \ldots, N$ and all $j = i+1, \ldots, N$. This result implies

$$\operatorname{var}(A_{N,t}) \le \frac{1}{N^2}\sum_{i=1}^{N} \operatorname{var}\left(1_{\{X_i\in\mathcal{C}_t\}}W_i(1+K_{N,i})^2\right). \tag{12}$$

To finish the proof, we will show that $\operatorname{var}(1_{\{X_i\in\mathcal{C}_t\}}W_i(1+K_{N,i})^2)$ is uniformly bounded in (i, N). Because

$$\begin{aligned}
\operatorname{var}\left(1_{\{X_i\in\mathcal{C}_t\}}W_i(1+K_{N,i})^2\right) &\le E\left[1_{\{X_i\in\mathcal{C}_t\}}W_i(1+K_{N,i})^4\right] \\
&= E\left[(1+K_{N,i})^4 \,|\, 1_{\{X_i\in\mathcal{C}_t\}}W_i = 1\right]\Pr(1_{\{X_i\in\mathcal{C}_t\}}W_i = 1),
\end{aligned}$$

and because $E[K_{N,i}^4 | 1_{\{X_i\in\mathcal{C}_t\}}W_i = 1]$ is uniformly bounded in (i, N), we obtain $\operatorname{var}(A_{N,t}) \to 0$. □

Proof of Theorem 2: First, notice that, because $(1+K_{N,i})^2 \ge (1+K_{N,i})$ and $\sum_{i=1}^{N} 1_{\{X_i\in\mathcal{C}_t\}}W_i(1+K_{N,i}) = N_t$, we obtain:

$$\sigma^2 \ge E\left[\sum_{t=1}^{T}\left(\frac{N_t}{N}\right)(\mu_t - \mu)^2\right] + E\left[\sum_{t=1}^{T}\sigma_t^2\frac{1}{N}\sum_{i=1}^{N}1_{\{X_i\in\mathcal{C}_t\}}W_i(1+K_{N,i})\right]$$

$$= E\left[\sum_{t=1}^{T}\left(\frac{N_t}{N}\right)(\mu_t-\mu)^2\right] + E\left[\sum_{t=1}^{T}\left(\frac{N_t}{N}\right)\sigma_t^2\right] = \mathrm{var}(Y) > 0,$$

and the sequence $\{\xi_{N,k}\}_{k=1}^{2N}$ is well-defined. Now, applying Proposition 2 we obtain:

$$\begin{aligned}\sum_{k=1}^{2N} E[\xi_{N,k}^2|\mathcal{F}_{N,k-1}] &= \frac{1}{\sigma^2 N}\sum_{k=1}^{N} E[(\mu(X_k)-\mu)^2] \\ &+ \frac{1}{\sigma^2 N}\sum_{k=N+1}^{2N}\sum_{t=1}^{T} 1_{\{X_{k-N}\in\mathcal{C}_t\}} W_{k-N}(1+K_{N,k-N})^2\sigma_t^2 \\ &= \frac{1}{\sigma^2}E\left[\frac{1}{N}\sum_{k=1}^{N}\sum_{t=1}^{T} 1_{\{X_k\in\mathcal{C}_t\}}(\mu_t-\mu)^2\right] \\ &+ \frac{1}{\sigma^2}\sum_{t=1}^{T}\sigma_t^2\frac{1}{N}\sum_{k=1}^{N} 1_{\{X_k\in\mathcal{C}_t\}} W_k(1+K_{N,k})^2 \xrightarrow{p} 1.\end{aligned}$$

Jensen's inequality implies: $E[|\mu(X_i)|^{2+\delta}] \le E[|Y_i|^{2+\delta}] < \infty$. Because $E[|Y_i-\mu(X_i)|^{2+\delta}] < \infty$ and because all positive moments of $K_{N,i}$ are bounded (uniformly in N and i), Holder's Inequality implies that $E[W_i(1+K_{N,i})^{2+\delta/2}|Y_i-\mu(X_i)|^{2+\delta/2}]$ is bounded (uniformly in N and i). As a result, we obtain the Lyapunov condition:

$$\sum_{k=1}^{2N} E[\xi_{N,k}^{2+\delta/2}] \to 0.$$

The result of Theorem 2 follows now from Theorem 35.12 in Billingsley (1995). □

References

Abadie, A. and Imbens, G.W. (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica* , vol. 74, no. 1, 235-267.

Abadie, A. and Imbens, G.W. (2008), "On the Failure of the Bootstrap for Matching Estimators," *Econometrica* , vol. 76, no. 6, 1537-1558.

Abadie, A. and Imbens, G.W. (2009), "Bias Corrected Matching Estimators for Average Treatment Effects," *Journal of Business and Economic Statistics* (forthcoming).

Abadie, A. and Imbens, G.W. (2010), "Matching on the Estimated Propensity Score," mimeo.

Andridge, R.R. and Little, R.J.A. (2010), "A Review of Hot Deck Imputation for Survey Non-response," *International Statistical Review* (forthcoming).

Avery, R.B., Beeson, P.E., and Calem, P.S. (1997), "Using HMDA Data as a Regulatory Screen for Fair Lending Compliance," *Journal of Financial Services Research*, vol. 11, 9-42.

Avery, R.B, Canner, G.B., and Cook, R.E. (2005), "New Information Reported Under HMDA and Its Application in Fair Lending Enforcement," *Federal Reserve Bulletin*, vol. 91, 344-394.

Billingsley, P. (1995), *Probability and Measure*, third edition. Wiley, New York.

Bollinger, C.R. and Hirsch, B.T. (2009), "Wage Gap Estimation with Proxies and Nonresponse," mimeo.

Dehejia, R. and Wahba, S. (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053-1062.

Diamond, A. and Sekhon, J.S. (2008), "Genetic Matching for Estimating Causal Effects: A New Method of Achieving Balance in Observational Studies," UC Berkeley.

Gu, X.S. and Rosenbaum, P.R. (1993), "Comparison of Multivariate Matching Methods: Structures, Distances and Algorithms," *Journal of Computational and Graphical Statistics*, 2, 405-420.

Hall, P. and Heyde C.C. (1980), *Martingale Limit Theory and its Applications.* Academic Press, New York.

Hansen, B.B. (2004), "Full Matching in an Observational Study of Coaching for the SAT," *Journal of the American Statistical Association*, 99, 609-618.

Heckman, J., Ichimura, H., and Todd, P. (1998), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, vol. 65, 261-294.

Hirsch, B.T. and Schumacher, E.J. (2004), "Match Bias in Wage Gap Estimates Due to Earnings Imputation," *Journal of Labor Economics*, vol. 22, no. 3, 689-722.

Iacus, S.M., King, G., and Porro, G. (2009),"Causal Inference Without Balance Checking: Coarsened Exact Matching," mimeo.

Imbens, G.W. (2004), "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *Review of Economics and Statistics* , vol. 86, no. 1, 4-29.

Joag-Dev, K. and Proschan, F. (1983), "Negative Association of Random Variables with Applications," *Annals of Statistics*, vol. 11, no. 1, 286-295.

Little, R.J.A. and Rubin, D.B. (2002), *Statistical Analysis with Missing Data*, second edition. Wiley-Interscience, New York.

Motwani, R. and Raghavan, P. (1995), *Randomized Algorithms.* Cambridge University Press, New York.

Munnell, A.H., Tootell, G.M.B., Browne, L.E. and McEneaney, J. (1996), "Mortgage Lending in Boston: Interpreting HMDA Data," *American Economic Review*, vol. 86, no. 1, 25-53.

Quade, D. (1982), "Nonparametric Analysis of Covariance by Matching", *Biometrics*, 38, 597-611.

Rosenbaum, P.R. (2002), *Observational Studies*, second edition. Springer, New York.

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

Rubin, D.B. (1973a), "Matching to Reduce Bias in Observational Studies," *Biometrics*, 29, 159-183.

Rubin, D.B. (1973b), "The Use of Matched Sampling and Regression Adjustments to Remove Bias in Observational Studies," *Biometrics*, 29, 185-203.

Rubin, D.B. (1977), "Assignment to Treatment Group on the Basis of a Covariate", *Journal of Educational Statistics*, 2, 1-26.

Rubin, D.B. (2006), *Matched Sampling for Causal Effects.* Cambridge University Press, New York.

Shorack, G.R. (2000), *Probability for Statisticians.* Springer, New York.

Shorack, G.R. and Wellner, J.A. (1986), *Empirical Processes with Applications to Statistics.* Wiley, New York.

Table I – Boston HMDA Data, Simulation Results
Black-White Difference in Mortgage Denial Probability for Matched Pairs
(Number of simulations = 10000)

Sample sizes		Bias		Variance		Coverage of 95% C.I.	
		(1)	(2)	(3)	(4)	(5)	(6)
		$\lvert E[\widehat{\tau}] - \tau \rvert$	$\lvert E[\widehat{\tau}_{\text{bc}}] - \tau \rvert$	$\text{var}(\widehat{\tau})$	$E[\widehat{\sigma}^2/N_1]$	$\widehat{\tau} \pm 1.96\,\widehat{\sigma}/\sqrt{N_1}$	$\widehat{\tau}_{\text{bc}} \pm 1.96\,\widehat{\sigma}/\sqrt{N_1}$
$N_1 = 25$	$N_0 = 250$	0.0143	0.0012	0.0091	0.0091	0.9225	0.9348
	$N_0 = 500$	0.0106	0.0001	0.0092	0.0091	0.9244	0.9394
	$N_0 = 1000$	0.0077	0.0002	0.0090	0.0091	0.9263	0.9430
$N_1 = 50$	$N_0 = 500$	0.0106	0.0011	0.0045	0.0045	0.9427	0.9458
	$N_0 = 1000$	0.0073	0.0009	0.0044	0.0046	0.9427	0.9456
$N_1 = 100$	$N_0 = 1000$	0.0090	0.0001	0.0023	0.0023	0.9436	0.9468

Table II – Current Population Survey Data, Simulation Results
Average Log Weekly Earnings
(Number of simulations = 50000)

Sample size	Variance			Ratio	Coverage of 95% C.I.	
	(1)	(2)	(3)	(4)	(5)	(6)
N	$\text{var}(\bar{Y})$	$E[\widehat{\sigma}^2_{\text{adj}}/N]$	$E[\widehat{\sigma}^2/N]$	(3)/(1)	$\bar{Y} \pm 1.96\,\widehat{\sigma}_{\text{adj}}/\sqrt{N}$	$\bar{Y} \pm 1.96\,\widehat{\sigma}/\sqrt{N}$
50	0.0072	0.0071	0.0052	0.7262	0.9436	0.8973
100	0.0039	0.0039	0.0026	0.6701	0.9476	0.8888
200	0.0021	0.0021	0.0013	0.6342	0.9492	0.8799
856	0.0005	0.0005	0.0003	0.5834	0.9482	0.8661

Matching on the Estimated Propensity Score

Alberto Abadie – Harvard University and NBER
Guido W. Imbens – Harvard University and NBER

June 2011

Abstract

Propensity score matching estimators (Rosenbaum and Rubin, 1983) are widely used in evaluation research to estimate average treatment effects. In this article, we derive the large sample distribution of propensity score matching estimators. Our derivations take into account that the propensity score is itself estimated in a first step, prior to matching. We prove that first step estimation of the propensity score affects the large sample distribution of propensity score matching estimators and derive adjustments to the large sample variances of two common propensity score matching estimators.

Alberto Abadie, John F. Kennedy School of Government, 79 John F. Kennedy Street, Cambridge MA 02138, USA. E-mail: alberto_abadie@harvard.edu, web: http://www.hks.harvard.edu/fs/aabadie/. Guido W. Imbens, Department of Economics, 1830 Cambridge Street, Cambridge MA 02138. E-mail: imbens@harvard.edu, web: http://www.economics.harvard.edu/faculty/imbens/imbens.html.

We are grateful to Ben Hansen, James Robins, Paul Rosenbaum, Donald Rubin, and participants in seminars at the Banff Center, Brown, Georgetown, Harvard/MIT, Montreal, the 2010 NBER Summer Institute, Stanford, UCL, and UPenn for comments and discussions. Financial support by the NSF through grants SES 0820361 and 0961707 is gratefully acknowledged. Software implementing these methods is available from the authors.

I. Introduction

Propensity score matching estimators (Rosenbaum and Rubin, 1983) are widely used to estimate treatment effects when all treatment confounders are measured.[1] Rosenbaum and Rubin (1983) define the propensity score as the conditional probability of assignment to a treatment given a vector of covariates including the values of all treatment confounders. Their key insight is that adjusting for the propensity score is sufficient to remove the bias associated with all treatment confounders. Relative to matching directly on the covariates, propensity score matching has the advantage of reducing the dimensionality of matching to a single dimension. This greatly facilitates the matching process, because units with dissimilar covariate values may nevertheless have similar values for their propensity scores.

In observational studies propensity scores are not known, so they have to be estimated prior to matching. In spite of the great popularity that propensity score matching methods have enjoyed since they were proposed by Rosenbaum and Rubin in 1983, their large sample distribution has not yet been derived for the case when the propensity score is estimated in a first step.[2] A possible reason for this void in the literature is that matching estimators are non-smooth functionals of the distribution of the matching variables, which makes it difficult to establish an asymptotic approximation to the distribution of matching estimators when a matching variable is estimated in a first step. This has motivated the use of bootstrap standard errors for propensity score matching estimators. However, recently it has been shown that the bootstrap is not in general valid for matching estimators (Abadie and Imbens, 2008).[3]

[1] Following the terminology in Abadie and Imbens (2006), the term "matching estimator" is reserved in this article to estimators that match each unit (or each unit of some sample subset, e.g., the treated) to a small number of units with similar characteristics in the opposite treatment arm. Thus our discussion does not refer to regression imputation methods, like the kernel matching method of Heckman, Ichimura and Todd (1998), which use use a large number of matches per unit and nonparametric smoothing techniques to consistently estimate unit-level regression values in the opposite treatment arms. See Heckman, Ichimura and Todd (1998), Imbens (2004), and Imbens and Wooldridge (2009) for a discussion of such estimators.

[2] Influential papers using matching on the estimated propensity score include Heckman, Ichimura, and Todd (1997), Dehejia and Wahba (1999), and Smith and Todd 92005).

[3] In contexts other than matching, Heckman, Ichimura and Todd (1998), Hirano, Imbens and Ridder (2003), Abadie (2005), Wooldridge (2007), and Angrist and Kuersteiner (2009) derive large sample properties of statistics based on a first step estimator of the propensity score. In all these cases, the second step statistics are smooth functionals of the propensity scores and, therefore, standard stochastic expansions

In this article, we derive the large sample approximations to the distribution of propensity score matching estimators. Our derivations take into account that the propensity score is itself estimated in a first step. We show that propensity matching estimators have approximately Normal distributions in large samples. We demonstrate that first step estimation of the propensity score affects the large sample distribution of propensity score matching estimators, and we derive adjustments to the large sample variance of propensity score matching estimators that correct for first step estimation of the propensity score. We do this for estimators of the average treatment effect (ATE) and the average treatment effect on the treated (ATET). The adjustment for the ATE estimator is negative (or zero in some special cases), implying that matching on the estimated propensity score is more efficient than matching on the true propensity score in large samples. As a result, treating the estimated propensity score as it was the true propensity score for estimating the variance of the ATE estimator leads to conservative confidence intervals. However, for the ATET estimator the sign of the adjustment depends on the data generating process, and ignoring the estimation error in the propensity score may lead to confidence intervals that are either too large or too small. We present the results from a small simulation exercise to illustrate the implications of our theoretical results.

To preview the main results, let τ be the average treatment effect, and let $\hat{\tau}_N^*$ be an estimator of τ obtained by matching on the true propensity score. For matching with replacement using the true propensity score as the only matching variables, the results in Abadie and Imbens (2006) imply:

$$\sqrt{N}(\hat{\tau}_N^* - \tau) \xrightarrow{d} N(0, \sigma^2),$$

for some σ^2. The particular form for σ^2 is given in Proposition 1 in Section 2, which weakens some of the regularity conditions from Abadie and Imbens (2006) in a way that is important in the current context. Now let $F(x'\theta)$ be a parametric model for the propensity score, with unknown parameter θ, and let $\hat{\theta}_N$ be the maximum likelihood estimator for θ. We show that, under regularity conditions, the estimator $\hat{\tau}_N$, based on matching with

for two-step estimators apply (see, e.g., Newey and McFadden, 1994).

replacement on the estimated propensity score $F(X_i'\hat{\theta}_N)$, satisfies

$$\sqrt{N}(\hat{\tau}_N - \tau) \xrightarrow{d} N(0, \sigma^2 - c' I_{\theta^*}^{-1} c).$$

The asymptotic variance for this estimator differs from the asymptotic variance for the estimator that matches on the true propensity score by $c' I_{\theta^*}^{-1} c$, where I_{θ^*} is the Fisher information matrix for the parametric model for the propensity score, and c is a vector that depends on the joint distribution of the outcome, the treatment, and the covariates. Thus, matching on the estimated propensity score has, in large samples, a weakly smaller asymptotic variance than matching on the true propensity score. This is in line with results in Rubin and Thomas (1992ab) who argue that, in settings with covariates with Normal distributions, matching on the estimated rather than the true propensity score improves the properties of matching estimators, and the result in Hahn (1998) that the efficiency bound for τ is not affected by knowledge of the propensity score. Hirano, Imbens and Ridder (2003) obtain a similar result for semiparametric weighting estimators.

Let $\hat{\tau}_{t,N}^*$ be an estimator of the average treatment effect on the treated, τ_t, obtained by matching on the true propensity score. Matching estimators of τ_t are constructed by matching treated units only. The large sample distribution of this estimator is

$$\sqrt{N}(\hat{\tau}_{t,N}^* - \tau_t) \xrightarrow{d} N(0, \sigma_t^2),$$

with the form for σ_t^2 given in Proposition 1 in Section 2. If one matches on the estimated propensity score,

$$\sqrt{N}(\hat{\tau}_{t,N} - \tau_t) \xrightarrow{d} N\left(0, \sigma_t^2 - c_t' I_{\theta^*}^{-1} c_t + \frac{\partial \tau_t(\theta^*)}{\partial \theta}' I_{\theta^*}^{-1} \frac{\partial \tau_t(\theta^*)}{\partial \theta}\right).$$

Here c_t (like c in the large sample distribution of $\hat{\tau}_N$) is a vector that depends on the joint distribution of the outcome, the treatment, and the covariates. In this case, the adjustment relative to the variance for the estimator based on matching on the true propensity score can be positive as well as negative, and ignoring the estimation error in the propensity score may lead to confidence intervals that are either too large or too small. This is consistent with Hahn's (1998) result that knowledge of the propensity score matters for the efficiency

bound for the average effect on the treated. We also propose estimators for the asymptotic variances of the two matching estimators.

The rest of the article is organized as follows. Section II provides an introduction to propensity score matching. In this section we provide a new result on the large sample distribution of matching estimators when the matching is on a single covariate. The technical contribution of this result over previous results in Abadie and Imbens (2006) is to allow for the possibility that the density of the matching variable goes to zero at the boundary of the support. This is important in the context of this article because the density of the propensity score is typically not bounded away from zero at the boundary of its support even if that is the case for the matching variables. Section III contains the main results of the article. In this section we derive the large sample properties of two estimators that match on estimated propensity scores, one estimating the overall average effect and one estimating the average effect on the treated. Section IV proposes estimators for the adjusted standard errors derived in section III. In section V we report the results of a small simulation exercise. Section VI concludes. Proofs are provided in an appendix.

II. Matching Estimators

The set up in this article is a standard one in the program evaluation literature. See Imbens and Wooldridge (2009) for a recent survey. In evaluation research the focus of the analysis is often the effect of a binary treatment, represented in this paper by the indicator variable W, on some outcome variable, Y. More specifically, $W = 1$ indicates exposure to the treatment, while $W = 0$ indicates lack of exposure to the treatment. Following Rubin (1974), we define treatment effects in terms of potential outcomes. We define $Y(1)$ as the potential outcome under exposure to treatment, and $Y(0)$ as the potential outcome under no exposure to treatment. Our goal is to estimate the average treatment effect,

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

where the expectation is taken over the population of interest, based on a random sample from this population. Alternatively the goal may be estimation of the average effect for

the treated,

$$\tau_t = E\Big[Y(1) - Y(0)\Big|W = 1\Big].$$

Estimation of these average treatment effects is complicated by the fact that for each unit in the population, we observe at most one of the potential outcomes:

$$Y = \begin{cases} Y(0) & \text{if } W = 0, \\ Y(1) & \text{if } W = 1. \end{cases}$$

Let X be a vector of covariates that representing all treatment confounders, that is, all variables that affect the probability of treatment exposure and the potential outcomes. The propensity score is $p(X) = \Pr(W = 1|X)$, and $p^* = E[W] = E[p(X)]$ is the marginal probability of being treated. The following assumption is often referred to as "strong ignorability" (Rosenbaum and Rubin, 1983).

Assumption 1: (i) $Y(1), Y(0) \perp\!\!\!\perp W|X$ almost surely; (ii) $\underline{p} \leq p(X) \leq \overline{p}$ almost surely, for some $\underline{p} > 0$ and $\overline{p} < 1$.

Assumption 1(i), also referred to as "unconfoundedness," will hold if all treatment confounders are included in X; so that after controlling for X, treatment exposure is independent of the potential outcomes. Assumption 1 (ii) implies that for almost all values of X the population includes treated and untreated units.

Define

$$\mu(w, x) = E[Y|W = w, X = x], \qquad \text{and} \qquad \sigma^2(w, x) = E[Y|W = w, X = x],$$

to be the conditional expectation and variance respectively of the outcome given treatment indicator and covariates, and define

$$\bar{\mu}(w, p) = E[Y|W = w, p(X) = p], \qquad \text{and} \qquad \bar{\sigma}^2(w, p) = E[Y|W = w, p(X) = p],$$

to be the conditional mean and variance respectively of the outcome given the treatment indicator and the propensity score. Rosenbaum and Rubin (1983) prove that, under Assumption 1,

$$\tau = E\Big[\bar{\mu}(1, p(X)) - \bar{\mu}(0, p(X))\Big], \qquad \text{and} \qquad \tau_t = E\Big[\bar{\mu}(1, p(X)) - \bar{\mu}(0, p(X))\Big|W = 1\Big].$$

In other words, adjusting for difference between treated and control units in the propensity score is sufficient to remove all biases associated with the observed treatment confounders.

This result by Rosenbaum and Rubin (1983) motivates the use of propensity score matching estimators. A propensity score matching estimator for the average treatment effect can be defined as:

$$\widehat{\tau}_N^* = \frac{1}{N}\sum_{i=1}^{N}(2W_i - 1)\left(Y_i - \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} Y_j\right), \tag{1}$$

where $\mathcal{J}_M(i)$ is the set of matches for unit i. (The superscript $*$ on $\hat{\tau}_N^*$ indicates that matching is done on the true propensity score.) In this article we will consider matching with replacement, so each unit in the sample can be used as a match multiple times. In the absence of matching ties, the set of matches $\mathcal{J}_M(i)$ can formally be defined as:

$$\mathcal{J}_M(i) = \left\{ j = 1,\ldots,N : W_j = 1 - W_i,\ \left(\sum_{k:W_k=1-W_i} 1_{|p(X_i)-p(X_k)|\leq|p(X_i)-p(X_j)|}\right) \leq M \right\}.$$

For the average effect on the treated, τ_t, the corresponding estimator is,

$$\widehat{\tau}_{t,N}^* = \frac{1}{N_1}\sum_{i=1}^{N} W_i\left(Y_i - \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} Y_j\right),$$

where $N_1 = \sum_{i=1}^{N} W_i$ is the number of treated units in the sample.

Let us now consider the large sample distributions of $\widehat{\tau}_N^*$ and $\widehat{\tau}_{t,N}^*$. Some version of these results is covered by the results in Abadie and Imbens (2006). However, one of their assumptions requires that the density of the matching variables is bounded away from zero. Although this assumption may be appropriate in settings where the matching is directly on covariates, it is much less appealing for propensity score matching estimators. For example, if the propensity score has the form $p(X) = F(X'\theta)$, then even if the density of X is bounded away from zero on its support, the density of $F(X'\theta)$ will generally not be bounded away from zero on its support. We therefore generalize the results in Abadie and Imbens (2006) to allow the density of the matching variable (the propensity score in our case) to go to zero at the boundary of its support. We first state an assumption about the propensity score.

ASSUMPTION 2: *(i) The propensity score $p(x)$ is continuously differentiable in x, with derivative bounded in absolute value by C_p, and (ii), a version of the density function of the propensity score, $f : [\underline{p}, \overline{p}] \mapsto [0, \infty)$, satisfies: (i) $f(p) \leq \overline{f}$ for all p, (ii) for any $\varepsilon > 0$, $\inf_{\underline{p}+\varepsilon<p<\overline{p}-\varepsilon} f(p) = f_\varepsilon > 0$, and $\sup_{\underline{p}+\varepsilon<p<\overline{p}-\varepsilon} |\partial f(p)/\partial p| = \overline{f}'_\varepsilon < \infty$.*

This assumption allows the density of the propensity score to go to zero at the boundary of its support. However, because the propensity score itself is bounded away from zero and one, the ratio of the conditional distribution of the propensity score in the two treatment groups is bounded and bounded away from zero. This is key for controlling the bias and variance of the matching estimator.

ASSUMPTION 3: *The support of X and Y is compact. $\bar{\mu}(w, p)$ and $\bar{\sigma}^2(w, p)$ are continuously differentiable in p on $[\underline{p}, \overline{p}]$ with derivatives with respect to p bounded by $\bar{\mu}'$ and $\bar{\sigma}^{2\prime}$respectively, for $w = 0, 1$.*

ASSUMPTION 4: *$\{(Y_i, W_i, X_i)\}_{i=1}^N$ are independent draws from the distribution of (Y, W, X).*

The next proposition reports the large sample distributions of $\hat{\tau}_N^*$ and $\widehat{\tau}_{t,N}^*$ under assumptions 1-4.

PROPOSITION 1: *Suppose assumptions 1-4 hold. Then, (i)*

$$\sqrt{N}\left(\hat{\tau}_N^* - \tau\right) \xrightarrow{d} N\left(0, \sigma^2\right),$$

where

$$\begin{aligned}
\sigma^2 = {} & E\left[\left(\bar{\mu}(1, p(X)) - \bar{\mu}(0, p(X)) - \tau\right)^2\right] && (2)\\
& + E\left[\bar{\sigma}^2(1, p(X))\left(\frac{1}{p(X)} + \frac{1}{2M}\left(\frac{1}{p(X)} - p(X)\right)\right)\right] && (3)\\
& + E\left[\bar{\sigma}^2(0, p(X))\left(\frac{1}{1-p(X)} + \frac{1}{2M}\left(\frac{1}{1-p(X)} - (1-p(X))\right)\right)\right], && (4)
\end{aligned}$$

and (ii),

$$\sqrt{N}\left(\widehat{\tau}_{t,N}^* - \tau_t\right) \xrightarrow{d} N\left(0, \sigma_t^2\right),$$

where

$$\sigma_t^2 = \frac{1}{E[p(X)]^2} E\left[p(X)\left(\bar{\mu}(1,p(X)) - \bar{\mu}(0,p(X)) - \tau_t\right)^2\right] + \frac{1}{E[p(X)]^2} E\left[\bar{\sigma}^2(1,p(X))p(X)\right] + \frac{1}{E[p(X)]^2} E\left[\bar{\sigma}^2(0,p(X))\left(\frac{p^2(X)}{1-p(X)} + \frac{1}{M}p(X) + \frac{1}{2M}\frac{p^2(X)}{1-p(X)}\right)\right].$$

The proof of this proposition is available as an additional web appendix (Abadie and Imbens, 2011).

Motivated by the fact that in observational studies propensity scores are not known, we are interested in the case where matching is not on the true propensity score $p(X)$, but on an estimate of the propensity score. Following Rosenbaum and Rubin (1983) and most of the empirical literature, we consider a generalized linear specification for the propensity score $p(x) = F(x'\theta)$. In empirical research the link function F is usually specified as Logit or Probit. It is straightforward to extend our results to more general parametric models for the propensity score. For unit i, and for arbitrary values for θ, let $\mathcal{J}_M(i,\theta)$ denote the set of M matches where we match on $F(X'\theta)$:

$$\mathcal{J}_M(i,\theta) = \left\{ j = 1,\ldots,N : W_j = 1 - W_i, \left(\sum_{k:W_k=1-W_i} 1_{|F(X_i'\theta)-F(X_k'\theta)| \leq |F(X_i'\theta)-F(X_j'\theta)|} \right) \leq M \right\}.$$

The matching estimator for the average treatment effect where we match on $F(X'\theta)$ is then:

$$\widehat{\tau}_N(\theta) = \frac{1}{N}\sum_{i=1}^{N}(2W_i - 1)\left(Y_i - \frac{1}{M}\sum_{j\in\mathcal{J}_M(i,\theta)} Y_j\right). \tag{5}$$

If θ^* denotes the true value of the parameter, the estimator based on matching on the true propensity score can also be written as $\hat{\tau}_N^* = \hat{\tau}_N(\theta^*)$. We are interested in the case where $\widehat{\tau}_N(\theta)$ is evaluated at an estimator $\hat{\theta}_N$ of θ^*, based on a sample $\{Y_i, W_i, X_i\}_{i=1}^N$. We focus on the case where $\widehat{\theta}_N$ is the maximum likelihood estimator of θ:[4]

$$\widehat{\theta}_N = \arg\max_\theta L(\theta | W_1, X_1, \ldots, W_N, X_N),$$

[4]It is straightforward to extend our results to alternative asymptotically linear estimators of θ^*.

where the log likelihood function is

$$L(\theta|W_1, X_1, \ldots, W_N, X_N) = \sum_{i=1}^{N} W_i \ln F(X_i'\theta) + (1 - W_i) \ln(1 - F(X_i'\theta)).$$

The propensity score matching estimator of τ that matches on the estimated propensity score can now be written as:

$$\widehat{\tau}_N = \widehat{\tau}_N(\widehat{\theta}_N) = \frac{1}{N} \sum_{i=1}^{N} (2W_i - 1) \left(Y_i - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i, \widehat{\theta}_N)} Y_j \right). \tag{6}$$

Similarly, the propensity score matching estimator of τ_t that matches on the estimated propensity score can be written as:

$$\widehat{\tau}_{t,N} = \widehat{\tau}_{t,N}(\widehat{\theta}_N) = \frac{1}{N_1} \sum_{i=1}^{N} W_i \left(Y_i - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i, \widehat{\theta}_N)} Y_j \right). \tag{7}$$

Whenever confusion is possible, we will be explicit in the dependence of the matching estimators on θ. If the argument is omitted, $\widehat{\tau}_N$ and $\widehat{\tau}_{t,N}$ are used as shorthand for $\widehat{\tau}_N(\widehat{\theta}_N)$ and $\widehat{\tau}_{t,N}(\widehat{\theta}_N)$ respectively.

The two main questions addressed in this paper are (i), do the estimators based on matching on the estimated propensity score have Normal large sample distributions, and (ii), if so, how do the large sample variance compare to those of the matching estimator based on matching on the true propensity score, given in Proposition 1? In the next section we answer these two questions and derive the large sample distribution of $\widehat{\tau}_N(\widehat{\theta}_N)$ and $\widehat{\tau}_{t,N}(\widehat{\theta}_N)$. Conventional linearization methods for two-step statistics are difficult to apply in the context of matching estimators because matching estimators are complicated functionals of the distribution of the data. We therefore follow a different route, building on work by Andreou and Werker (2005) on residual based statistics, and the martingale representations for matching estimators derived by Abadie and Imbens (2010).

III. Large Sample Distribution

In the first part of this section we derive the large sample approximation to the sampling distribution of $\widehat{\tau}_N(\widehat{\theta}_N)$, and in the second part we present the results for $\widehat{\tau}_{t,N}(\widehat{\theta}_N)$.

To derive the distributions for these estimators, we need to establish the properties of particular statistics in a neighborhood of the true data generating process. Let P^θ be the distribution of $Z = \{Y, W, X\}$ induced by the propensity score, $F(X'\theta)$, the marginal distribution of X, and the conditional distribution of Y given X and W. We index this distribution P^θ by θ, and will consider properties of estimators for different values of θ, under the same marginal distribution for X, and the same conditional distribution for Y given W and X. Let θ^* denote the true value of the parameter, so that $p(X) = F(X'\theta^*)$. Given assumption 1, the overall average treatment effect is

$$\tau = E[Y(1) - Y(0)] = E\left[E[Y|W=1,X] - E[Y|W=0,X]\right].$$

This parameter does not depend on θ^* because it depends only on the conditional distribution of Y given W and X and the marginal distribution of X. The average treatment effect for the treated is

$$\begin{aligned}\tau_t &= E[Y(1) - Y(0)|W=1] = E\left[E[Y|W=1,X] - E[Y|W=0,X]|\, W=1\right] \\ &= \frac{E\left[F(X'\theta^*)\Big(E[Y|W=1,X] - E[Y|W=0,X]\Big)\right]}{E[F(X'\theta^*)]}.\end{aligned}$$

In contrast to the average treatment effect, τ, the average treatment effect for the treated, τ_t, depends on the propensity score, and we make this dependence explicit by indexing τ_t by θ wherever appropriate. Let $\tau_t = \tau_t(\theta^*)$ be the true value for the average effect on the treated. Also, let $S_\theta = F(X'\theta)$ denote the random variable equal to the propensity score, evaluated at θ, for different values of θ.

Consider $Z_{N,i} = \{Y_{N,i}, W_{N,i}, X_{N,i}\}$ with distribution given by the local "shift" P^{θ_N} with $\theta_N = \theta^* + h/\sqrt{N}$, where h is a conformable vector of constants.

A. Large Sample Distribution for $\widehat{\tau}_N(\widehat{\theta}_N)$

First we restrict the functional form of the propensity score, $F(\cdot)$.

ASSUMPTION 5: *$F(a)$ is continously differentiable, with derivative $f(a)$ bounded, and bounded away from zero for $a \in [\underline{a}, \overline{a}]$, where, for some $\epsilon > 0$, $\underline{a} < \inf_{x \in \text{Supp}(X), \|\theta - \theta^*\| < \epsilon} x'\theta$ and $\overline{a} > \sup_{x \in \text{Supp}(X), \|\theta - \theta^*\| < \epsilon} x'\theta$.*

Next, we extend Assumptions 1, 2 and 3 to hold for all θ in a neighborhood of θ^*.

ASSUMPTION 6: *For some $\varepsilon > 0$, and for all $\|\theta - \theta^*\| \leq \varepsilon$,*
(i) The distribution of $S_\theta = F(X'\theta)$ is continuous with support equal to an interval $[\underline{s}_\theta, \overline{s}_\theta]$, with $\underline{s}_\theta > 0$ and $\overline{s}_\theta < 1$.
(ii) A version of the density function of the propensity score, $f_\theta(s)$, satisfies: (1) $f_\theta(s) \leq \overline{f}_\theta$ *for all* $s \in [\underline{s}_\theta, \overline{s}_\theta]$, *and* (2), *for any* $\varepsilon > 0$, $\inf_{\underline{s}_\theta + \varepsilon < s < \overline{s} - \varepsilon} f_\theta(s) = f_{\varepsilon,\theta} > 0$, *and* $\sup_{\underline{p}+\varepsilon<p<\overline{p}-\varepsilon} \left|\frac{\partial f}{\partial p}(p)\right| = \overline{f}'_\varepsilon < \infty$.
(ii) For all s in the support of S_θ, and $w = 0, 1$, the conditional expectation and variance of Y given $W = w$ and $S_\theta = s$, , under P^θ, $\bar{\mu}(w, s)$ and $\bar{\sigma}^2(w, s)$ are continuously differentiable in s, with derivatives bounded by $C_{s,\theta}$.

Next we impose some regularity conditions on the model for the propensity score. Let $\Lambda_N(\theta|\theta')$ be the difference in log likelihood functions,

$$\Lambda_N(\theta|\theta') = L(\theta|Z_{N,1}, \ldots, Z_{N,N}) - L(\theta'|Z_{N,1}, \ldots, Z_{N,N}),$$

let $\Delta_N(\theta)$ be the normalized score function, or the central sequence,

$$\Delta_N(\theta) = \frac{1}{\sqrt{N}} \frac{\partial}{\partial \theta} L(\theta|Z_{N,1}, \ldots, Z_{N,N}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_{N,i} \frac{W_{N,i} - F(X'_{N,i}\theta)}{F(X'_{N,i}\theta)(1 - F(X'_{N,i}\theta))} f(X'_{N,i}\theta),$$

and let

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

be the Fisher Information Matrix for θ. The expectation in this equation is taken over the marginal distribution of X, which does not depend on θ, so the indexing by θ solely reflects the value of θ where $f(X'\theta)$ and $F(X'\theta)$ are evaluated. The following assumption states the regularity conditions we impose on the parametric model for the propensity score. These conditions are sufficient for asymptotic Normality of the maximum likelihood estimator, but they are slightly stronger by also restricting the behavior of the maximum likelihood estimator in a shrinking neighborhood of the true value θ^*.

ASSUMPTION 7: *Under* P^{θ_N}*:*

$$\Lambda_N(\theta^*|\theta_N) = -h'\Delta_N(\theta_N) - \frac{1}{2}h'I_{\theta^*}h + o_p(1), \tag{8}$$

$$\Delta_N(\theta_N) \xrightarrow{d} N(0, I_{\theta^*}), \tag{9}$$

and

$$\sqrt{N}(\widehat{\theta}_N - \theta_N) = I_{\theta^*}^{-1}\Delta_N(\theta_N) + o_p(1). \tag{10}$$

For regular parametric models, equation (8) can be established using Proposition 2.1.2 in Bickel et al. (1998). Also for regular parametric models, equation (9) is derived in the proof of Proposition 2.1.2 in Bickel et al. (1998). Equation (10) can be established using the same set of results in combination with classical conditions for asymptotic linearity of maximum likelihood estimators (see, e.g., van der Vaart (1998) Theorem 5.39; Lehmann and Romano (2005) Theorem 12.4.1).

The following assumption is a regularity condition that will be used later in this section.

ASSUMPTION 8: *For all bounded functions* $h(y, w, x)$*, the sequence of functions* $k(s, w, \theta_N) = E_{\theta_N}[h(Y, W, X)|F(X'\theta_N) = s, W = w]$ *converges to* $k(s, w, \theta^*) = E[h(Y, W, X)|F(X'\theta^*) = s, W = w]$ *(where* E_{θ_N} *denotes an expectation with respect to* P^{θ_N}*).*

Primitive conditions for this assumption can be established using the results in Ganssler and Pfanzagl (1971).

Our derivation of the limit distribution of $\sqrt{N}(\widehat{\tau}_N - \tau)$ is based on the techniques developed in Andreou and Werker (2005) to analyze the limit distribution of residual-based statistics. We proceed in three steps. First, we derive the joint limit distribution of $(\sqrt{N}\big(\widehat{\tau}_N(\theta_N) - \tau\big), \sqrt{N}\big(\widehat{\theta} - \theta_N\big), \Lambda_N(\theta^*|\theta_N))$ under P^{θ_N}.

PROPOSITION 2: *Suppose that Assumptions 1–8 hold. Then, under* P^{θ_N}*:*

$$\begin{pmatrix} \sqrt{N}\big(\widehat{\tau}_N(\theta_N) - \tau\big) \\ \sqrt{N}(\widehat{\theta}_N - \theta_N) \\ \Lambda_N(\theta^*|\theta_N) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \\ -h'I_{\theta^*}h/2 \end{pmatrix}, \begin{pmatrix} \sigma^2 & c'I_{\theta^*}^{-1} & -c'h \\ I_{\theta^*}^{-1}c & I_{\theta^*}^{-1} & -h \\ -h'c & -h' & h'I_{\theta^*}h \end{pmatrix}\right),$$

where

$$c = E\left[\operatorname{cov}\left(X, Y\middle|F(X'\theta^*), W\right) f(X'\theta^*)\left(\frac{W}{F(X'\theta^*)^2} + \frac{1-W}{(1-F(X'\theta^*))^2}\right)\right]. \tag{11}$$

PROOF: See Appendix.

Normality of the first component, $\sqrt{N}\big(\widehat{\tau}_N(\theta_N) - \tau\big)$, follows from Proposition 1. Joint Normality of the last two components, $\sqrt{N}(\widehat{\theta}_N - \theta_N)$ and $\Lambda_N(\theta^*|\theta_N)$, follows directly from Assumption 7. The key result in the proposition is that the first component and the last two components have a joint Normal distribution. If $\widehat{\tau}_N(\theta_N)$ were asymptotically linear, this would be straightforward to establish. Given that asymptotic linearity has not been established, and in fact may not hold, the result is more subtle. In the proof, extending Abadie and Imbens (2010), we use martingale limit theory to derive this joint Normality.

In the second step we use LeCam's third lemma (e.g., Van der Vaart, 1998, p. 90). LeCam's third lemma implies that, if under P^{θ_N}

$$\begin{pmatrix} \sqrt{N}\big(\widehat{\tau}_N(\theta_N) - \tau\big) \\ \sqrt{N}(\widehat{\theta}_N - \theta_N) \\ \Lambda_N(\theta^*|\theta_N) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \\ -h' I_{\theta^*} h/2 \end{pmatrix}, \begin{pmatrix} \sigma^2 & c' I_{\theta^*}^{-1} & -c'h \\ I_{\theta^*}^{-1} c & I_{\theta^*}^{-1} & -h \\ -h'c & -h' & h' I_{\theta^*} h \end{pmatrix}\right),$$

(as implied by Proposition 2) then, under P^{θ^*} instead of under P^{θ_N},

$$\begin{pmatrix} \sqrt{N}\big(\widehat{\tau}_N(\theta_N) - \tau\big) \\ \sqrt{N}(\widehat{\theta}_N - \theta_N) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} -c'h \\ -h \end{pmatrix}, \begin{pmatrix} \sigma^2 & c' I_{\theta^*}^{-1} \\ I_{\theta^*}^{-1} c & I_{\theta^*}^{-1} \end{pmatrix}\right).$$

Substituting $\theta_N = \theta^* + h/\sqrt{N}$, this implies that (still under P^{θ^*}) for any $h \in \mathbb{R}^k$,

$$\begin{pmatrix} \sqrt{N}\big(\widehat{\tau}_N(\theta^* + h/\sqrt{N}) - \tau\big) \\ \sqrt{N}(\widehat{\theta}_N - \theta^*) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} -c'h \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & c' I_{\theta^*}^{-1} \\ I_{\theta^*}^{-1} c & I_{\theta^*}^{-1} \end{pmatrix}\right). \tag{12}$$

The distribution in (12) is key. If this Normal distribution were exact rather than an approximation based on convergence in distribution, this would directly lead to the result of interest. In that case it would follow that

$$\sqrt{N}\big(\widehat{\tau}_N(\theta^* + h/\sqrt{N}) - \tau\big)|\sqrt{N}(\widehat{\theta}_N - \theta^*) \sim N\left(-c'h + c'\sqrt{N}(\widehat{\theta}_N - \theta^*), \sigma^2 - c' I_{\theta^*}^{-1} c\right).$$

Thus, conditional on $\sqrt{N}(\widehat{\theta}_N - \theta^*) = h$,

$$\sqrt{N}\big(\widehat{\tau}_N(\theta^* + h/\sqrt{N}) - \tau\big)|\sqrt{N}(\widehat{\theta}_N - \theta) = h \sim N\left(0, \sigma^2 - c' I_{\theta^*}^{-1} c\right).$$

Because $\sqrt{N}(\widehat{\theta}_N - \theta^*) = h$ implies $\theta^* + h/\sqrt{N} = \widehat{\theta}_N$, and thus implies that $\widehat{\tau}_N(\theta^* + h/\sqrt{N}) = \widehat{\tau}_N(\widehat{\theta}_N) = \widehat{\tau}_N$, this implies that

$$\sqrt{N}\big(\widehat{\tau}_N - \tau\big)|\sqrt{N}(\widehat{\theta}_N - \theta) = h \sim N\left(0, \sigma^2 - c' I_{\theta^*}^{-1} c\right).$$

Because the right hand side does not depend on h, this in turn implies that, under P^{θ^*}, unconditionally,

$$\sqrt{N}\left(\widehat{\tau}_N - \tau\right) \sim N\left(0, \sigma^2 - c' I_{\theta^*}^{-1} c\right),$$

which is the result we are looking for: the distribution of the matching estimator based on matching on the estimated propensity score.

The complication is that although the joint distribution of $\sqrt{N}\big(\widehat{\tau}_N(\theta + h/\sqrt{N}) - \tau\big)$ and $\sqrt{N}(\widehat{\theta}_N - \theta)$ converges to a joint Normal distribution, that does not imply that the conditional distribution of $\sqrt{N}\big(\widehat{\tau}_N(\theta+h/\sqrt{N})-\tau\big)$ given $\sqrt{N}(\widehat{\theta}_N-\theta)$ converges to a Normal distribution. Formally, we need to discretize the first step estimators $\widehat{\theta}_N$, as in Andreou and Werker (2005), and this is what we do in the third step.

Consider a grid of cubes in $\mathbb{R}^k$ with sides of length $d/\sqrt{N}$, for arbitrary positive d. Then $\bar{\theta}_N$ is the discretized estimator, defined as the midpoint of the cube $\widehat{\theta}_N$ belongs to. If $\hat{\theta}_{N,j}$ is the jth component of the k-vector $\hat{\theta}_N$, then the jth component of the k-vector $\bar{\theta}_N$ is $\bar{\theta}_{N,j} = (d/\sqrt{N})(\lfloor\sqrt{N}\hat{\theta}_{N,j}/d\rfloor + \lceil\sqrt{N}\hat{\theta}_{N,j}/d\rceil)/2$, where $\lfloor a\rfloor$ and $\lceil a\rceil$ are the largest integer smaller than a and the smallest integer at least equal to a respectively. Now we can state the main result of the paper.

THEOREM 1: *Suppose Assumptions 1–8 hold. Then, under P^{θ^*},*

$$\lim_{d\downarrow 0}\lim_{N\to\infty} \Pr\left(\sqrt{N}\left(\sigma^2 - c' I_{\theta^*}^{-1} c\right)^{-1/2}\left(\widehat{\tau}_N(\bar{\theta}_N) - \tau\right) \le z\right) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) dx.$$

PROOF: See Appendix.

The implication is that we can approximate the distribution of of $\sqrt{N}(\widehat{\tau}_N(\hat{\theta}_N) - \tau)$ by a Normal distribution with mean zero and variance $\sigma^2 - c' I_{\theta^*}^{-1} c$.

B. Large Sample Distribution for $\widehat{\tau}_{t,N}$

In this section we consider the asymptotic distribution for $\sqrt{N}(\widehat{\tau}_{t,N} - \tau_t(\theta^*))$. First we present the equivalent to Proposition 2 for the average effect on the treated. A key difference with the result for the average effect is that $\tau_t(\theta)$ depends explicitly on θ, as long as the treatment effect varies by the propensity score. This dependence shows up in the asymptotic distribution.

PROPOSITION 3: *Suppose Assumptions 1–8 hold. Then, under* P^{θ_N}*:*

$$\begin{pmatrix} \sqrt{N}\big(\widehat{\tau}_{t,N}(\theta_N) - \tau_t(\theta_N)\big) \\ \sqrt{N}(\widehat{\theta}_N - \theta_N) \\ \Lambda_N(\theta^*|\theta_N) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \\ -h'I_{\theta^*}h/2 \end{pmatrix}, \begin{pmatrix} \sigma_t^2 & c_t'I_{\theta^*}^{-1} & -c_t'h \\ I_{\theta^*}^{-1}c_t & I_{\theta^*}^{-1} & -h \\ -c_t'h & -h & h'I_{\theta^*}h \end{pmatrix}\right),$$

where

$$c_t = \frac{1}{E[F(X'\theta^*)]}E\left[Xf(X'\theta^*)\left(\bar{\mu}(1,F(X'\theta^*)) - \bar{\mu}(0,F(X'\theta^*)) - \tau_t\right)\right] \tag{13}$$
$$+\frac{1}{E[F(X'\theta^*)]}E\left[\operatorname{cov}\left(X, Y\middle|F(X'\theta^*), W\right)f(X'\theta^*)\left(\frac{W}{F(X'\theta^*)} + \frac{(1-W)F(X'\theta^*)}{(1-F(X'\theta^*))^2}\right)\right].$$

PROOF: See Appendix.

Le Cam's third lemma now implies that, under P^{θ^*} instead of under P^{θ_N},

$$\begin{pmatrix} \sqrt{N}\big(\widehat{\tau}_{t,N}(\theta_N) - \tau_t(\theta_N)\big) \\ \sqrt{N}(\widehat{\theta}_N - \theta_N) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} -c_t'h \\ -h \end{pmatrix}, \begin{pmatrix} \sigma_t^2 & c_t'I_{\theta^*}^{-1} \\ I_{\theta^*}^{-1}c_t & I_{\theta^*}^{-1} \end{pmatrix}\right).$$

Because $\tau_t(\theta_N) = \tau_t(\theta^*) + \frac{\partial\tau_t(\theta^*)}{\partial\theta}'h/\sqrt{N} + o(N^{-1/2})$, and using the shorthand $\tau_t = \tau_t(\theta^*)$, this implies that, under P^{θ^*},

$$\begin{pmatrix} \sqrt{N}\big(\widehat{\tau}_{t,N}(\theta_N) - \tau_t\big) \\ \sqrt{N}(\widehat{\theta}_N - \theta^*) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} -c_t'h + \dfrac{\partial\tau_t(\theta^*)}{\partial\theta}'h \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_t^2 & c_t'I_{\theta^*}^{-1} \\ I_{\theta^*}^{-1}c_t & I_{\theta^*}^{-1} \end{pmatrix}\right).$$

The same heuristic argument used for Proposition 2 implies that if this Normality was exact, then the conditional distribution of $\sqrt{N}\big(\widehat{\tau}_{t,N}(\theta_N) - \tau_t\big)$ given $\sqrt{N}(\widehat{\theta}_N - \theta^*)$ would be

$$\sqrt{N}\big(\widehat{\tau}_N(\theta_N) - \tau_t\big)|\sqrt{N}(\widehat{\theta}_N - \theta^*) \sim N\left(-c_t'h + \frac{\partial\tau_t(\theta^*)}{'}h + c_t'\sqrt{N}(\widehat{\theta}_N - \theta^*), \sigma_t^2 - c_t'I_{\theta^*}^{-1}c_t\right).$$

Thus, conditional on $\sqrt{N}(\widehat{\theta}_N - \theta^*) = h$,

$$\sqrt{N}\big(\widehat{\tau}_N(\theta_N) - \tau_t\big)|\sqrt{N}(\widehat{\theta}_N - \theta^*) = h \sim N\left(\frac{\partial\tau_t(\theta^*)}{\partial\theta}, \sigma^2 - c_t'I_{\theta^*}^{-1}c_t\right).$$

Given $\sqrt{N}(\widehat{\theta}_N - \theta^*) = h$, $\theta_N = \widehat{\theta}_N$, and so $\widehat{\tau}_N(\theta_N) = \widehat{\tau}_N(\widehat{\theta}_N) = \widehat{\tau}_N$. Hence, under P^{θ^*},

$$\sqrt{N}\big(\widehat{\tau}_N - \tau_t\big)|\sqrt{N}(\widehat{\theta}_N - \theta^*) = h \sim N\left(\frac{\partial\tau_t(\theta^*)}{\partial\theta}'h, \sigma^2 - c_t'I_{\theta^*}^{-1}c_t\right).$$

Note that in contrast to the analysis for $\hat{\tau}$ in Section II.A, the conditional distribution of $\sqrt{N}\big(\widehat{\tau}_N - \tau_t\big)$ given $\sqrt{N}(\widehat{\theta}_N - \theta^*) = h$ does vary with h. In this case the unconditional distribution of $\sqrt{N}\big(\widehat{\tau}_N - \tau_t\big)$ under P^{θ^*} is

$$\sqrt{N}\left(\widehat{\tau}_{t,N} - \tau_t\right) \xrightarrow{d} N\left(0, \sigma_t^2 - c_t' I_{\theta^*}^{-1} c_t + \frac{\partial \tau_t(\theta^*)}{\partial \theta}' I_{\theta^*}^{-1} \frac{\partial \tau_t(\theta^*)}{\partial \theta}\right).$$

Formally we need again to discretize the estimator. This leads to the following theorem that covers the case of matching on the estimated propensity score for the average effect for the treated.

THEOREM 2: *Suppose Assumptions 1–8 hold. Then, under* P^{θ^*},

$$\lim_{d \downarrow 0} \lim_{N \to \infty} \Pr\left(\sqrt{N}\left(\sigma_t^2 - c_t' I_{\theta^*}^{-1} c_t + \frac{\partial \tau_t(\theta^*)}{\partial \theta}' I_{\theta^*}^{-1} \frac{\partial \tau_t(\theta^*)}{\partial \theta}\right)^{-1/2} \big(\widehat{\tau}_N(\bar{\theta}_N) - \tau_t\big) \leq z\right)$$
$$= \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) dx.$$

The proof follows the same argument as that for 1, and is omitted.

IV. Estimation of the Asymptotic Variance

Here we discuss estimation of the two variances. As shown in the previous section, the asymptotic variance for $\sqrt{N}(\hat{\tau}_N - \tau)$ is

$$\sigma^2_{\text{adj}} = \sigma^2 - c' I_{\theta^*}^{-1} c, \tag{14}$$

and the asymptotic variance for $\sqrt{N}(\hat{\tau}_{t,N} - \tau_t)$ is

$$\sigma^2_{t,\text{adj}} = \sigma_t^2 - c_t' I_{\theta^*}^{-1} c_t + \frac{\partial \tau_t(\theta^*)}{\partial \theta}' I_{\theta^*}^{-1} \frac{\partial \tau_t(\theta^*)}{\partial \theta}. \tag{15}$$

The unknown components of the two expressions for the variance are I_θ, σ_τ^2, σ_t^2, c, c_t, and $\partial \tau_t(\theta^*)/\partial \theta$. We propose estimators for each of these components, and propose substituting these estimators into the variance expressions in (14) and (15).

First, estimation of the information matrix I_θ is standard:

$$\widehat{I}_\theta = \frac{1}{N} \sum_{i=1}^{N} \frac{f(X_i'\widehat{\theta}_N)^2}{F(X_i'\widehat{\theta}_N)(1 - F(X_i'\widehat{\theta}_N))} X_i X_i'.$$

Second, consider estimation of the variances corresponding to matching on the true propensity score, σ^2 and σ_t^2. Here we adapt the variance estimators proposed in Abadie and Imbens (2006). Let $\mathcal{H}_L(i,\theta)$ be the set of the L units with $W = W_i$ and closest values of $F(X'\theta)$ to $F(X_i'\theta)$,

$$\mathcal{H}_L(i,\theta) = \left\{ j = 1,\ldots,N : i \neq j, W_j = W_i, \left(\sum_{k:W_k=W_i} 1_{|F(X_i'\theta)-F(X_k'\theta)| \leq |F(X_i'\theta)-F(X_j'\theta)|} \right) \leq L \right\}.$$

Consider the following estimator of $\bar{\sigma}^2(W_i, F(X_i'\theta))$:

$$\widehat{\bar{\sigma}}^2(W_i, F(X_i'\theta)) = \frac{1}{L} \sum_{j \in \{i \cup \mathcal{H}_L(i,\theta)\}} \left(Y_j - \frac{1}{L+1} \sum_{k \in \{i \cup \mathcal{H}_L(i,\theta)\}} Y_k \right)^2.$$

Abadie and Imbens (2006) show how these estimates of the conditional variance can be used to estimate σ^2 and σ_t^2. A key step is the following representation of the matching estimators

$$\hat{\tau}_N = \frac{1}{N} \sum_{i=1}^{N} (2W_i - 1) \left(1 + \frac{K_{M,\theta}(i)}{M} \right) Y_i$$

and

$$\hat{\tau}_{t,N} = \frac{1}{N_1} \sum_{i=1}^{N} \left(W_i - (1 - W_i) \frac{K_{M,\theta}(i)}{M} \right) Y_i$$

where $K_{M,\theta}(i)$ is the number of times that observation i is used as a match (when matching on $F(X'\theta)$):

$$K_{M,\theta}(i) = \sum_{j=1}^{N} 1_{i \in \mathcal{J}_M(j,\theta)}. \tag{16}$$

Then we estimate σ^2 and σ_t^2 as

$$\begin{aligned} \widehat{\sigma}^2 = & \frac{1}{N} \sum_{i=1}^{N} \left((2W_i - 1) \left(Y_i - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i,\widehat{\theta})} Y_j \right) - \widehat{\tau}_N \right)^2 \\ & + \frac{1}{N} \sum_{i=1}^{N} \left(\left(\frac{K_{M,\widehat{\theta}}(i)}{M} \right)^2 + \frac{2M-1}{M} \left(\frac{K_{M,\widehat{\theta}}(i)}{M} \right) \right) \cdot \widehat{\bar{\sigma}}^2(W_i, F(X_i'\widehat{\theta})), \end{aligned}$$

and

$$\widehat{\sigma}_t^2 = \frac{N}{N_1^2} \sum_{i=1}^{N} W_i \cdot \left(Y_i - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i,\widehat{\theta})} Y_j \right) - \widehat{\tau}_{t,N} \Bigg)^2$$

$$+\frac{N}{N_1^2}\sum_{i=1}^{N}(1-W_i)\cdot\left(\frac{K_{M,\widehat{\theta}}(i)\cdot\left(K_{M,\widehat{\theta}}(i)-1\right)}{M^2}\right)\cdot\widehat{\bar{\sigma}}^2(W_i,F(X_i'\widehat{\theta})).$$

Third, consider estimation of c. Let $\bar{X}_i^{(L,\theta)}$ be the averages of X for for the $L+1$ units in the set $\{i\cup\mathcal{H}_L(i,\theta)\}$. Consider the following estimator for the covariance $\text{cov}(X,Y|F(X'\theta),W)$:

$$\widehat{\text{cov}}(X_i,Y_i|F(X_i'\theta),W_i)=\frac{1}{L}\sum_{j\in\{i\cup\mathcal{H}_L(i,\theta)\}}\left(X_j-\frac{1}{L+1}\sum_{k\in\{i\cup\mathcal{H}_L(i,\theta)\}}X_k\right)\left(Y_j-\frac{1}{L+1}\sum_{k\in\{i\cup\mathcal{H}_L(i,\theta)\}}Y_k\right).$$

Our estimator of c is:

$$\widehat{c}=\frac{1}{N}\sum_{i=1}^{N}\widehat{\text{cov}}(X_i,Y_i|F(X_i'\widehat{\theta}_N),W_i)f(X_i'\widehat{\theta}_N)\left(\frac{W_i}{F(X_i'\widehat{\theta}_N)^2}+\frac{1-W_i}{(1-F(X_i'\widehat{\theta}_N))^2}\right).$$

Fourth, consider estimation of c_t. We consider separately the two components of $c_t=c_{t,1}+c_{t,2}$, where

$$c_{t,1}=\frac{1}{E[F(X'\theta^*)]}E\left[Xf(X'\theta^*)\left(\bar{\mu}(1,F(X'\theta^*))-\bar{\mu}(0,F(X'\theta^*))-\tau_t\right)\right],$$

and

$$c_{t,2}=\frac{1}{E[F(X'\theta^*)]}E\left[\text{cov}\left(X,Y\middle|F(X'\theta^*),W\right)f(X'\theta^*)\left(\frac{W}{F(X'\theta^*)}+\frac{(1-W)F(X'\theta^*)}{(1-F(X'\theta^*))^2}\right)\right].$$

The second component, $c_{t,2}$ is similar to c_t, and our proposed estimator for $c_{t,2}$ is correspondingly similar to the estimator for c:

$$\widehat{c}_{t,2}=\frac{1}{N_1}\sum_{i=1}^{N}\widehat{\text{cov}}(X_i,Y_i|F(X_i'\widehat{\theta}),W_i)f(X_i'\widehat{\theta}_N)\left(\frac{W_i}{F(X_i'\widehat{\theta}_N)}+\frac{(1-W_i)F(X_i'\widehat{\theta}_N)}{(1-F(X_i'\widehat{\theta}_N))^2}\right).$$

The first component of c_t, $c_{t,1}$, is different. The key issue is to estimate $\bar{\mu}(w,F(X'\theta^*))$. For units with $W_i=0$, we estimate $\bar{\mu}(1,F(X'\theta^*))$ as $\widehat{Y}_i(1)$, and $\bar{\mu}(0,F(X'\theta^*))$ as $\bar{Y}_{(i)}^{(L,\theta^*)}$. Formally, define

$$\widehat{\bar{\mu}}(0,F(X'\theta))=\begin{cases}\dfrac{1}{L}\displaystyle\sum_{j\in\mathcal{H}_L(i,\theta)}Y_j & \text{if } W_i=0\\ \dfrac{1}{L}\displaystyle\sum_{j\in\mathcal{J}_L(i,\theta)}Y_j & \text{if } W_i=1,\end{cases}$$

and

$$\widehat{\bar{\mu}}(1, F(X'\theta)) = \begin{cases} \dfrac{1}{L} \displaystyle\sum_{j \in \mathcal{J}_L(i,\theta)} Y_j & \text{if } W_i = 0, \\ \dfrac{1}{L} \displaystyle\sum_{j \in \mathcal{H}_L(i,\theta)} Y_j & \text{if } W_i = 1. \end{cases}$$

Our proposed estimator for $c_{t,1}$ is

$$\widehat{c}_{t,1} = \frac{1}{N_1} \sum_{i=1}^{N} X_i f(X_i'\widehat{\theta}_N) \left(\widehat{\bar{\mu}}(1, F(X'\widehat{\theta})) - \widehat{\bar{\mu}}(0, F(X'\widehat{\theta})) - \widehat{\tau}_t \right).$$

The fifth component is $\partial\tau(\theta^*)/\partial\theta$. We can write this as

$$\frac{\partial \tau}{\partial \theta}(\theta^*) = \frac{1}{E[F(X'\theta^*)]} \cdot E\left[X f(X'\theta^*) \left(Y(1) - Y(0) - \tau_t \right) \right].$$

To estimate this component we need to match on the covariates rather than on the propensity score. Define the matching set when we match on all covariates,

$$\mathcal{J}_M^X(i) = \left\{ j = 1, \ldots, N : W_j = 1 - W_i, \left(\sum_{k: W_k = 1 - W_i} 1_{\|X_i - X_k\| \leq \|X_i - X_j\|} \right) \leq M \right\}.$$

Then define,

$$\widehat{\frac{\partial \tau}{\partial \theta}}(\theta) = \frac{1}{N_1} \sum_{i=1}^{N} X_i f(X_i'\theta) \left((2W_i - 1) \left(Y_i - \frac{1}{M} \sum_{j \in \mathcal{J}_M^X(i)} Y_j \right) - \widehat{\tau}_t \right),$$

and we estimate $\partial\tau(\theta^*)/\partial\theta$ as $\widehat{\partial\tau}(\widehat{\theta})/\partial\theta$.

Our estimator of the large sample variance of the propensity score matching estimator, adjusted for first step estimation of the propensity score, is:

$$\widehat{\sigma}^2_{\text{adj}} = \widehat{\sigma^2} - \widehat{c}' \widehat{I}_\theta^{-1} \widehat{c}. \tag{17}$$

The corresponding estimator for the variance for the estimator for the average effect for the treated is

$$\widehat{\sigma}^2_{\text{adj},t} = \widehat{\sigma}^2_t - \widehat{c}_t' \widehat{I}_\theta^{-1} \widehat{c}_t + \widehat{\frac{\partial \tau}{\partial \theta}}(\hat{\theta}_N)' \widehat{I}_\theta^{-1} \widehat{\frac{\partial \tau}{\partial \theta}}(\hat{\theta}_N). \tag{18}$$

Consistency of these estimators can be shown using the results in Abadie and Imbens (2006) and the contiguity arguments employed in section III.

V. A Small Simulation Exercise

In this section we report the results of a small simulation exercise designed to investigate the sampling distribution of propensity score matching estimators and the quality of the approximation to that distribution that is proposed in this article. In particular we use simulation designs to illustrate the importance of adjusting for the estimation error in the propensity score and to confirm our theoretical results.

We report results for two designs, for the two estimators considered in section II. In all cases we use a single match ($M = 1$) and $N = 5000$ observations. For each design and each estimand we calculate two estimators and two variances. The first estimator is based on matching on the true propensity score, $\hat{\tau}_N^* = \hat{\tau}_N(\theta^*)$ for the average effect and $\hat{\tau}_{t,N}^* = \hat{\tau}_{t,N}(\theta^*)$ for the average effect on the treated. The second estimator is based on matching on the estimated propensity score, $\hat{\tau}_N = \hat{\tau}_N(\hat{\theta}_N)$ for the average effect and $\hat{\tau}_{t,N} = \hat{\tau}_{t,N}(\hat{\theta}_N)$ for the average effect on the treated. The first pair of variance estimators we consider is $\hat{\sigma}_\tau^2$ and $\hat{\sigma_t^2}$. These are appropriate for matching on the true propensity score. We estimate these using the true propensity score. The second pair of variance estimators we consider is $\hat{\sigma}_{\text{adj},\tau}^2$ and $\hat{\sigma}_{\text{adj},\tau,t}^2$. These are appropriate for matching on the estimated propensity score. We estimate these using the estimated propensity score. We then evaluate the performance of two confidence intervals for each estimand, all based on estimators that match on the estimated propensity score, $\hat{\tau}_N$ and $\hat{\tau}_{t,N}$. For τ the first confidence interval is

$$\left(\hat{\tau}_N - 1.96 \times \frac{\hat{\sigma}_\tau}{\sqrt{N}}, \hat{\tau}_N + 1.96 \times \frac{\hat{\sigma}_\tau}{\sqrt{N}}\right),$$

which uses the variance estimator that does not take account of estimation error in the propensity score. This confidence interval is not valid in general, and we report coverage rates for it to assess the problems with ignoring estimation of the propensity score. The second confidence interval is

$$\left(\hat{\tau}_N - 1.96 \times \frac{\hat{\sigma}_{\text{adj},\tau}}{\sqrt{N}}, \hat{\tau}_N + 1.96 \times \frac{\hat{\sigma}_{\text{adj},\tau}}{\sqrt{N}}\right),$$

which does take account of the estimation error in the propensity score.

We use two simple Monte Carlo designs. In both designs there are two covariates, X_1 and X_2, both uniformly distributed on $[-1/2, 1/2]$ and independent of each other. In Design I, the two potential outcomes are generated by $Y(0) = 3X_1 - 3X_2 + U_0$ and $Y(1) = 5 + 5X_1 + X_2 + U_1$, and U_0 and U_1 are independent standard Normal random variables, independent of (W, X_1, X_2). The treatment variable, W, is related to (X_1, X_2) through the propensity score, which is logistic

$$\Pr(W = 1|X_1 = x_1, X_2 = x_2) = \frac{\exp(x_1 + 2x_2)}{1 + \exp(x_1 + 2x_2)}.$$

In this design the standard deviation of the estimator for the average effect τ, based on matching on the true propensity score, is equal to 0.056 (row 1 in the table). Estimating the standard error that does not account for the estimation error in the propensity score, $\sigma_\tau/\sqrt{N}$ gives on average 0.055 (row 2), centered almost exactly at the actual standard deviation. Matching on the estimated propensity score leads to a smaller standard deviation, 0.046, (row 3 in the table). Taking into account the estimation error in the propensity score leads to the adjusted standard error $\sigma_{\text{adj},\tau}/\sqrt{N}$. The standard error is centered at 0.045 (row 4), centered close to the true standard deviation. The last two rows of the table give the coverage rates for the two nominal 95% confidence intervals. First we use the interval $(\hat{\tau}_N - 1.96 \times \hat{\sigma}_\tau/\sqrt{N}, \hat{\tau}_N + 1.96 \times \hat{\sigma}_\tau/\sqrt{N})$, which does not take account of the estimation error in the propensity score (row 5). This leads to overcoverage, with actual coverage equal to 0.983. Second we use the interval $(\hat{\tau}_N - 1.96 \times \hat{\sigma}_{\text{adj},\tau}/\sqrt{N}, \hat{\tau}_N + 1.96 \times \hat{\sigma}_{\text{adj},\tau}/\sqrt{N})$, which does take account of the estimation error in the propensity score (row 6). This leads to an actual coverage rate of 0.949, very close to the nominal level of 0.950. The second column presents the corresponding results for the average effect for the treated under the same design.

In Design II, the two potential outcomes are generated by $Y(0) = 10X_1 + U_0$ and $Y(1) = 5 - 10X_1 + U_1$, and U_0 and U_1 are independent standard Normal random variables, independent of (W, X_1, X_2). The treatment variable, W, is related to (X_1, X_2) through the propensity score, which is logistic

$$\Pr(W = 1|X_1 = x_1, X_2 = x_2) = \frac{\exp(2x_2)}{1 + \exp(2x_2)}.$$

This is a special design in the sense that one can show that the adjustment to the variance for $\hat{\tau}$ when matching on the estimated propensity score vanishes ($c = (0, 0, 0)'$). As a result both the adjusted and unadjusted variance estimator for $\hat{\tau}$ perform well. When we look at the average effect for the treated, however, the adjustment is important ($\frac{\partial \tau_t(\theta^*)}{\partial \theta} \neq 0$), and leads to an increase in the variance. This shows up in the actual standard deviation (0.116 for matching on the true propensity score and 0.141 for matching on the estimated propensity score) as well as in the standard errors ($\hat{\sigma}_{\tau,t}$ is on average 0.116 and $\hat{\sigma}_{\text{adj},\tau,t}$ is on average 0.141). It also shows up in the unadjusted confidence intervals having coverage 0.896, less than the nominal level of 0.95.

VI. Conclusions

In this article, we derive the large sample distribution of propensity score for matching estimators for the case where the propensity score is unknown and needs to be estimated in a first step prior to matching. We show that first step estimation of the propensity score generally affects the asymptotic variance of matching estimators, and derive adjustments for propensity score matching estimators of ATE and ATET. These results allow, for the first time, valid large sample inference for estimators that use matching on the estimated propensity score.

Appendix

Before proving Proposition 2, we introduce some additional notation. Using the definition of $K_{M,\theta}(i)$ in (16), The estimator $\widehat{\tau}_N(\theta)$ can be written as:

$$\widehat{\tau}_N(\theta) = \frac{1}{N}\sum_{i=1}^{N}(2W_i - 1)\left(1 + \frac{K_{M,\theta}(i)}{M}\right) Y_i.$$

Define

$$\begin{aligned} D_N(\theta) = &\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Big(\bar{\mu}(1, F(X_i'\theta)) - \bar{\mu}(0, F(X_i'\theta)) - \tau\Big) \\ &+ \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(2W_i - 1)\left(1 + \frac{K_{M,\theta}(i)}{M}\right)\Big(Y_i - \bar{\mu}(W_i, F(X_i'\theta))\Big), \end{aligned}$$

and

$$R_N(\theta) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(2W_i - 1)\Big(\bar{\mu}(1 - W_i, F(X_i'\theta)) - \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)}\bar{\mu}(1 - W_i, F(X_i'\theta))\Big).$$

Now the normalized estimator can be written as the sum of two parts, the first $D_N(\theta)$ a variance component, and the second, $R_N(\theta)$, a one bias component:

$$\sqrt{N}\big(\widehat{\tau}_N(\theta) - \tau\big) = D_N(\theta) + R_N(\theta).$$

Proof of Proposition 2: The conditions for Lemma 5 in the additional appendix to this paper (Abadie and Imbens, 2011) hold uniformly in θ, for $|\theta - \theta^*| \leq \epsilon$, so that the results in Lemma 5 imply that $R_N(\theta_N) \xrightarrow{p} 0$. Therefore, in order to prove the result in the proposition it suffices to prove that, under P^{θ_N}

$$\begin{pmatrix} D_N(\theta_N) \\ \sqrt{N}(\widehat{\theta}_N - \theta_N) \\ \Lambda_N(\theta^*|\theta_N) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \\ -h'I_{\theta^*}h/2 \end{pmatrix}, \begin{pmatrix} \sigma^2 & c'I_{\theta^*}^{-1} & -c'h \\ I_{\theta^*}^{-1}c & I_{\theta^*}^{-1} & -h \\ -h'c & -h' & h'I_{\theta^*}h \end{pmatrix}\right).$$

By Assumption 7, under P^{θ_N}

$$\Lambda_N(\theta^*|\theta_N) = -h'\Delta_N(\theta_N) - \frac{1}{2}h'I_{\theta^*}h + o_p(1),$$

and

$$\sqrt{N}(\widehat{\theta}_N - \theta_N) = I_{\theta^*}^{-1}\Delta_N(\theta_N) + o_p(1).$$

Therefore, it suffices to prove that, under P^{θ_N}:

$$\begin{pmatrix} D_N(\theta_N) \\ \Delta_N(\theta_N) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & c' \\ c & I_{\theta^*} \end{pmatrix}\right). \tag{A.1}$$

To prove A.1 we extend the martingale representation of matching estimators (Abadie and Imbens, 2010) to allow for estimation of the propensity score. Consider the linear combination $C_N = z_1(D_{N,1}(\theta_N) + D_{N,2}(\theta_N)) + z_2'\Delta_N(\theta_N)$:

$$\begin{aligned} C_N &= z_1 \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Big(\bar{\mu}(1, F(X_{N,i}'\theta_N)) - \bar{\mu}(0, F(X_{N,i}'\theta_N)) - \tau \Big) \\ &+ z_1 \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (2W_{N,i} - 1) \left(1 + \frac{K_{M,\theta_N}(i)}{M} \right) \Big(Y_{N,i} - \bar{\mu}(W_{N,i}, F(X_{N,i}'\theta_N)) \Big) \\ &+ z_2' \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_{N,i} \frac{W_{N,i} - F(X_{N,i}'\theta_N)}{F(X_{N,i}'\theta_N)(1 - F(X_{N,i}'\theta_N))} f(X_{N,i}'\theta_N). \end{aligned}$$

We analyze C_N using martingale methods. First:

$$C_N = \sum_{k=1}^{3N} \xi_{N,k},$$

where

$$\begin{aligned} \xi_{N,k} &= z_1 \frac{1}{\sqrt{N}} \Big(\bar{\mu}(1, F(X_{N,k}'\theta_N)) - \bar{\mu}(0, F(X_{N,k}'\theta_N)) - \tau \Big) \\ &+ z_2' \frac{1}{\sqrt{N}} E_{\theta_N}[X_{N,k} \,|\, F(X_{N,k}'\theta_N)] \frac{W_{N,k} - F(X_{N,k}'\theta_N)}{F(X_{N,k}'\theta_N)(1 - F(X_{N,k}'\theta_N))} f(X_{N,k}'\theta_N), \end{aligned}$$

for $1 \leq k \leq N$,

$$\begin{aligned} \xi_{N,k} &= z_2' \frac{1}{\sqrt{N}} \left(X_{N,k-N} - E_{\theta_N}[X_{N,k-N} \,|F(X_{N,k-N}'\theta_N)] \right) \frac{(W_{N,k-N} - F(X_{N,k-N}'\theta_N)) f(X_{N,k-N}'\theta_N)}{F(X_{N,k-N}'\theta_N)(1 - F(X_{N,k-N}'\theta_N))} \\ &+ z_1 \frac{1}{\sqrt{N}} (2W_{N,k-N} - 1) \left(1 + \frac{K_{M,\theta_N}(k-N)}{M} \right) \Big(\mu(W_{N,k-N}, X_{N,k-N}) - \bar{\mu}(W_{N,k-N}, F(X_{N,k-N}'\theta_N)) \Big). \end{aligned}$$

for $N+1 \leq k \leq 2N$,

$$\xi_{N,k} = z_1 \frac{1}{\sqrt{N}} (2W_{N,k-2N} - 1) \left(1 + \frac{K_{M,\theta_N}(k-2N)}{M} \right) \Big(Y_{N,k-2N} - \mu(W_{N,k-2N}, X_{N,k-2N}) \Big),$$

for $2N+1 \leq k \leq 3N$. Consider the σ-fields $\mathcal{F}_{N,k} = \sigma\{W_{N,1}, \ldots, W_{N,k}, X_{N,1}'\theta_N, \ldots, X_{N,k}'\theta_N\}$ for $1 \leq k \leq N$, $\mathcal{F}_{N,k} = \sigma\{W_{N,1}, \ldots, W_{N,N}, X_{N,1}'\theta_N, \ldots, X_{N,N}'\theta_N, X_{N,1}, \ldots, X_{N,k-N}\}$ for $N+1 \leq k \leq 2N$, and $\mathcal{F}_{N,k} = \sigma\{W_{N,1}, \ldots, W_{N,N}, X_{N,1}, \ldots, X_{N,N}, Y_{N,1}, \ldots, Y_{N,k-N}\}$ for $2N+1 \leq k \leq 3N$. Then,

$$\left\{ \sum_{j=1}^{i} \xi_{N,j}, \mathcal{F}_{N,i}, 1 \leq i \leq 3N \right\}$$

is a martingale for each $N \geq 1$. Therefore, the limiting distribution of C_N can be studied using Martingale Central Limit Theorem (e.g., Theorem 35.12 in Billingsley (1995), p. 476; importantly, notice that this theorem allows that the probability space varies with N). Because $Y_{N,i}$, $X_{N,i}$,

and $W_{N,i}$ are bounded random variables (uniformly in N), and because $K_{M,\theta}(i)$ has uniformly bounded moments (see Abadie and Imbens, 2009), it follows that:

$$\sum_{k=1}^{3N} E\left[\left|\xi_{N,k}\right|^{2+\delta}\right] \to 0 \qquad \text{for some } \delta > 0.$$

Lindeberg's condition in Billingsley's theorem follows easily from the last equation (Lyapounov's condition). As a result, we obtain that under P^{θ_N}

$$C_N \xrightarrow{d} N\left(0, \sigma_1^2 + \sigma_2^2 + \sigma_3^2\right),$$

where

$$\begin{aligned}
\sigma_1^2 &= \text{plim} \sum_{k=1}^{N} E_{\theta_N}\left[\xi_{N,k}^2 \middle|\, \mathcal{F}_{N,k-1}\right], \\
\sigma_2^2 &= \text{plim} \sum_{k=N+1}^{2N} E_{\theta_N}\left[\xi_{N,k}^2 \middle|\, \mathcal{F}_{N,k-1}\right], \\
\sigma_3^2 &= \text{plim} \sum_{k=2N+1}^{3N} E_{\theta_N}\left[\xi_{N,k}^2 \middle|\, \mathcal{F}_{N,k-1}\right].
\end{aligned}$$

After some algebra, we obtain:

$$\begin{aligned}
\sigma_1^2 &= z_1^2\, E\Big[\big(\bar{\mu}(1, F(X'\theta^*)) - \bar{\mu}(0, F(X'\theta^*)) - \tau\big)^2\Big] \\
&+ z_2' E\left[\frac{f^2(X'\theta^*)}{F(X'\theta^*)(1 - F(X'\theta^*))}\, E[X \mid F(X'\theta^*)] E[X' \mid F(X'\theta^*)]\right] z_2.
\end{aligned}$$

Following the calculations in the additional appendix (Abadie and Imbens (2011) for the expectation of $(1 + K_{M,\theta^*}(i)/M)^2$:

$$\begin{aligned}
\sigma_2^2 &= z_2' E\left[\frac{f^2(X'\theta^*)}{F(X'\theta^*)(1 - F(X'\theta^*))}\, \text{var}(X \mid F(X'\theta^*))\right] z_2 \\
&+ z_1^2 E\left[\frac{\text{var}(\mu(1, X)|F(X'\theta^*))}{F(X'\theta^*)} + \frac{\text{var}(\mu(0, X)|F(X'\theta^*))}{1 - F(X'\theta^*)}\right] \\
&+ z_1^2 \frac{1}{2M} E\left[\left(\frac{1}{F(X'\theta^*)} - F(X'\theta^*)\right) \text{var}(\mu(1, X)|F(X'\theta^*))\right] \\
&+ z_1^2 \frac{1}{2M} E\left[\left(\frac{1}{1 - F(X'\theta^*)} - (1 - F(X'\theta^*))\right) \text{var}(\mu(0, X)|F(X'\theta^*))\right] \\
&+ 2\, z_2' E\left[\text{cov}(X, \mu(W, X)|F(X'\theta^*), W) \frac{f(X'\theta^*)}{F(X'\theta^*)(1 - F(X'\theta^*))}\right] z_1.
\end{aligned}$$

Here we use the fact that, conditional on the propensity score, X is independent of W. To derive the constant vector of the cross-product notice that:

$$E\left[\text{cov}\big(X, \mu(X, W)\big|F(X'\theta^*), W\big) \frac{(W - F(X'\theta^*))(2W - 1)}{F(X'\theta^*)(1 - F(X'\theta^*))} f(X'\theta^*) \left(1 + \frac{K_N(\theta^*)}{M}\right)\right]$$

$$
\begin{aligned}
&= E\left[\operatorname{cov}\big(X,\mu(X,1)\big|F(X'\theta^*)\big)\frac{f(X'\theta^*)}{F(X'\theta^*)}\left(1+\frac{K_N(\theta^*)}{M}\right)\Big|W=1\right]p\\
&+ E\left[\operatorname{cov}\big(X,\mu(X,0)\big|F(X'\theta^*)\big)\frac{f(X'\theta^*)}{1-F(X'\theta^*)}\left(1+\frac{K_N(\theta^*)}{M}\right)\Big|W=0\right](1-p)\\
&\to E\left[\operatorname{cov}\big(X,\mu(X,1)\big|F(X'\theta^*)\big)\frac{f(X'\theta^*)}{F(X'\theta^*)^2}\Big|W=1\right]p\\
&+ E\left[\operatorname{cov}\big(X,\mu(X,0)\big|F(X'\theta^*)\big)\frac{f(X'\theta^*)}{(1-F(X'\theta^*))^2}\Big|W=0\right](1-p)\\
&= E\left[\operatorname{cov}(X,\mu(W,X)|F(X'\theta^*),W)f(X'\theta^*)\left(\frac{W}{F(X'\theta^*)^2}+\frac{1-W}{(1-F(X'\theta^*))^2}\right)\right]\\
&= E\left[\operatorname{cov}\left(X,Y\Big|F(X'\theta^*),W\right)f(X'\theta^*)\left(\frac{W}{F(X'\theta^*)^2}+\frac{1-W}{(1-F(X'\theta^*))^2}\right)\right]=c.
\end{aligned}
$$

Finally,

$$
\begin{aligned}
\sigma_3^2 &= z_1^2\,E\left[\frac{\operatorname{var}(Y|X,W=1)}{F(X'\theta^*)}+\frac{\operatorname{var}(Y|X,W=0)}{1-F(X'\theta^*)}\right]\\
&+ z_1^2\frac{1}{2M}E\left[\left(\frac{1}{F(X'\theta^*)}-F(X'\theta^*)\right)\operatorname{var}(Y|X,W=1)\right]\\
&+ z_1^2\frac{1}{2M}E\left[\left(\frac{1}{1-F(X'\theta^*)}-(1-F(X'\theta^*))\right)\operatorname{var}(Y|X,W=0)\right].
\end{aligned}
$$

Notice that for any integrable function $g(F(X'\theta^*))$:

$$
\begin{aligned}
&E\Big[g(F(X'\theta^*))\Big(\operatorname{var}(\mu(w,X)|F(X'\theta^*))+\operatorname{var}(Y|X,W=w)\Big)\Big]\\
&\quad= E\Big[g(F(X'\theta^*))\Big(\operatorname{var}(\mu(w,X)|F(X'\theta^*))+E\Big[\operatorname{var}(Y|X,W=w)\Big|F(X'\theta^*)\Big]\Big)\Big]\\
&= E\Big[g(F(X'\theta^*))\Big(\operatorname{var}(\mu(w,X)|F(X'\theta^*),W=w)+E\Big[\operatorname{var}(Y|X,W=w)\Big|F(X'\theta^*),W=w\Big]\Big)\Big]\\
&\quad= E\Big[g(F(X'\theta^*))\operatorname{var}(Y|F(X'\theta^*),W=w)\Big].
\end{aligned}
$$

Collecting the terms in $\sigma_1^2+\sigma_2^2+\sigma_3^2$, we find that the quadratic form in z_1 is

$$
\begin{aligned}
&z_1^2\,E\Big[\big(\bar\mu(1,F(X'\theta^*))-\bar\mu(0,F(X'\theta^*))-\tau\big)^2\Big]\\
&+z_1^2E\left[\frac{\operatorname{var}(\mu(1,X)|F(X'\theta^*))}{F(X'\theta^*)}+\frac{\operatorname{var}(\mu(0,X)|F(X'\theta^*))}{1-F(X'\theta^*)}\right]\\
&+z_1^2\frac{1}{2M}E\left[\left(\frac{1}{F(X'\theta^*)}-F(X'\theta^*)\right)\operatorname{var}(\mu(1,X)|F(X'\theta^*))\right]\\
&+z_1^2\frac{1}{2M}E\left[\left(\frac{1}{1-F(X'\theta^*)}-(1-F(X'\theta^*))\right)\operatorname{var}(\mu(0,X)|F(X'\theta^*))\right]\\
&z_1^2\,E\left[\frac{\operatorname{var}(Y|X,W=1)}{F(X'\theta^*)}+\frac{\operatorname{var}(Y|X,W=0)}{1-F(X'\theta^*)}\right]
\end{aligned}
$$

$$
\begin{aligned}
&+z_1^2 \frac{1}{2M} E\left[\left(\frac{1}{F(X'\theta^*)} - F(X'\theta^*)\right) \mathrm{var}(Y|X, W=1)\right] \\
&+z_1^2 \frac{1}{2M} E\left[\left(\frac{1}{1-F(X'\theta^*)} - (1-F(X'\theta^*))\right) \mathrm{var}(Y|X, W=0)\right] \\
&= z_1^2 E\left[(\bar{\mu}(1, p(X)) - \bar{\mu}(0, p(X)) - \tau)^2\right] \\
&+z_1^2 E\left[\bar{\sigma}^2(1, p(X)) \left(\frac{1}{p(X)} + \frac{1}{2M}\left(\frac{1}{p(X)} - p(X)\right)\right)\right] \\
&+z_1^2 E\left[\bar{\sigma}^2(0, p(X)) \left(\frac{1}{1-p(X)} + \frac{1}{2M}\left(\frac{1}{1-p(X)} - (1-p(X))\right)\right)\right] \\
&= \sigma^2,
\end{aligned}
$$

by Proposition 1 and Equation (A.3). Collecting the quadratic form in z_2, we find

$$
\begin{aligned}
&z_2' E\left[\frac{f^2(X'\theta^*)}{F(X'\theta^*)(1-F(X'\theta^*))} E[X \,|\, F(X'\theta^*)] E[X' \,|\, F(X'\theta^*)]\right] z_2 \\
&+z_2' E\left[\frac{f^2(X'\theta^*)}{F(X'\theta^*)(1-F(X'\theta^*))} \mathrm{var}(X \,|\, F(X'\theta^*))\right] z_2 \\
&= z_2' I_{\theta^*} z_2.
\end{aligned}
$$

Finally, the cross term is $2z_1 c' z_2$, so that the asymptotic variance of C_N is $z_1^2\sigma^2 + 2z_2 c' z_2 + z_2 I_{\theta^*} z_2$. Hence, by the martingale central limit theorem, under P^{θ_N}:

$$
C_N \xrightarrow{d} N\left(0, \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}' \begin{pmatrix} \sigma^2 & c' \\ c & I_{\theta^*} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}\right).
$$

Applying the Cramer-Wold device, under P^{θ_N}:

$$
\begin{pmatrix} D_N(\theta_N) \\ \Delta_N(\theta_N) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & c' \\ c & I_{\theta^*} \end{pmatrix}\right),
$$

proving A.1 and thus Proposition 2. □

Proof of Theorem 1: Given our preliminary results, Theorem 1 follows from Andreou and Werker (2005). □

Proof of Proposition 3: First, define

$$
\varphi(\theta) = E\left[Y \frac{W - F(X'\theta)}{1 - F(X'\theta)}\right],
$$

$$
P(\theta) = E[F(X'\theta)],
$$

and the estimators

$$
\widehat{\varphi}_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left(W_i - (1-W_i) \cdot \frac{K_{M,\theta}(i)}{M}\right) \cdot Y_i,
$$

and

$$
\widehat{P}_N = \bar{W} = N_1/N.
$$

Note that $P(\theta^*) = E[p(X)] = E[F(X'\theta^*)]$ and $\tau_t(\theta) = \varphi(\theta)/P(\theta)$, and also

$$\widehat{\tau}_t(\theta) = \frac{\widehat{\varphi}(\theta)}{\widehat{P}_N}.$$

Define also

$$a = (1 - P(\theta^*))E\left[F(X'\theta^*)\left(\bar{\mu}(1, F(X'\theta^*)) - \bar{\mu}(0, F(X'\theta^*))\right)\right],$$

$$b = E\left[E[X|F(X'\theta^*)]f(X'\theta^*)\right],$$

$$d = E\left[E[X|F(X'\theta^*)]f(X'\theta^*)(\bar{\mu}(1, F(X'\theta^*)) - \bar{\mu}(0, F(X'\theta^*))\right]$$
$$+E\left[\left(\operatorname{cov}\left(X, \mu(1, X)|F(X'\theta^*)\right) + \frac{F(X'\theta^*)}{1 - F(X'\theta^*)}\operatorname{cov}\left(X, \mu(0, X)|F(X'\theta^*)\right)\right) f(X'\theta^*)\right],$$

$$\sigma_\varphi^2 = E\left[\left(W \cdot \left(\bar{\mu}(1, F(X'\theta^*) - \bar{\mu}(0, F(X'\theta^*)\right) - \varphi(\theta^*)\right)^2\right]$$
$$+E\left[F(X'\theta)\bar{\sigma}^2(1, F(X'\theta^*)) + \frac{F(X'\theta^*)^2}{1 - F(X'\theta^*)}\bar{\sigma}^2(0, F(X'\theta^*))\right]$$
$$+\frac{1}{M} \cdot E\left[\left(F(X'\theta^*) + \frac{F(X'\theta^*)^2}{2(1 - F(X'\theta^*))}\right) \cdot \bar{\sigma}^2(0, F(X'\theta^*))\right].$$

Finally, define

$$D_{t,N}(\theta) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} W_i\Big(\bar{\mu}(1, F(X_i'\theta)) - \bar{\mu}(0, F(X_i'\theta)) - \tau_t(\theta)\Big)$$
$$+ \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(W_i - (1 - W_i) \cdot \frac{K_{M,\theta}(i)}{M}\right) \cdot \left(Y_i - \bar{\mu}(W_i, F(X_i'\theta))\right),$$

and

$$R_{t,N}(\theta) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} W_i\Big(\bar{\mu}(1 - W_i, F(X_i'\theta)) - \frac{1}{M}\sum_{j \in \mathcal{J}_M(i)} \bar{\mu}(1 - W_i, F(X_i'\theta))\Big),$$

so that

$$\sqrt{N}\left(\hat{\varphi}(\theta) - \varphi(\theta)\right) = D_{t,N}(\theta) + R_{t,N}.$$

In the first step we prove that under P^{θ_N},

$$\begin{pmatrix} \sqrt{N}\left(\widehat{P}_N - P(\theta_N)\right) \\ D_{t,N}(\theta_N) \\ \Delta_N(\theta_N) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} P(\theta^*)(1 - P(\theta^*)) & a & b' \\ & \sigma_\varphi^2 & d' \\ & & I_{\theta^*} \end{pmatrix}\right). \tag{A.2}$$

In the second step we prove that (A.2) implies the claim in the lemma that under P^{θ_N}:

$$\begin{pmatrix} \sqrt{N}\left(\widehat{\tau}_{t,N}(\theta_N) - \tau_t(\theta_N)\right) \\ \sqrt{N}(\widehat{\theta}_N - \theta_N) \\ \Lambda_N(\theta^*|\theta_N) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \\ -h'I_{\theta^*}h/2 \end{pmatrix}, \begin{pmatrix} \sigma_t^2 & c_t'I_{\theta^*}^{-1} & -c_t'h \\ I_{\theta^*}^{-1}c_t & I_{\theta^*}^{-1} & -h \\ -c_t'h & -h & h'I_{\theta^*}h \end{pmatrix}\right). \tag{A.3}$$

where c_t is as defined before in (13).

First we prove (A.2). Here we use the martingale representation we also used in the proof of Lemma 2. Define

$$C_N = z_1\sqrt{N}\big(\widehat{P}_N - P(\theta_N)\big) + z_2 D_{t,N}(\theta_N) + z_3'\Delta_N(\theta_N)$$

$$= z_1\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(W_{N,i} - P(\theta_N)\right)$$

$$+z_2\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Big(W_{N,i}\left(\bar{\mu}(1,F(X'_{N,i}\theta_N)) - \bar{\mu}(0,F(X'_{N,i}\theta_N))\right) - \varphi(\theta_N)\Big)$$

$$+z_2\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(W_{N,i} - (1-W_{N,i})\frac{K_{M,\theta}(i)}{M}\right)\left(Y_{N,i} - \bar{\mu}(W_{N,i},\bar{\mu}(W_{N,i},F(X'_{N,i}\theta_N))\right)$$

$$+z_3'\frac{1}{\sqrt{N}}\sum_{i=1}^{N}X_{N,i}\frac{W_{N,i} - F(X'_{N,i}\theta_N)}{F(X'_{N,i}\theta_N)(1-F(X'_{N,i}\theta_N))}f(X'_{N,i}\theta).$$

Notice that

$$C_N = \sum_{k=1}^{3N}\xi_{N,k},$$

where

$$\begin{aligned}
\xi_{N,k} &= z_1\frac{1}{\sqrt{N}}\left(W_{N,k} - P(\theta_N)\right)\\
&+ z_2\frac{1}{\sqrt{N}}W_{N,k}\Big(\bar{\mu}(1,F(X'_{N,k}\theta_N)) - \bar{\mu}(0,F(X'_{N,k}\theta_N))\Big) - \varphi(\theta_N)\\
&+ z_3'\frac{1}{\sqrt{N}}E_{\theta_N}[X_{N,k}\,|\,F(X'_{N,k}\theta_N)]\frac{W_{N,k} - F(X'_{N,k}\theta_N)}{F(X'_{N,k}\theta_N)(1-F(X'_{N,k}\theta_N))}f(X'_{N,k}\theta_N),
\end{aligned}$$

for $1 \le k \le N$,

$$\begin{aligned}
\xi_{N,k} &= z_3'\frac{1}{\sqrt{N}}\left(X_{N,k-N} - E_{\theta_N}[X_{N,k-N}\,|F(X'_{N,k-N}\theta_N)]\right)\frac{(W_{N,k-N} - F(X'_{N,k-N}\theta_N))f(X'_{N,k-N}\theta_N)}{F(X'_{N,k-N}\theta_N)(1-F(X'_{N,k-N}\theta_N))}\\
&+ z_2\frac{1}{\sqrt{N}}\left(W_{N,k-N} - (1-W_{N,k-N})\frac{K_{M,\theta_N}(k-N)}{M}\right)\Big(\mu(W_{N,k-N},X_{N,k-N}) - \bar{\mu}(W_{N,k-N},F(X'_{N,k-N}\theta_N))\Big),
\end{aligned}$$

for $N+1 \le k \le 2N$,

$$\xi_{N,k} = z_2\frac{1}{\sqrt{N}}\left(W_{N,k-2N} - (1-W_{N,k-2N})\frac{K_{M,\theta_N}(k-2N)}{M}\right)\Big(Y_{N,k-2N} - \mu(W_{N,k-2N},X_{N,k-2N})\Big),$$

for $2N+1 \le k \le 3N$. Consider the σ-fields $\mathcal{F}_{N,k} = \sigma\{W_{N,1},\ldots,W_{N,k},X'_{N,1}\theta_N,\ldots,X'_{N,k}\theta_N\}$ for $1 \le k \le N$, $\mathcal{F}_{N,k} = \sigma\{\mathbf{W}_N, X'_{N,1}\theta_N,\ldots,X'_{N,N}\theta_N, X_{N,1},\ldots,X_{N,k-N}\}$ for $N+1 \le k \le 2N$, and $\mathcal{F}_{N,k} = \sigma\{\mathbf{W}_N,\mathbf{X}_N,Y_{N,1},\ldots,Y_{N,k-2N}\}$ for $2N+1 \le k \le 3N$. Then,

$$\left\{\sum_{j=1}^{i}\xi_{N,j},\mathcal{F}_{N,i},1 \le i \le 3N\right\}$$

is a martingale for each $N \geq 1$. Therefore, the limiting distribution of C_N is given by the Martingale CLT (e.g., Theorem 35.12 in Billingsley, p. 476; importantly, notice that this theorem allows that the probability space varies with N): under P^{θ_N}

$$C_N \xrightarrow{d} N\left(0, \sigma_1^2 + \sigma_2^2 + \sigma_3^2\right),$$

where

$$\sigma_1^2 = \text{plim} \sum_{k=1}^{N} E_{\theta_N}[\xi_{N,k}^2 | \, \mathcal{F}_{N,k-1}],$$

$$\sigma_2^2 = \text{plim} \sum_{k=N+1}^{2N} E_{\theta_N}[\xi_{N,k}^2 | \, \mathcal{F}_{N,k-1}],$$

and

$$\sigma_3^2 = \text{plim} \sum_{k=2N+1}^{3N} E_{\theta_N}[\xi_{N,k}^2 | \, \mathcal{F}_{N,k-1}].$$

After some algebra, we obtain:

$$\begin{aligned}
\sigma_1^2 &= z_1^2 \, P(\theta^*)(1 - P(\theta^*)) \\
&+ z_2^2 \, E\Big[\Big(W\big(\bar{\mu}(1, F(X'\theta^*)) - \bar{\mu}(0, F(X'\theta^*))\big) - \varphi(\theta^*)\Big)^2\Big] \\
&+ z_3' E\left[\frac{f^2(X'\theta^*)}{F(X'\theta^*)(1 - F(X'\theta^*))} \, E[X \,|\, F(X'\theta^*)] E[X' \,|\, F(X'\theta^*)]\right] z_3 \\
&+ 2 z_1 \, z_2 (1 - P(\theta^*)) E\Big[F(X'\theta^*)\big(\bar{\mu}(1, F(X'\theta^*)) - \bar{\mu}(0, F(X'\theta^*))\big)\Big] \\
&+ 2 z_3' E\left[E[X \,|\, F(X'\theta^*)] f(X'\theta^*)\right] z_1 \\
&+ 2 z_3' E\left[E[X \,|\, F(X'\theta^*)] f(X'\theta^*)\big(\bar{\mu}(1, F(X'\theta^*)) - \bar{\mu}(0, F(X'\theta^*))\big)\right] z_2,
\end{aligned}$$

$$\begin{aligned}
\sigma_2^2 &= z_3' E\left[\frac{f^2(X'\theta^*)}{F(X'\theta^*)(1 - F(X'\theta^*))} \, \text{var}(X \,|\, F(X'\theta^*))\right] z_3 \\
&+ z_2^2 E\left[F(X'\theta^*) \, \text{var}(\mu(1, X) | F(X'\theta^*)) + \left(\frac{F(X'\theta^*)}{M} + \frac{2M+1}{2M} \frac{F^2(X'\theta^*)}{1 - F(X'\theta^*)}\right) \text{var}(\mu(0, X) | F(X'\theta^*))\right] \\
&+ 2 \, z_2 E\left[f(X'\theta^*)\left(\text{cov}\big(\mu(1, X), X' \big| F(X'\theta^*)\big) + \frac{F(X'\theta^*)}{1 - F(X'\theta^*)} \, \text{cov}\big(\mu(0, X), X' \big| F(X'\theta^*)\big)\right)\right] z_3,
\end{aligned}$$

and

$$\sigma_3^2 = z_2^2 \, E\left[F(X'\theta^*) \, \text{var}(Y | X, W = 1) + \left(\frac{F(X'\theta^*)}{M} + \frac{2M+1}{2M} \frac{F^2(X'\theta^*)}{1 - F(X'\theta^*)}\right) \text{var}(Y | X, W = 0)\right].$$

Notice that

$$\text{var}(Y | F(X'\theta^*), W) = E[\text{var}(Y | X, W) | F(X'\theta^*), W] + \text{var}(\mu(W, X) | F(X'\theta^*), W).$$

As a result, under P^{θ_N}, for arbitrary $z = (z_1, z_2, z_3')'$,

$$C_N \xrightarrow{d} N\left(0, \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix}' \begin{pmatrix} P(\theta^*)(1-P(\theta^*)) & a & b' \\ a & \sigma_\varphi^2 & d' \\ b & d & I_\theta \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix}\right).$$

Applying the Cramer-Wold device, this implies that under P^{θ_N}:

$$\begin{pmatrix} \sqrt{N}(\widehat{P}_N - P(\theta_N)) \\ D_{t,N}(\theta_N) \\ \Delta_N(\theta_N) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} P(\theta^*)(1-P(\theta^*)) & a & b' \\ a & \sigma_\varphi^2 & d' \\ b & d & I_\theta \end{pmatrix}\right),$$

finishing the proof of (A.2).
Now we consider the second step, that (A.2) implies (A.3). Using the definitions of $D_{N,t,1}(\theta)$, $D_{N,t,2}(\theta)$ and $R_{t,N}(\theta)$, we can write the estimator $\hat{\varphi}(\theta)$ as

$$\sqrt{N}\big(\widehat{\varphi}_N(\theta) - \varphi(\theta)\big) = D_{t,N}(\theta) + R_{t,N}(\theta).$$

The conditions for Lemma 5 in the additional appendix to this paper (Abadie and Imbens, 2011) hold uniformly in θ, for $|\theta - \theta^*| \leq \epsilon$, so that the results in Lemma 5 imply that $R_{t,N}(\theta_N) \xrightarrow{p} 0$. Therefore, (A.2) implies that under P^{θ_N}:

$$\begin{pmatrix} \sqrt{N}(\widehat{P}_N - P(\theta_N)) \\ \sqrt{N}(\widehat{\varphi}_N(\theta_N) - \varphi(\theta_N)) \\ \Delta_N(\theta_N) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} P(\theta^*)(1-P(\theta^*)) & a & b' \\ a & \sigma_\varphi^2 & d' \\ b & d & I_\theta \end{pmatrix}\right). \tag{A.4}$$

Using Assumption 7,

$$\Lambda_N(\theta^*|\theta_N) = -h'\Delta_N(\theta_N) - \frac{1}{2}h' I_{\theta^*} h + o_p(1),$$

and

$$\sqrt{N}(\widehat{\theta}_N - \theta_N) = I_{\theta^*}^{-1}\Delta_N(\theta_N) + o_p(1),$$

it follows that (A.4) implies that, under P^{θ_N},

$$\begin{pmatrix} \sqrt{N}\big(\widehat{P}_N(\theta_N) - P(\theta_N)\big) \\ \widehat{\varphi}_N(\theta_N) - \varphi(\theta_N) \\ \sqrt{N}(\widehat{\theta}_N - \theta_N) \\ \Lambda_N(\theta^*|\theta_N) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ h' I_{\theta^*} h/2 \end{pmatrix}, \begin{pmatrix} P(\theta^*)(1-P(\theta^*)) & a & b' I_{\theta^*}^{-1} & -b'h \\ & \sigma_\varphi^2 & d' I_{\theta^*}^{-1} & -d'h \\ & & I_{\theta^*}^{-1} & -h \\ & & & h' I_{\theta^*} h \end{pmatrix}\right). \tag{A.5}$$

Note that under P^θ,

$$\sqrt{N}\big(\widehat{\tau}_{t,N}(\widehat{\theta}_N) - \tau_t(\theta)\big) = \frac{1}{P(\theta)}\sqrt{N}\big(\widehat{\varphi}_N(\widehat{\theta}_N) - \varphi(\theta)\big) - \frac{\varphi(\theta)}{P(\theta)^2}\sqrt{N}\big(\widehat{P}_N(\widehat{\theta}_N) - P(\theta)\big) + o_p(1),$$

so that (A.5) implies that, under P^{θ_N},

$$\begin{pmatrix} \sqrt{N}\big(\widehat{\tau}_{t,N}(\theta_N) - \tau_t(\theta_N)\big) \\ \sqrt{N}(\widehat{\theta}_N - \theta_N) \\ \Lambda_N(\theta^*|\theta_N) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \\ -h' I_{\theta^*} h/2 \end{pmatrix},\right.$$

$$\left(\begin{array}{cc} \frac{1}{P(\theta^*)^2}\sigma_\varphi^2 - 2\frac{\varphi(\theta^*)}{P(\theta^*)^2}a + \frac{\varphi(\theta^*)^2}{P(\theta^*)^2}P(\theta^*)(1-P(\theta^*)) & \left(\frac{1}{P(\theta^*)}d - \frac{\tau}{P(\theta^*)}b\right)' I_{\theta^*}^{-1} \quad -\left(\frac{1}{P(\theta^*)}d - \frac{\tau}{P(\theta^*)}b\right)' h \\ & I_{\theta^*}^{-1} \qquad\qquad -h \\ & \qquad\qquad\qquad\quad h' I_{\theta^*} h \end{array} \right). \tag{A.6}$$

Thus, to complete the proof of the second step we need to show that the two covariance matrices in (A.3) and (A.6) are equal, which follows if the following two equations hold:

$$\sigma_t^2 = \frac{1}{P(\theta^*)^2}\sigma_\varphi^2 - 2\frac{\varphi(\theta^*)}{P(\theta^*)^3}a + \frac{\varphi(\theta^*)^2}{P(\theta^*)^4}P(\theta^*)(1-P(\theta^*)) \tag{A.7}$$

and

$$c_t = \frac{1}{P(\theta^*)}d - \frac{\tau}{P(\theta^*)}b. \tag{A.8}$$

To prove (A.7) we substitute the definitions of σ_t^2, σ_φ^2, a, and $\varphi(\theta)$ into the equality, so that checking (A.7) is equivalent to checking that

$$\begin{aligned} &\frac{1}{P(\theta^*)}E\left[F(X'\theta^*)\left(\bar{\mu}(1,F(X'\theta^*)) - \bar{\mu}(0,F(X'\theta^*)) - \tau_t\right)^2\right] \\ &\quad + \frac{1}{P(\theta^*)}E\left[\bar{\sigma}^2(1,F(X'\theta^*))F(X'\theta^*)\right] \\ &\quad + \frac{1}{P(\theta^*)}E\left[\bar{\sigma}^2(0,F(X'\theta^*))\left(\frac{F(X'\theta^*)^2}{1-F(X'\theta^*)} + \frac{1}{M}F(X'\theta^*) + \frac{1}{2M}\frac{F(X'\theta^*)^2}{1-F(X'\theta^*)}\right)\right] \\ &= \frac{1}{P(\theta^*)^2}E\left[\left(W\cdot\left(\bar{\mu}(1,F(X'\theta^*) - \bar{\mu}(0,F(X'\theta^*)\right) - \varphi(\theta)\right)^2\right] \\ &\quad + \frac{1}{P(\theta^*)^2}E\left[F(X'\theta^*)\bar{\sigma}^2(1,F(X'\theta^*)) + \frac{F(X'\theta^*)^2}{1-F(X'\theta^*)}\bar{\sigma}^2(0,F(X'\theta^*))\right] \\ &\quad + \frac{1}{P(\theta^*)^2}\frac{1}{M}\cdot E\left[\left(F(X'\theta^*) + \frac{F(X'\theta^*)^2}{2(1-F(X'\theta^*))}\right)\cdot\bar{\sigma}^2(0,F(X'\theta^*))\right] \\ &\quad - 2\frac{\tau_t}{P(\theta^*)^2}(1-P(\theta^*))E\left[F(X'\theta^*)\left(\bar{\mu}(1,F(X'\theta^*)) - \bar{\mu}(0,F(X'\theta^*))\right)\right] \\ &\quad + \frac{\tau_t^2}{P(\theta^*)^2}P(\theta^*)(1-P(\theta^*)). \end{aligned}$$

This equality holds if

$$\begin{aligned} &E\left[F(X'\theta^*)\left(\bar{\mu}(1,F(X'\theta^*)) - \bar{\mu}(0,F(X'\theta^*)) - \tau_t\right)^2\right] \\ &= E\left[\left(W\cdot\left(\bar{\mu}(1,F(X'\theta^*) - \bar{\mu}(0,F(X'\theta^*)\right) - \varphi(\theta^*)\right)^2\right] \\ &\quad - 2\tau_t(1-P(\theta^*))E\left[F(X'\theta^*)\left(\bar{\mu}(1,F(X'\theta^*)) - \bar{\mu}(0,F(X'\theta^*))\right)\right] \\ &\quad + \tau_t^2 P(\theta^*)(1-P(\theta^*)). \end{aligned}$$

Taking the difference, and using the fact that

$$\tau_t = \tau_t(\theta^*) = E\left[\frac{1}{P(\theta^*)}F(X'\theta^*)\left(\bar{\mu}(1, F(X'\theta^*)) - \bar{\mu}(1, F(X'\theta^*))\right)\right],$$

we have

$$\begin{aligned}
&E\left[F(X'\theta^*)\left(\bar{\mu}(1, F(X'\theta^*)) - \bar{\mu}(0, F(X'\theta^*)) - \tau_t\right)^2\right] \\
&-E\left[\left(W \cdot \left(\bar{\mu}(1, F(X'\theta^*) - \bar{\mu}(0, F(X'\theta^*)\right) - \varphi(\theta^*)\right)^2\right] \\
&+2\tau_t E\left[F(X'\theta^*)(1 - P(\theta^*))\left(\bar{\mu}(1, F(X'\theta^*)) - \bar{\mu}(0, F(X'\theta^*))\right)\right] \\
&-\tau_t^2 P(\theta^*)(1 - P(\theta^*)). \\
&= E\left[F(X'\theta^*)\left(\bar{\mu}(1, F(X'\theta^*)) - \bar{\mu}(0, F(X'\theta^*))\right)^2\right] \\
&-2\tau_t E\left[F(X'\theta^*)\left(\bar{\mu}(1, F(X'\theta^*)) - \bar{\mu}(0, F(X'\theta^*))\right)\right] + E[F(X'\theta^*)]\tau_t^2 \\
&-E\left[W \cdot \left(\bar{\mu}(1, F(X'\theta^*) - \bar{\mu}(0, F(X'\theta^*)\right)^2\right] \\
&+2E\left[W \cdot \left(\bar{\mu}(1, F(X'\theta^*) - \bar{\mu}(0, F(X'\theta^*)\right)\varphi(\theta)\right] - \varphi(\theta)^2 \\
&+2\tau_t(1 - P(\theta^{)})E\left[F(X'\theta^*)\left(\bar{\mu}(1, F(X'\theta^*)) - \bar{\mu}(0, F(X'\theta^*))\right)\right] \\
&-\tau_t^2 P(\theta^*)(1 - P(\theta^*)) \\
&= -2\tau_t^2 P(\theta^*) + P(\theta^*)\tau_t^2 \\
&+2P(\theta^*)^2\tau_t^2 - P(\theta^*)^2\tau_t^2 \\
&+2\tau_t^2 P(\theta^*)(1 - P(\theta^*)) \\
&-\tau_t^2 P(\theta^*)(1 - P(\theta^*)). \\
&= 0,
\end{aligned}$$

which proves (A.7). Checking (A.8) is equivalent to verifying that

$$\begin{aligned}
&\frac{1}{P(\theta^*)}E\left[Xf(X'\theta^*)\left(\bar{\mu}(1, F(X'\theta^*)) - \bar{\mu}(0, F(X'\theta^*)) - \tau_t\right)\right] \\
&+\frac{1}{P(\theta^*)}E\left[\mathrm{cov}\left(X, Y\middle|F(X'\theta^*), W\right)f(X'\theta^*)\left(\frac{W}{F(X'\theta^*)} + \frac{(1-W)F(X'\theta^*)}{(1 - F(X'\theta^*))^2}\right)\right] \\
&= \frac{1}{P(\theta^*)}E\left[E[X|F(X'\theta^*)]f(X'\theta^*)(\bar{\mu}(1, F(X'\theta^*)) - \bar{\mu}(0, F(X'\theta^*))\right] \\
&+\frac{1}{P(\theta^*)}E\left[\left(\mathrm{cov}\left(X, \mu(1, X)|F(X'\theta^*)\right) + \frac{F(X'\theta^*)}{1 - F(X'\theta^*)}\mathrm{cov}\left(X, \mu(0, X)|F(X'\theta^*)\right)\right)f(X'\theta^*)\right] \\
&-\frac{\tau_t}{P(\theta^*)}E\left[E[X|F(X'\theta^*)]f(X'\theta^*)\right].
\end{aligned}$$

This in turn holds if

$$E\left[\mathrm{cov}\left(X, Y\middle|F(X'\theta^*), W\right)f(X'\theta^*)\left(\frac{W}{F(X'\theta^*)} + \frac{(1-W)F(X'\theta^*)}{(1 - F(X'\theta^*))^2}\right)\right]$$

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

To prove this,

$$E\left[\text{cov}\left(X, Y\middle|F(X'\theta^*), W\right) f(X'\theta^*)\left(\frac{W}{F(X'\theta^*)} + \frac{(1-W)F(X'\theta^*)}{(1-F(X'\theta^*))^2}\right)\right]$$

$$= E\left[\text{cov}\left(X, Y\middle|F(X'\theta^*), W=1\right) f(X'\theta^*)\right]$$

$$+E\left[\text{cov}\left(X, Y\middle|F(X'\theta^*), W=0\right) f(X'\theta^*)\frac{F(X'\theta^*)}{(1-F(X'\theta^*))}\right]$$

$$= E\left[\text{cov}\left(X, \mu(1,X)\middle|F(X'\theta^*), W=1\right) f(X'\theta^*)\right]$$

$$+E\left[\text{cov}\left(X, \mu(0,X)\middle|F(X'\theta^*), W=0\right) f(X'\theta^*)\frac{F(X'\theta^*)}{(1-F(X'\theta^*))}\right]$$

$$= E\left[\text{cov}\left(X, \mu(1,X)\middle|F(X'\theta^*)\right) f(X'\theta^*)\right]$$

$$+E\left[\text{cov}\left(X, \mu(0,X)\middle|F(X'\theta^*)\right) f(X'\theta^*)\frac{F(X'\theta^*)}{(1-F(X'\theta^*))}\right]$$

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

which finishes the proof of (A.8), and thus of the second step, that (A.2) implies (A.3), and therefore finishes the proof of the lemma. □

References

Abadie, A. (2005) "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies*, vol. 72, no. 1, 1-19.

Abadie, A. and Imbens, G.W. (2006) "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica* , vol. 74, no. 1, 235-267.

Abadie, A. and Imbens, G.W. (2008) "On the Failure of the Bootstrap for Matching Estimators," *Econometrica* , vol. 76, no. 6, 1537-1558.

Abadie, A. and Imbens, G.W. (2009) "A Martingale Representation for Matching Estimators," NBER working paper, no. 14756.

Abadie, A. and Imbens, G.W. (2011) "Additional Proofs for 'Matching on the Estimated Propensity Score'," www.hks.harvard.edu/fs/aabadie/,
and http://www.economics.harvard.edu/faculty/imbens/imbens.html.

Andreou, E. and Werker, B.J.M. (forthcoming) "An Alternative Asymptotic Analysis of Residual-Based Statistics," *Review of Economics and Statistics.*

Angrist, J.D. and Kuersteiner, G.M. (forthcoming) "Causal Effects of Monetary Shocks: Semiparametric Conditional Independence Tests with a Multinomial Propensity Score," *Review of Economics and Statistics.*

Bickel, P.J., Klaassen, C.A., Ritov, Y. and Wellner, J.A. (1998) *Efficient and Adaptive Estimation for Semiparametric Models*, Springer, New York.

Billingsley, P. (1995), *Probability and Measure*, third edition. Wiley, New York.

Ganssler, P. and Pfanzagl, J. (1971) "Convergence of Conditional Expectations," *The Annals of Mathematical Statistics*, vol. 42, no. 1, 315-324.

Dehejia, R. and Wahba, S. (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053-1062.

Hahn, J., (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66 (2), 315-331.

Heckman, J., H. Ichimura, and P. Todd, (1997), "Matching as an Econometric Evaluation Estimator: Evidence from a Job Training Program," *Review of Economic Studies* 64, 605-654.

Heckman, J., H. Ichimura, and P. Todd, (1998), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies* 65, 261–294.

Hirano, K., G. Imbens, and G. Ridder (2003) "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, vol. 71, no. 4, 1161-1189.

Imbens, G., (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics and Statistics*, 86(1): 1-29.

Imbens, G., and J. Wooldridge (2009) "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*.

Johnson, N. and Kotz, S. (1977) *Urn Models and Their Applications*, John Wiley & Sons, New York.

Newey, W.K. and McFadden, D. (1994) "Large sample estimation and hypothesis testing." In: Engle, R.F., McFadden, D. (Eds.), *Handbook of Econometrics*, vol. 4. Elsevier Science, Amsterdam.

Lehmann, E.L. and Romano, J.P. (2005) *Testing Statistical Hypothesis*. Springer, New York.

Rosenbaum, P. and Rubin, D.B. (1983) "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, vol. 70, 4155.

Rubin, D.B. (1974) "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology*, vol. 66, 688-701.

Rubin, D., and N. Thomas (1992a) "Characterizing the effect of matching using linear propensity score methods with normal distributions," *Biometrika*, vol. 79, 797-809.

Rubin, D., and N. Thomas (1992b) "Affinely Invariant Matching Methods with Ellipsoidal Distributions," *Annals of Statistics*, vol. 20, no. 2, 1079-1093.

Smith, J., and P. Todd (2005) "Does matching overcome LaLondes critique of nonexperimental estimators?," *Journal of Econometrics*, Vol. 125 305353.

van der Vaart, A. (1998), *Asymptotic Statistics*, Cambridge University Press, New York.

van der Vaart, A.W. and Wellner, J.A. (1996), *Weak Convergence and Empirical Processes*, Springer-Verlag, New York.

Wooldridge, J.M. (2007) "Inverse Probability Weighted Estimation for General Missing Data Problems," *Journal of Econometrics*, vol. 141, 1281-1301.

Simulation Results
($N = 5,000$, Number of simulations $= 50,000$)

	Design I		Design II	
	τ	τ_t	τ	τ_t
Standard Deviation, Matching on True P-Score	0.056	0.066	0.103	0.117
Standard Error, Ignoring Estimation of P-score	0.055	0.065	0.102	0.116
Standard Deviation, Matching on Estimated P-Score	0.046	0.057	0.102	0.141
Standard Error, Accounting for Estimation of P-score	0.045	0.057	0.102	0.141
Coverage Rate 95% Confidence Interval Using Matching on the Estimated Propensity Score				
Ignoring Estimation of P-score	0.983	0.977	0.950	0.893
Accounting for Estimation of P-score	0.949	0.947	0.950	0.950