Econometrica, Vol. 74, No. 1 (January, 2006), 235–267

LARGE SAMPLE PROPERTIES OF MATCHING ESTIMATORS FOR AVERAGE TREATMENT EFFECTS

BY ALBERTO ABADIE AND GUIDO W. IMBENS[1]

Matching estimators for average treatment effects are widely used in evaluation research despite the fact that their large sample properties have not been established in many cases. The absence of formal results in this area may be partly due to the fact that standard asymptotic expansions do not apply to matching estimators with a fixed number of matches because such estimators are highly nonsmooth functionals of the data. In this article we develop new methods for analyzing the large sample properties of matching estimators and establish a number of new results. We focus on matching with replacement with a fixed number of matches. First, we show that matching estimators are not $N^{1/2}$-consistent in general and describe conditions under which matching estimators do attain $N^{1/2}$-consistency. Second, we show that even in settings where matching estimators are $N^{1/2}$-consistent, simple matching estimators with a fixed number of matches do not attain the semiparametric efficiency bound. Third, we provide a consistent estimator for the large sample variance that does not require consistent nonparametric estimation of unknown functions. Software for implementing these methods is available in Matlab, Stata, and R.

KEYWORDS: Matching estimators, average treatment effects, unconfoundedness, selection on observables, potential outcomes.

1. INTRODUCTION

ESTIMATION OF AVERAGE TREATMENT EFFECTS is an important goal of much evaluation research, both in academic studies, as well as in substantive evaluations of social programs. Often, analyses are based on the assumptions that (i) assignment to treatment is unconfounded or exogenous, that is, independent of potential outcomes conditional on observed pretreatment variables, and (ii) there is sufficient overlap in the distributions of the pretreatment variables. Methods for estimating average treatment effects in parametric settings under these assumptions have a long history (see, e.g., Cochran and Rubin (1973), Rubin (1977), Barnow, Cain, and Goldberger (1980), Rosenbaum and Rubin (1983), Heckman and Robb (1984), and Rosenbaum (1995)). Recently, a number of nonparametric implementations of this idea have been proposed. Hahn (1998) calculates the efficiency bound and proposes an asymptotically efficient estimator based on nonparametric series estimation. Heckman,

[1]We wish to thank Donald Andrews, Joshua Angrist, Gary Chamberlain, Geert Dhaene, Jinyong Hahn, James Heckman, Keisuke Hirano, Hidehiko Ichimura, Whitney Newey, Jack Porter, James Powell, Geert Ridder, Paul Rosenbaum, Edward Vytlacil, a co-editor and two anonymous referees, and seminar participants at various universities for comments, and Don Rubin for many discussions on the topic of this article. Financial support for this research was generously provided through National Science Foundation Grants SES-0350645 (Abadie), SBR-9818644, and SES-0136789 (Imbens). Imbens also acknowledges financial support from the Giannini Foundation and the Agricultural Experimental Station at UC Berkeley.

Ichimura, and Todd (1998) focus on the average effect on the treated and consider estimators based on local linear kernel regression methods. Hirano, Imbens, and Ridder (2003) propose an estimator that weights the units by the inverse of their assignment probabilities and show that nonparametric series estimation of this conditional probability, labeled the propensity score by Rosenbaum and Rubin (1983), leads to an efficient estimator of average treatment effects.

Empirical researchers, however, often use simple matching procedures to estimate average treatment effects when assignment for treatment is believed to be unconfounded. Much like nearest neighbor estimators, these procedures match each treated unit to a fixed number of untreated units with similar values for the pretreatment variables. The average effect of the treatment is then estimated by averaging within-match differences in the outcome variable between the treated and the untreated units (see, e.g., Rosenbaum (1995), Dehejia and Wahba (1999)). Matching estimators have great intuitive appeal and are widely used in practice. However, their formal large sample properties have not been established. Part of the reason may be that matching estimators with a fixed number of matches are highly nonsmooth functionals of the distribution of the data, not amenable to standard asymptotic methods for smooth functionals. In this article we study the large sample properties of matching estimators of average treatment effects and establish a number of new results. Like most of the econometric literature, but in contrast with some of the statistics literature, we focus on matching with replacement.

Our results show that some of the formal large sample properties of matching estimators are not very attractive. First, we show that matching estimators include a conditional bias term whose stochastic order increases with the number of continuous matching variables. We show that the order of this conditional bias term may be greater than $N^{-1/2}$, where N is the sample size. As a result, matching estimators are not $N^{1/2}$-consistent in general. Second, even when the simple matching estimator is $N^{1/2}$-consistent, we show that it does not achieve the semiparametric efficiency bound as calculated by Hahn (1998). However, for the case when only a single continuous covariate is used to match, we show that the efficiency loss can be made arbitrarily close to zero by allowing a sufficiently large number of matches. Despite these poor formal properties, matching estimators do have some attractive features that may account for their popularity. In particular, matching estimators are extremely easy to implement and they do not require consistent nonparametric estimation of unknown functions. In this article we also propose a consistent estimator for the variance of matching estimators that does not require consistent nonparametric estimation of unknown functions. This result is particularly relevant because the standard bootstrap does not lead to valid confidence intervals for the

simple matching estimator studied in this article (Abadie and Imbens (2005)). Software for implementing these methods is available in Matlab, Stata, and R.[2]

2. NOTATION AND BASIC IDEAS

2.1. *Notation*

We are interested in estimating the average effect of a binary treatment on some outcome. For unit i, with $i = 1, \ldots, N$, following Rubin (1973), let $Y_i(0)$ and $Y_i(1)$ denote the two potential outcomes given the control treatment and given the active treatment, respectively. The variable W_i, with $W_i \in \{0, 1\}$, indicates the treatment received. For unit i, we observe W_i and the outcome for this treatment,

$$Y_i = \begin{cases} Y_i(0), & \text{if } W_i = 0, \\ Y_i(1), & \text{if } W_i = 1, \end{cases}$$

as well as a vector of pretreatment variables or covariates, denoted by X_i. Our main focus is on the population average treatment effect and its counterpart for the population of the treated:

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)] \quad \text{and} \quad \tau^t = \mathbb{E}\big[Y_i(1) - Y_i(0) | W_i = 1\big].$$

See Rubin (1977), Heckman and Robb (1984), and Imbens (2004) for discussion of these estimands.

We assume that assignment to treatment is unconfounded (Rosenbaum and Rubin (1983)), and that the probability of assignment is bounded away from 0 and 1.

ASSUMPTION 1: Let X be a random vector of dimension k of continuous covariates distributed on $\mathbb{R}^k$ with compact and convex support $\mathbb{X}$, with (a version of the) density bounded and bounded away from zero on its support.

ASSUMPTION 2: For almost every $x \in \mathbb{X}$, where $\mathbb{X}$ is the support of X,

(i) (unconfoundedness) W is independent of $(Y(0), Y(1))$ conditional on $X = x$;

(ii) (overlap) $\eta < \Pr(W = 1 | X = x) < 1 - \eta$ for some $\eta > 0$.

The dimension of X, denoted by k, will be seen to play an important role in the properties of matching estimators. We assume that all covariates have

[2]Software for STATA and Matlab is available at *http://emlab.berkeley.edu/users/imbens/estimators.shtml*. Software for R is available at *http://jsekhon.fas.harvard.edu/matching/Match.html*. Abadie, Drukker, Herr, and Imbens (2004) discuss the implementation in STATA.

continuous distributions.[3] Compactness and convexity of the support of the covariates are convenient regularity conditions. The combination of the two conditions in Assumption 2 is referred to as strong ignorability (Rosenbaum and Rubin (1983)). These conditions are strong and in many cases may not be satisfied.

Heckman, Ichimura, and Todd (1998) point out that for identification of the average treatment effect, τ, Assumption 2(i) can be weakened to mean independence ($\mathbb{E}[Y(w)|W,X] = \mathbb{E}[Y(w)|X]$ for $w = 0, 1$). For simplicity, we assume full independence, although for most of the results, mean independence is sufficient. When the parameter of interest is the average effect for the treated, τ^t, Assumption 2(i) can be relaxed to require only that $Y(0)$ is independent of W conditional on X. Also, when the parameter of interest is τ^t, Assumption 2(ii) can be relaxed so that the support of X for the treated ($\mathbb{X}_1$) is a subset of the support of X for the untreated ($\mathbb{X}_0$).

ASSUMPTION 2′: For almost every $x \in \mathbb{X}$,
(i) W is independent of $Y(0)$ conditional on $X = x$;
(ii) $\Pr(W = 1|X = x) < 1 - \eta$ for some $\eta > 0$.

Under Assumption 2(i), the average treatment effect for the subpopulation with $X = x$ equals

$$\begin{aligned}(1)\qquad \tau(x) &= \mathbb{E}\big[Y(1) - Y(0)|X = x\big] \\ &= \mathbb{E}[Y|W = 1, X = x] - \mathbb{E}[Y|W = 0, X = x]\end{aligned}$$

almost surely. Under Assumption 2(ii), the difference on the right-hand side of (1) is identified for almost all x in $\mathbb{X}$. Therefore, the average effect of the treatment can be recovered by averaging $\mathbb{E}[Y|W = 1, X = x] - \mathbb{E}[Y|W = 0, X = x]$ over the distribution of X:

$$\tau = \mathbb{E}[\tau(X)] = \mathbb{E}\big[\mathbb{E}[Y|W = 1, X = x] - \mathbb{E}[Y|W = 0, X = x]\big].$$

Under Assumption 2′(i), the average treatment effect for the subpopulation with $X = x$ and $W = 1$ is equal to

$$\begin{aligned}(2)\qquad \tau^t(x) &= \mathbb{E}\big[Y(1) - Y(0)|W = 1, X = x\big] \\ &= \mathbb{E}[Y|W = 1, X = x] - \mathbb{E}[Y|W = 0, X = x]\end{aligned}$$

[3]Discrete covariates with a finite number of support points can be easily dealt with by analyzing estimation of average treatment effects within subsamples defined by their values. The number of such covariates does not affect the asymptotic properties of the estimators. In small samples, however, matches along discrete covariates may not be exact, so discrete covariates may create the same type of biases as continuous covariates.

almost surely. Under Assumption 2′(ii), the difference on the right-hand side of (2) is identified for almost all x in $\mathbb{X}_1$. Therefore, the average effect of the treatment on the treated can be recovered by averaging $\mathbb{E}[Y|W=1, X=x] - \mathbb{E}[Y|W=0, X=x]$ over the distribution of X conditional on $W=1$:

$$\begin{aligned}\tau^t &= \mathbb{E}[\tau^t(X)|W=1] \\ &= \mathbb{E}\big[\mathbb{E}[Y|W=1, X=x] - \mathbb{E}[Y|W=0, X=x]\big|W=1\big].\end{aligned}$$

Next, we introduce some additional notation. For $x \in \mathbb{X}$ and $w \in \{0,1\}$, let $\mu(x,w) = \mathbb{E}[Y|X=x, W=w]$, $\mu_w(x) = \mathbb{E}[Y(w)|X=x]$, $\sigma^2(x,w) = \mathbb{V}(Y|X=x, W=w)$, $\sigma_w^2(x) = \mathbb{V}(Y(w)|X=x)$, and $\varepsilon_i = Y_i - \mu_{W_i}(X_i)$. Under Assumption 2, $\mu(x,w) = \mu_w(x)$ and $\sigma^2(x,w) = \sigma_w^2(x)$. Let $f_w(x)$ be the conditional density of X given $W=w$ and let $e(x) = \Pr(W=1|X=x)$ be the propensity score (Rosenbaum and Rubin (1983)). In part of our analysis, we adopt the following assumption.

ASSUMPTION 3: Assume $\{(Y_i, W_i, X_i)\}_{i=1}^N$ are independent draws from the distribution of (Y, W, X).

In some cases, however, treated and untreated are sampled separately and their proportions in the sample may not reflect their proportions in the population. Therefore, we relax Assumption 3 so that conditional on W_i, sampling is random. As we will show later, relaxing Assumption 3 is particularly useful when the parameter of interest is the average treatment effect on the treated. The numbers of control and treated units are N_0 and N_1, respectively, with $N = N_0 + N_1$. We assume that N_0 is at least of the same order of magnitude as N_1.

ASSUMPTION 3′: Conditional on $W_i = w$, the sample consists of independent draws from $Y, X|W=w$ for $w=0,1$. For some $r \geq 1$, $N_1^r/N_0 \to \theta$ with $0 < \theta < \infty$.

In this article we focus on matching with replacement, allowing each unit to be used as a match more than once. For $x \in \mathbb{X}$, let $\|x\| = (x'x)^{1/2}$ be the standard Euclidean vector norm.[4] Let $j_m(i)$ be the index $j \in \{1, 2, \ldots, N\}$ that solves $W_j = 1 - W_i$ and

$$\sum_{l: W_l = 1 - W_i} \mathbb{1}\big\{\|X_l - X_i\| \leq \|X_j - X_i\|\big\} = m,$$

[4]Alternative norms of the form $\|x\|_V = (x'Vx)^{1/2}$ for some positive definite symmetric matrix V are also covered by the results below, because $\|x\|_V = ((Px)'(Px))^{1/2}$ for P such that $P'P = V$.

where $\mathbb{1}\{\cdot\}$ is the indicator function, equal to 1 if the expression in brackets is true and 0 otherwise. In other words, $j_m(i)$ is the index of the unit that is the mth closest to unit i in terms of the covariate values, among the units with the treatment opposite to that of unit i. In particular, $j_1(i)$, which will be sometimes denoted by $j(i)$, is the nearest match for unit i. For notational simplicity and because we consider only continuous covariates, we ignore the possibility of ties, which happen with probability 0. Let $\mathcal{J}_M(i)$ denote the set of indices for the first M matches for unit i: $\mathcal{J}_M(i) = \{j_1(i), \ldots, j_M(i)\}$.[5] Finally, let $K_M(i)$ denote the number of times unit i is used as a match given that M matches per unit are used:

$$K_M(i) = \sum_{l=1}^{N} \mathbb{1}\{i \in \mathcal{J}_M(l)\}.$$

The distribution of $K_M(i)$ will play an important role in the variance of the estimators.

In many analyses of matching methods (e.g., Rosenbaum (1995)), matching is carried out without replacement, so that every unit is used as a match at most once and $K_M(i) \leq 1$. In this article, however, we focus on matching with replacement, allowing each unit to be used as a match more than once. Matching with replacement produces matches of higher quality than matching without replacement by increasing the set of possible matches.[6] In addition, matching with replacement has the advantage that it allows us to consider estimators that match all units, treated as well as controls, so that the estimand is identical to the population average treatment effect.

2.2. *The Matching Estimator*

The unit-level treatment effect is $\tau_i = Y_i(1) - Y_i(0)$. For the units in the sample, only one of the potential outcomes, $Y_i(0)$ and $Y_i(1)$, is observed and the other is unobserved or missing. The matching estimator imputes the missing potential outcomes as

$$\widehat{Y}_i(0) = \begin{cases} Y_i, & \text{if } W_i = 0, \\ \dfrac{1}{M} \displaystyle\sum_{j \in \mathcal{J}_M(i)} Y_j, & \text{if } W_i = 1, \end{cases}$$

[5]For this definition to make sense, we assume that $N_0 \geq M$ and $N_1 \geq M$. We maintain this assumption implicitly throughout.

[6]As we show below, inexact matches generate bias in matching estimators. Therefore, expanding the set of possible matches will tend to produce smaller biases.

and

$$\widehat{Y}_i(1) = \begin{cases} \dfrac{1}{M} \displaystyle\sum_{j \in \mathcal{J}_M(i)} Y_j, & \text{if } W_i = 0, \\ Y_i, & \text{if } W_i = 1, \end{cases}$$

leading to the following estimator for the average treatment effect:

$$\widehat{\tau}_M = \frac{1}{N} \sum_{i=1}^{N} \big(\widehat{Y}_i(1) - \widehat{Y}_i(0)\big) = \frac{1}{N} \sum_{i=1}^{N} (2W_i - 1)\left(1 + \frac{K_M(i)}{M}\right) Y_i. \tag{3}$$

This estimator can easily be modified to estimate the average treatment effect on the treated:

$$\widehat{\tau}^t_M = \frac{1}{N_1} \sum_{W_i=1} \big(Y_i - \widehat{Y}_i(0)\big) = \frac{1}{N_1} \sum_{i=1}^{N} \left(W_i - (1 - W_i)\frac{K_M(i)}{M}\right) Y_i. \tag{4}$$

It is useful to compare matching estimators to covariance-adjustment or regression imputation estimators. Let $\widehat{\mu}_w(X_i)$ be a consistent estimator of $\mu_w(X_i)$. Let

$$\bar{Y}_i(0) = \begin{cases} Y_i, & \text{if } W_i = 0, \\ \widehat{\mu}_0(X_i), & \text{if } W_i = 1, \end{cases} \tag{5}$$

$$\bar{Y}_i(1) = \begin{cases} \widehat{\mu}_1(X_i), & \text{if } W_i = 0, \\ Y_i, & \text{if } W_i = 1. \end{cases}$$

The regression imputation estimators of τ and τ^t are

$$\widehat{\tau}^{\text{reg}} = \frac{1}{N} \sum_{i=1}^{N} \big(\bar{Y}_i(1) - \bar{Y}_i(0)\big) \quad \text{and} \quad \widehat{\tau}^{\text{reg},t} = \frac{1}{N_1} \sum_{W_i=1} \big(Y_i - \bar{Y}_i(0)\big). \tag{6}$$

In our discussion we classify as regression imputation estimators those for which $\widehat{\mu}_w(x)$ is a consistent estimator of $\mu_w(x)$. The estimators proposed by Hahn (1998) and some of those proposed by Heckman, Ichimura, and Todd (1998) fall into this category.[7]

If $\mu_w(X_i)$ is estimated using a nearest neighbor estimator with a fixed number of neighbors, then the regression imputation estimator is identical to the matching estimator with the same number of matches. The two estimators

[7]In a working paper version (Abadie and Imbens (2002)), we consider a bias-corrected version of the matching estimator that combines some of the feature of matching and regression estimators.

differ in the way they change with the sample size. We classify as matching estimators those estimators that use a finite and fixed number of matches. Interpreting matching estimators in this way may provide some intuition for some of the subsequent results. In nonparametric regression methods one typically chooses smoothing parameters to balance bias and variance of the estimated regression function. For example, in kernel regression a smaller bandwidth leads to lower bias but higher variance. A nearest neighbor estimator with a single neighbor is at the extreme end of this. The bias is minimized within the class of nearest neighbor estimators, but the variance of $\widehat{\mu}_w(x)$ no longer vanishes with the sample size. Nevertheless, as we shall show, matching estimators of average treatment effects are consistent under weak regularity conditions. The variance of matching estimators, however, is still relatively high and, as a result, matching with a fixed number of matches does not lead to an efficient estimator.

The first goal of this article is to derive the properties of the simple matching estimator in large samples, that is, as N increases, for fixed M. The motivation for our fixed-M asymptotics is to provide an approximation to the sampling distribution of matching estimators with a small number of matches. Such matching estimators have been widely used in practice. The properties of interest include bias and variance. Of particular interest is the dependence of these results on the dimension of the covariates. A second goal is to provide methods for conducting inference through estimation of the large sample variance of the matching estimator.

3. LARGE SAMPLE PROPERTIES OF THE MATCHING ESTIMATOR

In this section we investigate the properties of the matching estimator, $\widehat{\tau}_M$, defined in (3). We can decompose the difference between the matching estimator $\widehat{\tau}_M$ and the population average treatment effect τ as

$$\widehat{\tau}_M - \tau = \left(\overline{\tau(X)} - \tau\right) + E_M + B_M, \tag{7}$$

where $\overline{\tau(X)}$ is the average conditional treatment effect,

$$\overline{\tau(X)} = \frac{1}{N}\sum_{i=1}^{N}(\mu_1(X_i) - \mu_0(X_i)), \tag{8}$$

E_M is a weighted average of the residuals,

$$E_M = \frac{1}{N}\sum_{i=1}^{N} E_{M,i} = \frac{1}{N}\sum_{i=1}^{N}(2W_i - 1)\left(1 + \frac{K_M(i)}{M}\right)\varepsilon_i, \tag{9}$$

and B_M is the conditional bias relative to $\overline{\tau(X)}$,

$$
\begin{aligned}
(10) \qquad B_M &= \frac{1}{N}\sum_{i=1}^{N} B_{M,i} \\
&= \frac{1}{N}\sum_{i=1}^{N}(2W_i - 1)\cdot\left[\frac{1}{M}\sum_{m=1}^{M}\bigl(\mu_{1-W_i}(X_i) - \mu_{1-W_i}(X_{j_m(i)})\bigr)\right].
\end{aligned}
$$

The first two terms on the right-hand side of (7), $(\overline{\tau(X)} - \tau)$ and E_M, have zero mean. They will be shown to be of order $N^{-1/2}$ and asymptotically normal. The first term depends only on the covariates, and its variance is $V^{\tau(X)}/N$, where $V^{\tau(X)} = \mathbb{E}[(\tau(X) - \tau)^2]$ is the variance of the conditional average treatment effect $\tau(X)$. Conditional on $\mathbf{X}$ and $\mathbf{W}$ (the matrix and vector with ith row equal to X_i' and W_i, respectively), the variance of $\widehat{\tau}_M$ is equal to the conditional variance of the second term, $\mathbb{V}(E_M|\mathbf{X}, \mathbf{W})$. We will analyze this variance in Section 3.2. We will refer to the third term on the right-hand side of (7), B_M, as the conditional bias, and to $\mathbb{E}[B_M]$ as the (unconditional) bias. If matching is exact, $X_i = X_{j_m(i)}$ for all i and the conditional bias is equal to zero. In general it differs from zero and its properties, in particular its stochastic order, will be analyzed in Section 3.1.

Similarly, we can decompose the estimator for the average effect for the treated, (4), as

$$
(11) \qquad \widehat{\tau}_M^t - \tau^t = \bigl(\overline{\tau(X)}^{\,t} - \tau^t\bigr) + E_M^t + B_M^t,
$$

where

$$
\overline{\tau(X)}^{\,t} = \frac{1}{N_1}\sum_{i=1}^{N} W_i(\mu(X_i, 1) - \mu_0(X_i)),
$$

$$
E_M^t = \frac{1}{N_1}\sum_{i=1}^{N} E_{M,i}^t = \frac{1}{N_1}\sum_{i=1}^{N}\left(W_i - (1 - W_i)\frac{K_M(i)}{M}\right)\varepsilon_i,
$$

and

$$
B_M^t = \frac{1}{N_1}\sum_{i=1}^{N} B_{M,i}^t = \frac{1}{N_1}\sum_{i=1}^{N} W_i\frac{1}{M}\sum_{m=1}^{M}\bigl(\mu_0(X_i) - \mu_0(X_{j_m(i)})\bigr).
$$

3.1. *Bias*

Here we investigate the stochastic order of the conditional bias (10) and its counterpart for the average treatment effect for the treated. The conditional bias consists of sums of terms of the form $\mu_1(X_{j_m(i)}) - \mu_1(X_i)$ or

$\mu_0(X_i) - \mu_0(X_{j_m(i)})$. To investigate the nature of these terms, expand the difference $\mu_1(X_{j_m(i)}) - \mu_1(X_i)$ around X_i:

$$\begin{aligned}
&\mu_1(X_{j_m(i)}) - \mu_1(X_i) \\
&\quad = (X_{j_m(i)} - X_i)' \frac{\partial \mu_1}{\partial x}(X_i) \\
&\qquad + \frac{1}{2}(X_{j_m(i)} - X_i)' \frac{\partial^2 \mu_1}{\partial x \, \partial x'}(X_i)(X_{j_m(i)} - X_i) + O\big(\|X_{j_m(i)} - X_i\|^3\big).
\end{aligned}$$

To study the components of the bias, it is therefore useful to analyze the distribution of the k vector $X_{j_m(i)} - X_i$, which we term the *matching discrepancy*.

First, let us analyze the matching discrepancy at a general level. Fix the covariate value at $X = z$ and suppose we have a random sample $X_1, \ldots, X_N$ with density $f(x)$ over a bounded support $\mathbb{X}$. Now consider the closest match to z in the sample. Let $j_1 = \arg\min_{j=1,\ldots,N} \|X_j - z\|$ and let $U_1 = X_{j_1} - z$ be the matching discrepancy. We are interested in the distribution of the k vector U_1. More generally, we are interested in the distribution of the mth closest matching discrepancy, $U_m = X_{j_m} - z$, where j_m is the mth closest match to z from the random sample of size N. The following lemma describes some key asymptotic properties of the matching discrepancy at interior points of the support of X.

LEMMA 1—Matching Discrepancy—Asymptotic Properties: *Suppose that f is differentiable in a neighborhood of z. Let $V_m = N^{1/k} U_m$ and let $f_{V_m}(v)$ be the density of V_m. Then*

$$\begin{aligned}
&\lim_{N\to\infty} f_{V_m}(v) \\
&\quad = \frac{f(z)}{(m-1)!}\left(\|v\|^k \frac{f(z)}{k}\frac{2\pi^{k/2}}{\Gamma(k/2)}\right)^{m-1} \exp\left(-\|v\|^k \frac{f(z)}{k}\frac{2\pi^{k/2}}{\Gamma(k/2)}\right),
\end{aligned}$$

where $\Gamma(y) = \int_0^\infty e^{-t} t^{y-1}\,dt$ (for $y > 0$) is Euler's gamma function. Hence $U_m = O_p(N^{-1/k})$. Moreover, the first three moments of U_m are

$$\begin{aligned}
\mathbb{E}[U_m] &= \Gamma\left(\frac{mk+2}{k}\right)\frac{1}{(m-1)!k}\left(f(z)\frac{\pi^{k/2}}{\Gamma(1+k/2)}\right)^{-2/k} \\
&\qquad \times \frac{1}{f(z)}\frac{\partial f}{\partial x}(z)\frac{1}{N^{2/k}} + o\left(\frac{1}{N^{2/k}}\right), \\
\mathbb{E}[U_m U_m'] &= \Gamma\left(\frac{mk+2}{k}\right)\frac{1}{(m-1)!k}\left(f(z)\frac{\pi^{k/2}}{\Gamma(1+k/2)}\right)^{-2/k}\frac{1}{N^{2/k}} I_k \\
&\qquad + o\left(\frac{1}{N^{2/k}}\right),
\end{aligned}$$

where I_k is the identity matrix of size k and $\mathbb{E}[\|U_m\|^3] = O(N^{-3/k})$.

(All proofs are given in the Appendix.)

This lemma shows how the order of the matching discrepancy increases with the number of continuous covariates. The lemma also shows that the first term in the stochastic expansion of $N^{1/k}U_m$ has a rotation invariant distribution with respect to the origin. The following lemma shows that for all points in the support, including the boundary points not covered by Lemma 1, the normalized moments of the matching discrepancies, U_m, are bounded.

LEMMA 2—Matching Discrepancy—Uniformly Bounded Moments: *If Assumption* 1 *holds, then all the moments of* $N^{1/k}\|U_m\|$ *are uniformly bounded in* N *and* $z \in \mathbb{X}$.

These results allow us to establish bounds on the stochastic order of the conditional bias.

THEOREM 1—Conditional Bias for the Average Treatment Effect: *Under Assumptions* 1, 2, *and* 3, (i) *if* $\mu_0(x)$ *and* $\mu_1(x)$ *are Lipschitz on* $\mathbb{X}$, *then* $B_M = O_p(N^{-1/k})$, *and* (ii) *the order of* $\mathbb{E}[B_M]$ *is not in general lower than* $N^{-2/k}$.

Consider the implications of this theorem for the asymptotic properties of the simple matching estimator. First notice that, under regularity conditions, $\sqrt{N}(\overline{\tau(X)} - \tau) = O_p(1)$ with a normal limiting distribution, by a standard central limit theorem. Also, it will be shown later that, under regularity conditions $\sqrt{N}E_M = O_p(1)$, again with a normal limiting distribution. However, the result of the theorem implies that $\sqrt{N}B_M$ is not $O_p(1)$ in general. In particular, if k is large enough, the asymptotic distribution of $\sqrt{N}(\widehat{\tau}_M - \tau)$ is dominated by the bias term and the simple matching estimator is not $N^{1/2}$-consistent. However, if only one of the covariates is continuously distributed, then $k = 1$ and $B_M = O_p(N^{-1})$, so $\sqrt{N}(\widehat{\tau}_M - \tau)$ will be asymptotically normal.

The following result describes the properties of the matching estimator for the average effect on the treated.

THEOREM 2—Conditional Bias for the Average Treatment Effect on the Treated: *Under Assumptions* 1, 2′, *and* 3′

(i) *if* $\mu_0(x)$ *is Lipschitz on* $\mathbb{X}_0$, *then* $B_M^t = O_p(N_1^{-r/k})$, *and*

(ii) *if* $\mathbb{X}_1$ *is a compact subset of the interior of* $\mathbb{X}_0$, $\mu_0(x)$ *has bounded third derivatives in the interior of* $\mathbb{X}_0$, *and* $f_0(x)$ *is differentiable in the interior of* $\mathbb{X}_0$ *with bounded derivatives, then*

$$\begin{aligned}\text{Bias}_M^t &= \mathbb{E}[B_M^t] \\ &= -\left(\frac{1}{M}\sum_{m=1}^{M}\Gamma\left(\frac{mk+2}{k}\right)\frac{1}{(m-1)!k}\right)\frac{1}{N_1^{2r/k}}\end{aligned}$$

$$\times \theta^{2/k} \int \left(f_0(x) \frac{\pi^{k/2}}{\Gamma(1+k/2)} \right)^{-2/k} \times \left\{ \frac{1}{f_0(x)} \frac{\partial f_0}{\partial x'}(x) \frac{\partial \mu_0}{\partial x}(x) + \frac{1}{2} \operatorname{tr}\left(\frac{\partial^2 \mu_0}{\partial x' \partial x}(x) \right) \right\} \times f_1(x)\, dx + o\left(\frac{1}{N_1^{2r/k}} \right).$$

This case is particularly relevant because often matching estimators have been used to estimate the average effect for the treated in settings in which a large number of controls are sampled separately. Typically in those cases the conditional bias term has been ignored in the asymptotic approximation to standard errors and confidence intervals. Theorem 2 shows that ignoring the conditional bias term in the first-order asymptotic approximation to the distribution of the simple matching estimator is justified if N_0 is of sufficiently high order relative to N_1 or, to be precise, if $r > k/2$. In that case it follows that $B_M^t = o_p(N_1^{-1/2})$ and the bias term will get dominated in the large sample distribution by the two other terms, $\overline{\tau(X)}^t - \tau^t$ and E_M^t, both of which are $O_p(N_1^{-1/2})$.

In part (ii) of Theorem 2, we show that a general expression of the bias, $\mathbb{E}[B_M^t]$, can be calculated if $\mathbb{X}_1$ is compact and $\mathbb{X}_1 \subset \operatorname{int} \mathbb{X}_0$ (so that the bias is not affected by the geometric characteristics of the boundary of $\mathbb{X}_0$). Under these conditions, the bias of the matching estimator is at most of order $N_1^{-2/k}$. This bias is further reduced when $\mu_0(x)$ is constant or when $\mu_0(x)$ is linear and $f_0(x)$ is constant, among other cases. Notice, however, that usual smoothness assumptions (existence of higher order derivatives) do not reduce the order of $\mathbb{E}[B_M^t]$.

3.2. *Variance*

In this section we investigate the variance of the matching estimator $\widehat{\tau}_M$. We focus on the first two terms of the representation of the estimator in (7), that is, the term that represents the heterogeneity in the treatment effect, (8), and the term that represents the residuals, (9), ignoring for the moment the conditional bias term (10). Conditional on $\mathbf{X}$ and $\mathbf{W}$, the matrix and vector with ith row equal to X_i' and W_i, respectively, the number of times a unit is used as a match, $K_M(i)$ is deterministic and hence the variance of $\widehat{\tau}_M$ is

$$\mathbb{V}(\widehat{\tau}_M | \mathbf{X}, \mathbf{W}) = \frac{1}{N^2} \sum_{i=1}^{N} \left(1 + \frac{K_M(i)}{M} \right)^2 \sigma^2(X_i, W_i). \tag{12}$$

For $\widehat{\tau}_M^t$ we obtain

$$\mathbb{V}(\widehat{\tau}_M^t|\mathbf{X},\mathbf{W}) = \frac{1}{N_1^2}\sum_{i=1}^{N}\left(W_i - (1-W_i)\frac{K_M(i)}{M}\right)^2\sigma^2(X_i,W_i). \tag{13}$$

Let $V^E = N\mathbb{V}(\widehat{\tau}_M|\mathbf{X},\mathbf{W})$ and $V^{E,t} = N_1\mathbb{V}(\widehat{\tau}_M^t|\mathbf{X},\mathbf{W})$ be the corresponding normalized variances. Ignoring the conditional bias term, B_M, the conditional expectation of $\widehat{\tau}_M$ is $\overline{\tau(X)}$. The variance of this conditional mean is therefore $V^{\tau(X)}/N$, where $V^{\tau(X)} = \mathbb{E}[(\tau(X)-\tau)^2]$. Hence the marginal variance of $\widehat{\tau}_M$, ignoring the conditional bias term, is $\mathbb{V}(\widehat{\tau}_M) = (\mathbb{E}[V^E] + V^{\tau(X)})/N$. For the estimator for the average effect on the treated, the marginal variance is, again ignoring the conditional bias term, $\mathbb{V}(\widehat{\tau}_M^t) = (\mathbb{E}[V^{E,t}] + V^{\tau(X),t})/N_1$, where $V^{\tau(X),t} = \mathbb{E}[(\tau^t(X)-\tau^t)^2|W=1]$.

The following lemma shows that the expectation of the normalized variance is finite. The key is that $K_M(i)$, the number of times that unit i is used as a match, is $O_p(1)$ with finite moments.[8]

LEMMA 3—Finite Variance: (i) *Suppose Assumptions* 1–3 *hold. Then* $K_M(i) = O_p(1)$ *and* $\mathbb{E}[K_M(i)^q]$ *is bounded uniformly in* N *for any* $q > 0$. (ii) *If, in addition,* $\sigma^2(x,w)$ *are Lipschitz in* $\mathbb{X}$ *for* $w = 0, 1$, *then* $\mathbb{E}[V^E + V^{\tau(X)}] = O(1)$. (iii) *Suppose Assumptions* 1, 2′, *and* 3′. *Then* $(N_0/N_1)\mathbb{E}[K_M(i)^q|W_i = 0]$ *is uniformly bounded in* N *for any* $q > 0$. (iv) *If, in addition,* $\sigma^2(x,w)$ *are Lipschitz in* $\mathbb{X}$ *for* $w = 0, 1$, *then* $\mathbb{E}[V^{E,t} + V^{\tau(X),t}] = O(1)$.

3.3. *Consistency and Asymptotic Normality*

In this section we show that the matching estimator is consistent for the average treatment effect and, without the conditional bias term, is $N^{1/2}$-consistent and asymptotically normal. The next assumption contains a set of weak smoothness restrictions on the conditional distribution of Y given X. Notice that it does not require the existence of higher order derivatives.

ASSUMPTION 4: For $w = 0, 1$, (i) $\mu(x,w)$ and $\sigma^2(x,w)$ are Lipschitz in $\mathbb{X}$, (ii) the fourth moments of the conditional distribution of Y given $W = w$ and $X = x$ exist and are bounded uniformly in x, and (iii) $\sigma^2(x,w)$ is bounded away from zero.

THEOREM 3—Consistency of the Matching Estimator:

(i) *Suppose Assumptions* 1–3 *and* 4(i) *hold. Then* $\widehat{\tau}_M - \tau \xrightarrow{p} 0$.

(ii) *Suppose Assumptions* 1, 2′, 3′, *and* 4(i) *hold. Then* $\widehat{\tau}_M^t - \tau^t \xrightarrow{p} 0$.

[8]Notice that, for $1 \le i \le N$, $K_M(i)$ are exchangeable random variables and therefore have identical marginal distributions.

Notice that the consistency result holds regardless of the dimension of the covariates.

Next, we state the formal result for asymptotic normality. The first result gives an asymptotic normality result for the estimators $\widehat{\tau}_M$ and $\widehat{\tau}_M^t$ after subtracting the bias term.

THEOREM 4—Asymptotic Normality for the Matching Estimator:

(i) *Suppose Assumptions* 1–4 *hold*. *Then*

$$(V^E + V^{\tau(X)})^{-1/2}\sqrt{N}(\widehat{\tau}_M - B_M - \tau) \xrightarrow{d} \mathcal{N}(0,1).$$

(ii) *Suppose Assumptions* 1, 2′, 3′, *and* 4 *hold*. *Then*

$$(V^{E,t} + V^{\tau(X),t})^{-1/2}\sqrt{N_1}(\widehat{\tau}_M^t - B_M^t - \tau^t) \xrightarrow{d} \mathcal{N}(0,1).$$

Although one generally does not know the conditional bias term, this result is useful for two reasons. First, in some cases the bias term can be ignored because it is of sufficiently low order (see Theorems 1 and 2). Second, as we show in Abadie and Imbens (2002), under some additional smoothness conditions, an estimate of the bias term based on nonparametric estimation of $\mu_0(x)$ and $\mu_1(x)$ can be used in the statement of Theorem 4 without changing the resulting asymptotic distribution.

In the scalar covariate case or when only the treated are matched and the size of the control group is of sufficient order of magnitude, there is no need to remove the bias.

COROLLARY 1—Asymptotic Normality for Matching Estimator—Vanishing Bias:

(i) *Suppose Assumptions* 1–4 *hold and* $k = 1$. *Then*

$$(V^E + V^{\tau(X)})^{-1/2}\sqrt{N}(\widehat{\tau}_M - \tau) \xrightarrow{d} \mathcal{N}(0,1).$$

(ii) *Suppose Assumptions* 1, 2′, 3′, *and* 4 *hold*, *and* $r > k/2$. *Then*

$$(V^{E,t} + V^{\tau(X),t})^{-1/2}\sqrt{N_1}(\widehat{\tau}_M^t - \tau^t) \xrightarrow{d} \mathcal{N}(0,1).$$

3.4. *Efficiency*

The asymptotic efficiency of the estimators considered here depends on the limit of $\mathbb{E}[V^E]$, which in turn depends on the limiting distribution of $K_M(i)$. It is difficult to work out the limiting distribution of this variable for the general

case.[9] Here we investigate the form of the variance for the special case with a scalar covariate ($k = 1$) and a general M.

THEOREM 5: *Suppose* $k = 1$. *If Assumptions* 1–4 *hold, and* $f_0(x)$ *and* $f_1(x)$ *are continuous on* $\text{int}\,\mathbb{X}$, *then*

$$N \cdot \mathbb{V}(\widehat{\tau}_M) = \mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)}\right] + V^{\tau(X)} + \frac{1}{2M}\mathbb{E}\left[\left(\frac{1}{e(X)} - e(X)\right)\sigma_1^2(X) + \left(\frac{1}{1-e(X)} - (1-e(X))\right)\sigma_0^2(X)\right] + o(1).$$

Note that with $k = 1$ we can ignore the conditional bias term, B_M. The semiparametric efficiency bound for this problem is, as established by Hahn (1998),

$$V^{\text{eff}} = \mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)}\right] + V^{\tau(X)}.$$

The limiting variance of the matching estimator is in general larger. Relative to the efficiency bound it can be written as

$$\lim_{N\to\infty} \frac{N \cdot \mathbb{V}(\widehat{\tau}_M) - V^{\text{eff}}}{V^{\text{eff}}} < \frac{1}{2M}.$$

The asymptotic efficiency loss disappears quickly if the number of matches is large enough and the efficiency loss from using a few matches is very small. For example, the asymptotic variance with a single match is less than 50% higher than the asymptotic variance of the efficient estimator and with five matches, the asymptotic variance is less than 10% higher.

4. ESTIMATING THE VARIANCE

Corollary 1 uses the square roots of $V^E + V^{\tau(X)}$ and $V^{E,t} + V^{\tau(X),t}$, respectively, as normalizing factors to obtain a limiting normal distribution for matching estimators. In this section, we show how to estimate these asymptotic variances.

[9]The key is the second moment of the volume of the "catchment area" $\mathbb{A}_M(i)$, defined as the subset of $\mathbb{X}$ such that each observation, j, with $W_j = 1 - W_i$ and $X_j \in \mathbb{A}_M(i)$ is matched to i. In the single match case with $M = 1$, these catchment areas are studied in stochastic geometry where they are known as Poisson–Voronoi tessellations (Okabe, Boots, Sugihara, and Nok Chiu (2000)). The variance of the volume of such objects under uniform $f_0(x)$ and $f_1(x)$, normalized by the mean volume, has been worked out analytically for the scalar case and numerically for the two- and three-dimensional cases.

4.1. *Estimating the Conditional Variance*

Estimating the conditional variance, $V^E = \sum_{i=1}^N (1 + K_M(i)/M)^2 \sigma^2(X_i, W_i)/N$, is complicated by the fact that it involves the conditional outcome variances, $\sigma^2(x, w)$. In principle, these conditional variances could be consistently estimated using nonparametric smoothing techniques. We propose, however, an estimator of the conditional variance of the simple matching estimator that does not require consistent nonparametric estimation of unknown functions. Our method uses a matching estimator for $\sigma^2(x, w)$, where instead of the original matching of treated to control units, we now match treated units to treated units and control units to control units.

Let $\ell_m(i)$ be the mth closest unit to unit i among the units with the same value for the treatment. Then, for fixed J, we estimate the conditional variance as

$$\widehat{\sigma}^2(X_i, W_i) = \frac{J}{J+1}\left(Y_i - \frac{1}{J}\sum_{m=1}^{J} Y_{\ell_j(i)}\right)^2. \tag{14}$$

Notice that if all matches are perfect so $X_{\ell_j(i)} = X_i$ for all $j = 1, \ldots, J$, then $\mathbb{E}[\widehat{\sigma}^2(X_i, W_i)|X_i = x, W_i = w] = \sigma^2(x, w)$. In practice, if the covariates are continuous, it will not be possible to find perfect matches, so $\widehat{\sigma}^2(X_i, W_i)$ will be only asymptotically unbiased. In addition, because $\widehat{\sigma}^2(X_i, W_i)$ is an average of a fixed number (i.e., J) of observations, this estimator will not be consistent for $\sigma^2(X_i, W_i)$. However, the next theorem shows that the appropriate averages of the $\widehat{\sigma}^2(X_i, W_i)$ over the sample are consistent for V^E and $V^{E,t}$.

THEOREM 6: *Let* $\widehat{\sigma}^2(X_i, W_i)$ *be as in* (14). *Define*

$$\widehat{V}^E = \frac{1}{N}\sum_{i=1}^{N}\left(1 + \frac{K_M(i)}{M}\right)^2 \widehat{\sigma}^2(X_i, W_i),$$

$$\widehat{V}^{E,t} = \frac{1}{N_1}\sum_{i=1}^{N}\left(W_i - (1 - W_i)\frac{K_M(i)}{M}\right)^2 \widehat{\sigma}^2(X_i, W_i).$$

If Assumptions 1–4 *hold, then* $|\widehat{V}^E - V^E| = o_p(1)$. *If Assumptions* 1, 2′, 3′, *and* 4 *hold, then* $|\widehat{V}^{E,t} - V^{E,t}| = o_p(1)$.

4.2. *Estimating the Marginal Variance*

Here we develop consistent estimators for $V = V^E + V^{\tau(X)}$ and $V^t = V^{E,t} + V^{\tau(X),t}$. The proposed estimators are based on the same matching approach to

estimating the conditional error variance $\sigma^2(x,w)$ as in Section 4.1. In addition, these estimators exploit the fact that

$$\mathbb{E}\big[\big(\widehat{Y}_i(1)-\widehat{Y}_i(0)-\tau\big)^2\big] \simeq V^{\tau(X)} + \mathbb{E}\left[\varepsilon_i^2 + \frac{1}{M^2}\sum_{m=1}^{M}\varepsilon_{j_m(i)}^2\right].$$

The average on the left-hand side can be estimated as $\sum_i(\widehat{Y}_i(1)-\widehat{Y}_i(0)-\widehat{\tau}_M)^2/N$. To estimate the second term on the right-hand side, we use the fact that

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\varepsilon_i^2 + \frac{1}{M^2}\sum_{m=1}^{M}\varepsilon_{j_m(i)}^2\,\middle|\,\mathbf{X},\mathbf{W}\right] = \frac{1}{N}\sum_{i=1}^{N}\left(1+\frac{K_M(i)}{M^2}\right)\sigma^2(X_i,W_i),$$

which can be estimated using the matching estimator for $\sigma^2(X_i,W_i)$. These two estimates can then be combined to estimate $V^{\tau(X)}$ and this in turn can be combined with the previously defined estimator for V^E to obtain an estimator of V.

THEOREM 7: *Let* $\widehat{\sigma}^2(X_i,W_i)$ *be as in* (14). *Define*

$$\widehat{V} = \frac{1}{N}\sum_{i=1}^{N}\big(\widehat{Y}_i(1)-\widehat{Y}_i(0)-\widehat{\tau}_M\big)^2 + \frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{K_M(i)}{M}\right)^2 + \left(\frac{2M-1}{M}\right)\left(\frac{K_M(i)}{M}\right)\right]\widehat{\sigma}^2(X_i,W_i)$$

and

$$\widehat{V}^t = \frac{1}{N_1}\sum_{W_i=1}\big(Y_i-\widehat{Y}_i(0)-\widehat{\tau}_M^t\big)^2 + \frac{1}{N_1}\sum_{i=1}^{N}(1-W_i)\left(\frac{K_M(i)(K_M(i)-1)}{M^2}\right)\widehat{\sigma}^2(X_i,W_i).$$

If Assumptions 1–4 *hold, then* $|\widehat{V}-V| = o_p(1)$. *If Assumptions* 1, 2′, 3′, *and* 4 *hold, then* $|\widehat{V}^t - V^t| = o_p(1)$.

5. CONCLUSION

In this article we derive large sample properties of matching estimators of average treatment effects that are widely used in applied evaluation research. The formal large sample properties of matching estimators are somewhat surprising in the light of this popularity. We show that matching estimators include

a conditional bias term that may be of order larger than $N^{-1/2}$. Therefore, matching estimators are not $N^{1/2}$-consistent in general and standard confidence intervals are not necessarily valid. We show, however, that when the set of matching variables contains at most one continuously distributed variable, the conditional bias term is $o_p(N^{-1/2})$, so that matching estimators are $N^{1/2}$-consistent in this case. We derive the asymptotic distribution of matching estimators for the cases when the conditional bias can be ignored and also show that matching estimators with a fixed number of matches do not reach the semiparametric efficiency bound. Finally, we propose an estimator of the asymptotic variance. This is particularly relevant because there is evidence that the bootstrap is not valid for matching estimators (Abadie and Imbens (2005)).

John F. Kennedy School of Government, Harvard University, 79 John F. Kennedy Street, Cambridge, MA 02138, U.S.A.; and NBER; alberto_abadie@harvard.edu; http://www.ksg.harvard.edu/fs/aabadie/

and

Dept. of Economics and Dept. of Agricultural and Resource Economics, University of California at Berkeley, 661 Evans Hall #3880, Berkeley, CA 94720-3880, U.S.A.; and NBER; imbens@econ.berkeley.edu; http://elsa.berkeley.edu/users/imbens/.

Manuscript received August, 2002; final revision received March, 2005.

APPENDIX

Before proving Lemma 1, we collect some results on integration using polar coordinates that will be useful. See, for example, Stroock (1994). Let $\mathbb{S}_k = \{\omega \in \mathbb{R}^k : \|\omega\| = 1\}$ be the unit k sphere and let $\lambda_{\mathbb{S}_k}$ be its surface measure. Then the area and volume of the unit k sphere are

$$\int_{\mathbb{S}_k} \lambda_{\mathbb{S}_k}(d\omega) = \frac{2\pi^{k/2}}{\Gamma(k/2)}$$

and

$$\int_0^1 r^{k-1} \int_{\mathbb{S}_k} \lambda_{\mathbb{S}_k}(d\omega)\, dr = \frac{2\pi^{k/2}}{k\Gamma(k/2)} = \frac{\pi^{k/2}}{\Gamma(1+k/2)},$$

respectively. In addition,

$$\int_{\mathbb{S}_k} \omega \lambda_{\mathbb{S}_k}(d\omega) = 0$$

and

$$\int_{\mathbb{S}_k} \omega\omega' \lambda_{\mathbb{S}_k}(d\omega) = \frac{\int_{\mathbb{S}_k} \lambda_{\mathbb{S}_k}(d\omega)}{k} I_k = \frac{\pi^{k/2}}{\Gamma(1+k/2)} I_k,$$

where I_k is the k-dimensional identity matrix. For any nonnegative measurable function $g(\cdot)$ on $\mathbb{R}^k$,

$$\int_{\mathbb{R}^k} g(x)\,dx = \int_0^\infty r^{k-1}\left(\int_{\mathbb{S}_k} g(r\omega)\lambda_{\mathbb{S}_k}(d\omega)\right)dr.$$

We will also use the following result on Laplace approximation of integrals.

LEMMA A.1: *Let $a(r)$ and $b(r)$ be two real functions; $a(r)$ is continuous in a neighborhood of zero and $b(r)$ has continuous first derivative in a neighborhood of zero. Suppose that $b(0) = 0$, $b(r) > 0$ for $r > 0$ and that for every $\tilde{r} > 0$, the infimum of $b(r)$ over $r \geq \tilde{r}$ as positive. Suppose also that there exist positive real numbers a_0, b_0, α, and β such that*

$$\lim_{r\to 0} a(r)r^{1-\alpha} = a_0, \quad \lim_{r\to 0} b(r)r^{-\beta} = b_0, \quad \textit{and} \quad \lim_{r\to 0} \frac{db}{dr}(r)r^{1-\beta} = b_0\beta.$$

Suppose also that $\int_0^\infty |a(r)|\exp(-Nb(r))\,dr < \infty$ for all sufficiently large N. Then, for $N \to \infty$,

$$\int_0^\infty a(r)\exp(-Nb(r))\,dr = \Gamma\left(\frac{\alpha}{\beta}\right)\frac{a_0}{\beta b_0^{\alpha/\beta}}\frac{1}{N^{\alpha/\beta}} + o\left(\frac{1}{N^{\alpha/\beta}}\right).$$

The proof follows from Theorem 7.1 in Olver (1997, p. 81).

PROOF OF LEMMA 1: First consider the conditional probability of unit i being the mth closest match to z, given $X_i = x$:

$$\begin{aligned}\Pr(j_m = i|X_i = x) = \binom{N-1}{m-1}&\left(\Pr(\|X - z\| > \|x - z\|)\right)^{N-m}\\ &\times\left(\Pr(\|X - z\| \leq \|x - z\|)\right)^{m-1}.\end{aligned}$$

Because the marginal probability of unit i being the mth closest match to z is $\Pr(j_m = i) = 1/N$ and because the density of X_i is $f(x)$, then the distribution of X_i conditional on it being the mth closest match is

$$\begin{aligned}f_{X_i|j_m=i}(x) &= Nf(x)\Pr(j_m = i|X_i = x)\\ &= Nf(x)\binom{N-1}{m-1}\left(1 - \Pr(\|X - z\| \leq \|x - z\|)\right)^{N-m}\\ &\quad\times\left(\Pr(\|X - z\| \leq \|x - z\|)\right)^{m-1},\end{aligned}$$

and this is also the distribution of X_{j_m}. Now transform to the matching discrepancy $U_m = X_{j_m} - z$ to get

$$f_{U_m}(u) = N\binom{N-1}{m-1} f(z+u)\big(1-\Pr(\|X-z\| \le \|u\|)\big)^{N-m} \times \big(\Pr(\|X-z\| \le \|u\|)\big)^{m-1}. \tag{A.1}$$

Transform to $V_m = N^{1/k}U_m$ with Jacobian N^{-1} to obtain

$$\begin{aligned} f_{V_m}(v) &= \binom{N-1}{m-1} f\left(z+\frac{v}{N^{1/k}}\right)\left(1-\Pr\left(\|X-z\| \le \frac{\|v\|}{N^{1/k}}\right)\right)^{N-m} \\ &\quad \times \left(\Pr\left(\|X-z\| \le \frac{\|v\|}{N^{1/k}}\right)\right)^{m-1} \\ &= N^{1-m}\binom{N-1}{m-1} f\left(z+\frac{v}{N^{1/k}}\right) \\ &\quad \times \left(1-\Pr\left(\|X-z\| \le \frac{\|v\|}{N^{1/k}}\right)\right)^{N}(1+o(1)) \\ &\quad \times \left(N\Pr\left(\|X-z\| \le \frac{\|v\|}{N^{1/k}}\right)\right)^{m-1}. \end{aligned}$$

Note that

$$\Pr\big(\|X-z\| \le \|v\|N^{-1/k}\big) = \int_0^{\|v\|/N^{1/k}} r^{k-1}\left(\int_{\mathbb{S}_k} f(z+r\omega)\lambda_{\mathbb{S}_k}(d\omega)\right)dr,$$

where as before $\mathbb{S}_k = \{\omega \in \mathbb{R}^k : \|\omega\| = 1\}$ is the unit k sphere, and $\lambda_{\mathbb{S}_k}$ is its surface measure. The derivative of $\Pr(\|X-z\| \le \|v\|N^{-1/k})$ with respect to N is

$$\left(-\frac{1}{N^2}\right)\frac{\|v\|^k}{k}\int_{\mathbb{S}_k} f\left(z+\frac{\|v\|^k}{N^{1/k}}\omega\right)\lambda_{\mathbb{S}_k}(d\omega).$$

Therefore, by l'Hospital's rule,

$$\lim_{N\to\infty}\frac{\Pr(\|X-z\| \le \|v\|N^{-1/k})}{1/N} = \frac{\|v\|^k}{k}f(z)\int_{\mathbb{S}_k}\lambda_{\mathbb{S}_k}(d\omega).$$

In addition, it is easy to check that for fixed m,

$$N^{1-m}\binom{N-1}{m-1} = \frac{1}{(m-1)!} + o(1).$$

Therefore,

$$\lim_{N\to\infty} f_{V_m}(v) = \frac{f(z)}{(m-1)!}\left(\|v\|^k \frac{f(z)}{k}\int_{\mathbb{S}_k}\lambda_{\mathbb{S}_k}(d\omega)\right)^{m-1} \times \exp\left(-\|v\|^k \frac{f(z)}{k}\int_{\mathbb{S}_k}\lambda_{\mathbb{S}_k}(d\omega)\right).$$

The previous equation shows that the density of V_m converges pointwise to a nonnegative function that is rotation invariant with respect to the origin. As a result, the matching discrepancy U_m is $O_p(N^{-1/k})$ and the limiting distribution of $N^{1/k}U_m$ is rotation invariant with respect to the origin. This finishes the proof of the first result.

Next, given $f_{U_m}(u)$ in (A.1),

$$\mathbb{E}[U_m] = N\binom{N-1}{m-1}A_m,$$

where

$$A_m = \int_{\mathbb{R}^k} uf(z+u)\big(1-\Pr(\|X-z\|\le\|u\|)\big)^{N-m} \times \big(\Pr(\|X-z\|\le\|u\|)\big)^{m-1}\,du.$$

Boundedness of $\mathbb{X}$ implies that A_m converges absolutely. It is easy to relax the bounded support condition here. We maintain it because it is used elsewhere in the article. Changing variables to polar coordinates gives

$$A_m = \int_0^\infty r^{k-1}\left(\int_{\mathbb{S}_k} r\omega f(z+r\omega)\lambda_{\mathbb{S}_k}(d\omega)\right) \times \big(1-\Pr(\|X-z\|\le r)\big)^{N-m}\big(\Pr(\|X-z\|\le r)\big)^{m-1}\,dr.$$

Then, rewriting the probability $\Pr(\|X-z\|\le r)$ as

$$\int_{\mathbb{R}^k} f(x)\mathbb{1}\{\|x-z\|\le r\}\,dx = \int_{\mathbb{R}^k} f(z+v)\mathbb{1}\{\|v\|\le r\}\,dv = \int_0^r s^{k-1}\left(\int_{\mathbb{S}_k} f(z+s\omega)\lambda_{\mathbb{S}_k}(d\omega)\right)ds$$

and substituting this into the expression for A_m gives

$$A_m = \int_0^\infty r^{k-1}\left(\int_{\mathbb{S}_k} r\omega f(z+r\omega)\lambda_{\mathbb{S}_k}(d\omega)\right) \times \left(1-\int_0^r s^{k-1}\left(\int_{\mathbb{S}_k} f(z+s\omega)\lambda_{\mathbb{S}_k}(d\omega)\right)ds\right)^{N-m}$$

$$\times \left(\int_0^r s^{k-1} \left(\int_{\mathbb{S}_k} f(z+s\omega) \lambda_{\mathbb{S}_k}(d\omega) \right) ds \right)^{m-1} dr$$

$$= \int_0^\infty e^{-Nb(r)} a(r)\, dr,$$

where

$$b(r) = -\log\left(1 - \int_0^r s^{k-1} \left(\int_{\mathbb{S}_k} f(z+s\omega) \lambda_{\mathbb{S}_k}(d\omega) \right) ds \right)$$

and

$$a(r) = r^k \left(\int_{\mathbb{S}_k} \omega f(z+r\omega) \lambda_{\mathbb{S}_k}(d\omega) \right)$$
$$\times \frac{\left(\int_0^r s^{k-1} \left(\int_{\mathbb{S}_k} f(z+s\omega) \lambda_{\mathbb{S}_k}(d\omega) \right) ds \right)^{m-1}}{\left(1 - \int_0^r s^{k-1} \left(\int_{\mathbb{S}_k} f(z+s\omega) \lambda_{\mathbb{S}_k}(d\omega) \right) ds \right)^m}.$$

That is, $a(r) = r^k c(r) g(r)^{m-1}$, where

$$c(r) = \frac{\int_{\mathbb{S}_k} \omega f(z+r\omega) \lambda_{\mathbb{S}_k}(d\omega)}{1 - \int_0^r s^{k-1} \left(\int_{\mathbb{S}_k} f(z+s\omega) \lambda_{\mathbb{S}_k}(d\omega) \right) ds},$$

$$g(r) = \frac{\int_0^r s^{k-1} \left(\int_{\mathbb{S}_k} f(z+s\omega) \lambda_{\mathbb{S}_k}(d\omega) \right) ds}{1 - \int_0^r s^{k-1} \left(\int_{\mathbb{S}_k} f(z+s\omega) \lambda_{\mathbb{S}_k}(d\omega) \right) ds}.$$

First notice that $b(r)$ is continuous in a neighborhood of zero and $b(0) = 0$. By Theorem 6.20 in Rudin (1976), $s^{k-1} \int_{\mathbb{S}_k} f(z+s\omega) \lambda_{\mathbb{S}_k}(d\omega)$ is continuous in s and

$$\frac{db}{dr}(r) = \frac{r^{k-1} \left(\int_{\mathbb{S}_k} f(z+r\omega) \lambda_{\mathbb{S}_k}(d\omega) \right)}{1 - \int_0^r s^{k-1} \left(\int_{\mathbb{S}_k} f(z+s\omega) \lambda_{\mathbb{S}_k}(d\omega) \right) ds},$$

which is continuous in r. Using l'Hospital's rule,

$$\lim_{r \to 0} b(r) r^{-k} = \lim_{r \to 0} \frac{1}{k r^{k-1}} \frac{db}{dr}(r) = \frac{1}{k} f(z) \int_{\mathbb{S}_k} \lambda_{\mathbb{S}_k}(d\omega).$$

Similarly, $c(r)$ is continuous in a neighborhood of zero, $c(0) = 0$, and

$$\lim_{r \to 0} c(r) r^{-1} = \lim_{r \to 0} \frac{dc}{dr}(r) = \int_{\mathbb{S}_k} \omega \omega' \lambda_{\mathbb{S}_k}(d\omega) \frac{\partial f}{\partial x}(z)$$
$$= \int_{\mathbb{S}_k} \lambda_{\mathbb{S}_k}(d\omega) \frac{I_k}{k} \frac{\partial f}{\partial x}(z) = \frac{1}{k} \frac{\partial f}{\partial x}(z) \int_{\mathbb{S}_k} \lambda_{\mathbb{S}_k}(d\omega).$$

Similarly, $g(r)$ is continuous in a neighborhood of zero and $g(0) = 0$, and

$$\lim_{r\to 0} g(r)r^{-k} = \lim_{r\to 0} \frac{1}{kr^{k-1}} \frac{dg}{dr}(r) = \frac{1}{k} f(z) \int_{\mathbb{S}_k} \lambda_{\mathbb{S}_k}(d\omega).$$

Therefore,

$$\lim_{r\to 0} g(r)^{m-1} r^{-(m-1)k} = \left(\lim_{r\to 0} \frac{g(r)}{r^k}\right)^{m-1} = \left(\frac{1}{k} f(z) \int_{\mathbb{S}_k} \lambda_{\mathbb{S}_k}(d\omega)\right)^{m-1}.$$

Now, it is clear that

$$\begin{aligned}
\lim_{r\to 0} a(r)r^{-(mk+1)} &= \left(\lim_{r\to 0} g(r)^{m-1} r^{-(m-1)k}\right)\left(\lim_{r\to 0} c(r)r^{-1}\right) \\
&= \left(\frac{1}{k} f(z) \int_{\mathbb{S}_k} \lambda_{\mathbb{S}_k}(d\omega)\right)^{m-1} \frac{1}{k} \frac{\partial f}{\partial x}(z) \int_{\mathbb{S}_k} \lambda_{\mathbb{S}_k}(d\omega) \\
&= \left(\frac{1}{k} f(z) \int_{\mathbb{S}_k} \lambda_{\mathbb{S}_k}(d\omega)\right)^{m} \frac{1}{f(z)} \frac{\partial f}{\partial x}(z).
\end{aligned}$$

Therefore, the conditions of Lemma A.1 hold for $\alpha = mk + 2$, $\beta = k$,

$$a_0 = \left(\frac{1}{k} f(z) \int_{\mathbb{S}_k} \lambda_{\mathbb{S}_k}(d\omega)\right)^{m} \frac{1}{f(z)} \frac{\partial f}{\partial x}(z)$$

and

$$b_0 = \frac{1}{k} f(z) \int_{\mathbb{S}_k} \lambda_{\mathbb{S}_k}(d\omega).$$

Applying Lemma A.1, we get

$$\begin{aligned}
A_m &= \Gamma\left(\frac{mk+2}{k}\right) \frac{a_0}{k b_0^{(mk+2)/k}} \frac{1}{N^{(mk+2)/k}} \\
&\quad + o\left(\frac{1}{N^{(mk+2)/k}}\right) \\
&= \Gamma\left(\frac{mk+2}{k}\right) \frac{1}{k} \left(f(z) \frac{\pi^{k/2}}{\Gamma(1+k/2)}\right)^{-2/k} \frac{1}{f(z)} \frac{df}{dx}(z) \frac{1}{N^{(mk+2)/k}} \\
&\quad + o\left(\frac{1}{N^{(mk+2)/k}}\right).
\end{aligned}$$

Therefore,

$$\mathbb{E}[U_m] = \Gamma\left(\frac{mk+2}{k}\right)\frac{1}{(m-1)!k}\left(f(z)\frac{\pi^{k/2}}{\Gamma(1+k/2)}\right)^{-2/k} \times \frac{1}{f(z)}\frac{df}{dx}(z)\frac{1}{N^{2/k}} + o\left(\frac{1}{N^{2/k}}\right),$$

which finishes the proof for the second result of the lemma. The results for $\mathbb{E}[U_m U_m']$ and $\mathbb{E}[\|U_m\|^3]$ follow from similar arguments. *Q.E.D.*

The proof of Lemma 2 is available on the authors' webpages.

PROOF OF THEOREM 1(i): Let the unit-level matching discrepancy $U_{m,i} = X_i - X_{j_m(i)}$. Define the unit-level conditional bias from the mth match as

$$\begin{aligned} B_{m,i} &= W_i\big(\mu_0(X_i) - \mu_0(X_{j_m(i)})\big) - (1-W_i)\big(\mu_1(X_i) - \mu_1(X_{j_m(i)})\big) \\ &= W_i\big(\mu_0(X_i) - \mu_0(X_i + U_{m,i})\big) \\ &\quad - (1-W_i)(\mu_1(X_i) - \mu_1(X_i + U_{m,i})). \end{aligned}$$

By the Lipschitz assumption on μ_0 and μ_1, we obtain $|B_{m,i}| \le C_1\|U_{m,i}\|$ for some positive constant C_1. The bias term is

$$B_M = \frac{1}{NM}\sum_{i=1}^{N}\sum_{m=1}^{M} B_{m,i}.$$

Using the Cauchy–Schwarz inequality and Lemma 2,

$$\begin{aligned} &\mathbb{E}[N^{2/k}(B_M)^2] \\ &\quad \le C_1^2 N^{2/k}\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\|U_{M,i}\|^2\right] \\ &\quad = C_1^2 N^{2/k-1}\mathbb{E}\left[\frac{1}{N_0^{2/k}}\sum_{W_i=1}\mathbb{E}\big[N_0^{2/k}\|U_{M,i}\|^2|W_1,\ldots,W_N,X_i\big]\right. \\ &\qquad\qquad \left. + \frac{1}{N_1^{2/k}}\sum_{W_i=0}\mathbb{E}\big[N_1^{2/k}\|U_{M,i}\|^2|W_1,\ldots,W_N,X_i\big]\right] \\ &\quad \le C_2\mathbb{E}\left[\left(\frac{N}{N_0}\right)^{2/k}\frac{N_1}{N} + \left(\frac{N}{N_1}\right)^{2/k}\frac{N_0}{N}\right] \end{aligned}$$

for some positive constant C_2. Using Chernoff's inequality, it can be seen that any moment of N/N_1 or N/N_0 is uniformly bounded in N (with $N_w \ge M$ for

$w = 0, 1$). The result of the theorem follows now from Markov's inequality. This proves part (i) of the theorem. We defer the proof of Theorem 1(ii) until after the proof of Theorem 2(ii), because the former will follow directly from the latter. *Q.E.D.*

LEMMA A.2: *Let X be distributed with density $f(x)$ on some compact set $\mathbb{X}$ of dimension k: $\mathbb{X} \subset \mathbb{R}^k$. Let $\mathbb{Z}$ be a compact set of dimension k that is a subset of* $\operatorname{int}\mathbb{X}$. *Suppose that $f(x)$ is bounded and bounded away from zero on $\mathbb{X}$, $0 < \underline{f} \leq f(x) \leq \bar{f} < \infty$ for all $x \in \mathbb{X}$. Suppose also that $f(x)$ is differentiable in the interior of $\mathbb{X}$ with bounded derivatives* $\sup_{x \in \operatorname{int}\mathbb{X}} \|\partial f(x)/\partial X\| < \infty$. *Then $N^{2/k}\|E[U_m]\|$ is bounded by a constant uniformly over $z \in \mathbb{Z}$ and $N > m$.*

The proof of Lemma A.2 is available on the authors' webpages.

PROOF OF THEOREM 2: The proof of the first part of Theorem 2 is very similar to the proof of Theorem 1(i) and therefore is omitted.

Consider the second part:

$$\mathbb{E}[B'_M] = \mathbb{E}\left[\frac{1}{N_1 M}\sum_{i=1}^{N}\sum_{m=1}^{M} W_i\big(\mu_0(X_i) - \mu_0(X_{j_m(i)})\big)\right]$$
$$= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\big[\mu_0(X_i) - \mu_0(X_{j_m(i)})|W_i = 1\big].$$

Applying a second-order Taylor expansion, we obtain

$$\mu_0(X_{j_m(i)}) - \mu_0(X_i)$$
$$= \frac{\partial \mu_0}{\partial x'}(X_i)U_{m,i} + \frac{1}{2}\operatorname{tr}\left(\frac{\partial^2 \mu_0}{\partial x\,\partial x'}(X_i)U_{m,i}U'_{m,i}\right) + O(\|U_{m,i}\|^3).$$

Therefore, because the trace is a linear operator,

$$\mathbb{E}\big[\mu_0(X_{j_m(i)}) - \mu_0(X_i)|X_i = z, W_i = 1\big]$$
$$= \frac{\partial \mu_0}{\partial x'}(z)\mathbb{E}[U_{m,i}|X_i = z, W_i = 1]$$
$$+ \frac{1}{2}\operatorname{tr}\left(\frac{\partial^2 \mu_0}{\partial x\,\partial x'}(z)\mathbb{E}[U_{m,i}U'_{m,i}|X_i = z, W_i = 1]\right)$$
$$+ O\big(\mathbb{E}\big[\|U_{m,i}\|^3|X_i = z, W_i = 1\big]\big).$$

Lemma 2 implies that the norms of $N_0^{2/k}\mathbb{E}[U_{m,i}U'_{m,i}|X_i = z, W_i = 1]$ and $N_0^{2/k}\mathbb{E}[\|U_{m,i}\|^3|X_i = z, W_i = 1]$ are uniformly bounded over $z \in \mathbb{X}_1$ and N_0.

Lemma A.2 implies the same result for $N_0^{2/k}\mathbb{E}[U_{m,i}|X_i = z, W_i = 1]$. As a result, $\|N_0^{2/k}\mathbb{E}[\mu_0(X_{j_m(i)}) - \mu_0(X_i)|X_i = z, W_i = 1]\|$ is uniformly bounded over $z \in \mathbb{X}_1$ and N_0. Applying Lebesgue's dominated convergence theorem along with Lemma 1, we obtain

$$
\begin{aligned}
&N_0^{2/k}\mathbb{E}\big[\mu_0(X_{j_m(i)}) - \mu_0(X_i)|W_i = 1\big] \\
&\quad = \Gamma\left(\frac{mk+2}{k}\right)\frac{1}{(m-1)!k} \\
&\qquad \times \int \left(f_0(x)\frac{\pi^{k/2}}{\Gamma(1+k/2)}\right)^{-2/k} \\
&\qquad\quad \times \left\{\frac{1}{f_0(x)}\frac{\partial f_0}{\partial x'}(x)\frac{\partial \mu_0}{\partial x}(x) + \frac{1}{2}\operatorname{tr}\left(\frac{\partial^2 \mu_0}{\partial x' \partial x}(x)\right)\right\} f_1(x)\,dx \\
&\qquad + o(1).
\end{aligned}
$$

Now the result follows easily from the conditions of the theorem. *Q.E.D.*

PROOF OF THEOREM 1(ii): Consider the special case where $\mu_1(x)$ is flat over $\mathbb{X}$ and $\mu_0(x)$ is flat in a neighborhood of the boundary, $\mathbb{B}$. Then matching the control units does not create bias. Matching the treated units creates a bias that is similar to the formula in Theorem 2(ii), but with $r = 1$, $\theta = p/(1-p)$, and the integral taken over $\mathbb{X} \cap \mathbb{B}^c$. *Q.E.D.*

PROOF OF LEMMA 3: Define $\underline{f} = \inf_{x,w} f_w(x)$ and $\bar{f} = \sup_{x,w} f_w(x)$, with $\underline{f} > 0$ and $\bar{f}$ finite. Let $\bar{u} = \sup_{x,y\in\mathbb{X}} \|x - y\|$. Consider the ball $B(x, u)$ with center $x \in \mathbb{X}$ and radius u. Let $c(u)$ $(0 < c(u) < 1)$ be the infimum over $x \in \mathbb{X}$ of the proportion that the intersection with $\mathbb{X}$ represents in volume of the balls. Note that, because $\mathbb{X}$ is convex, this proportion is nonincreasing in u, so let $\underline{c} = c(\bar{u})$ and $c(u) \geq \underline{c}$ for $u \leq \bar{u}$. The proof consists of three parts. First we derive an exponential bound for the probability that the distance to a match, $\|X_{j_m(i)} - X_i\|$, exceeds some value. Second, we use this to obtain an exponential bound on the volume of the catchment area, $\mathbb{A}_M(i)$, defined as the subset of $\mathbb{X}$ such that i is matched to each observation, j, with $W_j = 1 - W_i$ and $X_j \in \mathbb{A}_M(i)$. Formally,

$$
\mathbb{A}_M(i) = \left\{x \middle| \sum_{l|W_l=W_i} \mathbb{1}\{\|X_l - x\| \leq \|X_i - x\|\} \leq M\right\}.
$$

Thus, if $W_j = 1 - W_i$ and $X_j \in \mathbb{A}_M(i)$, then $i \in J_M(j)$. Third, we use the exponential bound on the volume of the catchment area to derive an exponential

bound on the probability of a large $K_M(i)$, which will be used to bound the moments of $K_M(i)$.

For the first part we bound the probability of the distance to a match. Let $x \in \mathbb{X}$ and $u < N_{1-W_i}^{1/k}\bar{u}$. Then

$$\begin{aligned}
&\Pr\bigl(\|X_j - X_i\| > uN_{1-W_i}^{-1/k} \big| W_1, \ldots, W_N, W_j = 1 - W_i, X_i = x\bigr) \\
&\quad = 1 - \int_0^{uN_{1-W_i}^{-1/k}} r^{k-1} \int_{S_k} f_{1-W_i}(x + r\omega)\lambda_{S_k}(d\omega)\, dr \\
&\quad \le 1 - \underline{c}\underline{f} \int_0^{uN_{1-W_i}^{-1/k}} r^{k-1} \int_{S_k} \lambda_{S_k}(d\omega)\, dr \\
&\quad = 1 - \underline{c}\underline{f} u^k N_{1-W_i}^{-1} \frac{\pi^{k/2}}{\Gamma(1 + k/2)}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\Pr\bigl(\|X_j - X_i\| \le uN_{1-W_i}^{-1/k} \big| W_1, \ldots, W_N, W_j = 1 - W_i, X_i = x\bigr) \\
&\quad \le \bar{f} u^k N_{1-W_i}^{-1} \frac{\pi^{k/2}}{\Gamma(1 + k/2)}.
\end{aligned}$$

Notice also that

$$\begin{aligned}
&\Pr\bigl(\|X_j - X_i\| > uN_{1-W_i}^{-1/k} \big| W_1, \ldots, W_N, X_i = x, j \in \mathcal{J}_M(i)\bigr) \\
&\quad \le \Pr\bigl(\|X_j - X_i\| > uN_{1-W_i}^{-1/k} \big| W_1, \ldots, W_N, X_i = x, j = j_M(i)\bigr) \\
&\quad = \sum_{m=0}^{M-1} \binom{N_{1-W_i}}{m} \Pr\bigl(\|X_j - X_i\| > uN_{1-W_i}^{-1/k} \big| \\
&\qquad\qquad W_1, \ldots, W_N, W_j = 1 - W_i, X_i = x\bigr)^{N_{1-W_i} - m} \\
&\qquad \times \Pr\bigl(\|X_j - X_i\| \le uN_{1-W_i}^{-1/k} \big| \\
&\qquad\qquad \times W_1, \ldots, W_N, W_j = 1 - W_i, X_i = x\bigr)^m.
\end{aligned}$$

In addition,

$$\begin{aligned}
&\binom{N_{1-W_i}}{m} \Pr\bigl(\|X_j - X_i\| \le uN_{1-W_i}^{-1/k} \big| W_1, \ldots, W_N, W_j = 1 - W_i, X_i = x\bigr)^m \\
&\quad \le \frac{1}{m!} \left(u^k \bar{f} \frac{\pi^{k/2}}{\Gamma(1 + k/2)} \right)^m.
\end{aligned}$$

Therefore,

$$\Pr\big(\|X_j - X_i\| > uN_{1-W_i}^{-1/k} \big| W_1, \ldots, W_N, X_i = x, j \in \mathcal{J}_M(i)\big)$$
$$\leq \sum_{m=0}^{M-1} \frac{1}{m!} \left(u^k \bar{f} \frac{\pi^{k/2}}{\Gamma(1+k/2)} \right)^m$$
$$\times \left(1 - u^k \underline{c}\, \underline{f} \frac{\pi^{k/2}}{\Gamma(1+k/2)} \cdot \frac{1}{N_{1-W_i}} \right)^{N_{1-W_i}-m}.$$

Then, for some constant $C_1 > 0$,

$$\Pr\big(\|X_j - X_i\| > uN_{1-W_i}^{-1/k} \big| W_1, \ldots, W_N, X_i = x, j \in \mathcal{J}_M(i)\big)$$
$$\leq C_1 \max\{1, u^{k(M-1)}\} \sum_{m=0}^{M-1} \left(1 - u^k \underline{c}\, \underline{f} \frac{\pi^{k/2}}{\Gamma(1+k/2)} \cdot \frac{1}{N_{1-W_i}} \right)^{N_{1-W_i}-m}$$
$$\leq C_1 M \max\{1, u^{k(M-1)}\} \exp\left(-\frac{u^k}{(M+1)} \underline{c}\, \underline{f} \frac{\pi^{k/2}}{\Gamma(1+k/2)} \right).$$

Notice that this bound also holds for $u \geq N_{1-W_i}^{1/k} \bar{u}$, because in that case the probability that $\|X_{j_m(i)} - X_i\| > uN_{1-W_i}^{-1/k}$ is zero.

Next, we consider for unit i, the volume $B_M(i)$ of the catchment area $\mathbb{A}_M(i)$, defined as $B_{M(i)} = \int_{\mathbb{A}_M(i)} dx$. Conditional on $W_1, \ldots, W_N$, $i \in \mathcal{J}_M(j)$, $X_i = x$, and $\mathbb{A}_M(i)$, the distribution of X_j is proportional to $f_{1-W_i}(x) \mathbb{1}\{x \in \mathbb{A}_M(i)\}$. Notice that a ball with radius $(b/2)^{1/k}/(\pi^{k/2}/\Gamma(1+k/2))^{1/k}$ has volume $b/2$. Therefore, for X_i in $\mathbb{A}_M(i)$ and $B_M(i) \geq b$, we obtain

$$\Pr\left(\|X_j - X_i\| > \frac{(b/2)^{1/k}}{(\pi^{k/2}/\Gamma(1+k/2))^{1/k}} \Big| \right.$$
$$\left. W_1, \ldots, W_N, X_i = x, \mathbb{A}_M(i), B_M(i) \geq b, i \in \mathcal{J}_M(j) \right) \geq \frac{\underline{f}}{2\bar{f}}.$$

The last inequality does not depend on $\mathbb{A}_m(i)$ (given $B_M(i) \geq b$). Therefore,

$$\Pr\left(\|X_j - X_i\| > \frac{(b/2)^{1/k}}{(\pi^{k/2}/\Gamma(1+k/2))^{1/k}} \Big| \right.$$
$$\left. W_1, \ldots, W_N, X_i = x, i \in \mathcal{J}_M(j), B_M(i) \geq b \right) \geq \frac{\underline{f}}{2\bar{f}}.$$

As a result, if

$$\text{(A.2)} \qquad \Pr\Bigg(\|X_j - X_i\| > \frac{(b/2)^{1/k}}{(\pi^{k/2}/\Gamma(1+k/2))^{1/k}} \Big| \; W_1, \ldots, W_N, X_i = x, i \in \mathcal{J}_M(j) \Bigg) \leq \delta \frac{\underline{f}}{2\bar{f}},$$

then it must be the case that $\Pr(B_M(i) \geq b | W_1, \ldots, W_N, X_i = x, i \in \mathcal{J}_M(j)) \leq \delta$. In fact, inequality (A.2) has been established above for

$$b = \frac{2u^k}{N_{W_i}} \left(\frac{\pi^{k/2}}{\Gamma(1+k/2)} \right)$$

and

$$\delta = \frac{2\bar{f}}{\underline{f}} C_1 M \max\{1, u^{k(M-1)}\} \exp\left(-\frac{u^k}{(M+1)} \frac{c}{} \underline{f} \frac{\pi^{k/2}}{\Gamma(1+k/2)} \right).$$

Let $t = 2u^k \pi^{k/2}/\Gamma(1+k/2)$. Then

$$\begin{aligned} &\Pr\big(N_{W_i} B_M(i) \geq t | W_1, \ldots, W_N, X_i = x, i \in \mathcal{J}_M(j)\big) \\ &\quad \leq C_2 \max\{1, C_3 t^{M-1}\} \exp(-C_4 t) \end{aligned}$$

for some positive constants, C_2, C_3, and C_4. This establishes an uniform exponential bound, so all the moments of $N_{W_i} B_M(i)$ exist conditional on $W_1, \ldots, W_N, X_i = x, i \in \mathcal{J}_M(j)$ (uniformly in N).

For the third part of the proof, consider the distribution of $K_M(i)$, the number of times unit i is used as a match. Let $P_M(i)$ be the probability that an observation with the opposite treatment is matched to observation i conditional on $\mathbb{A}_M(i)$:

$$P_M(i) = \int_{\mathbb{A}_M(i)} f_{1-W_i}(x)\, dx \leq \bar{f} B_M(i).$$

Note that for $n \geq 0$,

$$\begin{aligned} &\mathbb{E}\big[(N_{W_i} P_M(i))^n \big| X_i = x, W_1, \ldots, W_N\big] \\ &\quad \leq \mathbb{E}\big[(N_{W_i} P_M(i))^n \big| X_i = x, W_1, \ldots, W_N, i \in \mathcal{J}_M(j)\big] \\ &\quad \leq \bar{f}^n \mathbb{E}\big[(N_{W_i} B_M(i))^n \big| X_i = x, W_1, \ldots, W_N, i \in \mathcal{J}_M(j)\big]. \end{aligned}$$

As a result, $\mathbb{E}[(N_{W_i}P_M(i))^n|X_i = x, W_1, \ldots, W_N]$ is uniformly bounded. Conditional on $P_M(i)$ and on $X_i = x, W_1, \ldots, W_N$, the distribution of $K_M(i)$ is binomial with parameters N_{1-W_i} and $P_M(i)$. Therefore, conditional on $P_M(i)$ and $X_i = x, W_1, \ldots, W_N$, the qth moment of $K_M(i)$ is

$$\mathbb{E}\big[K_M^q(i)|P_M(i), X_i = x, W_1, \ldots, W_N\big] = \sum_{n=0}^{q} \frac{S(q,n)N_{1-W_i}!P_M(i)^n}{(N_{1-W_i}-n)!} \leq \sum_{n=0}^{q} S(q,n)\big(N_{1-W_i}P_M(i)\big)^n,$$

where $S(q,n)$ are Stirling numbers of the second kind and $q \geq 1$ (see, e.g., Johnson, Kotz, and Kemp (1992)). Then, because $S(q,0) = 0$ for $q \geq 1$,

$$\mathbb{E}\big[K_M^q(i)|X_i = x, W_1, \ldots, W_N\big] \leq C\sum_{n=1}^{q} S(q,n)\left(\frac{N_{1-W_i}}{N_{W_i}}\right)^n$$

for some positive constant C. Using Chernoff's bound for binomial tails, it can be easily seen that $E[(N_{1-W_i}/N_{W_i})^n|X_i = x, W_i] = E[(N_{1-W_i}/N_{W_i})^n|W_i]$ is uniformly bounded in N for all $n \geq 1$, so the result of the first part of the lemma follows. Because $K_M(i)^q \leq K_M(i)$ for $0 < q < 1$, this proof applies also to the case with $0 < q < 1$.

Next, consider part (ii) of Lemma 3. Because the variance $\sigma^2(x,w)$ is Lipschitz on a bounded set, it is therefore bounded by some constant, $\bar{\sigma}^2 = \sup_{w,x}\sigma^2(x,w)$. As a result, $\mathbb{E}[(1 + K_M/M)^2\sigma^2(x,w)]$ is bounded by $\bar{\sigma}^2\mathbb{E}[(1 + K_M/M)^2]$, which is uniformly bounded in N by the result in the first part of the lemma. Hence $\mathbb{E}[V^E] = O(1)$.

Next, consider part (iii) of Lemma 3. Using the same argument as for $\mathbb{E}[K_M^q(i)]$, we obtain

$$\mathbb{E}[K_M^q(i)|W_i = 0] \leq \sum_{n=1}^{q} S(q,n)\left(\frac{N_1}{N_0}\right)^n \mathbb{E}\big[(N_0P_M(i))^n|W_i = 0\big].$$

Therefore,

$$\left(\frac{N_0}{N_1}\right)\mathbb{E}[K_M^q(i)|W_i = 0] \leq \sum_{n=1}^{q} S(q,n)\left(\frac{N_1}{N_0}\right)^{n-1} \mathbb{E}\big[(N_0P_M(i))^n|W_i = 0\big],$$

which is uniformly bounded because $r \geq 1$.

For part (iv) notice that

$$\begin{aligned}\mathbb{E}[V^{E,t}] &= \mathbb{E}\left[\frac{1}{N_1}\sum_{i=1}^{N} W_i\sigma^2(X_i,W_i)\right] \\ &\quad + E\left[\frac{1}{N_1}\sum_{i=1}^{N}(1-W_i)\left(\frac{K_M(i)}{M}\right)^2\sigma^2_{W_i}(X_i)\right] \\ &\leq \bar{\sigma}^2 + \bar{\sigma}^2\left(\frac{N_0}{N_1}\right)E\left[\left(\frac{K_M(i)}{M}\right)^2\Big|W_i=0\right].\end{aligned}$$

Therefore, $\mathbb{E}[V^{E,t}]$ is uniformly bounded. *Q.E.D.*

PROOF OF THEOREM 3: We only prove the first part of the theorem. The second part follows the same argument. We can write $\widehat{\tau}_M - \tau = (\overline{\tau(X)} - \tau) + E_M + B_M$. We consider each of the three terms separately. First, by Assumptions 1 and 4(i), $\mu_w(x)$ is bounded over $x \in \mathbb{X}$ and $w = 0, 1$. Hence $\mu_1(X) - \mu_0(X) - \tau$ has mean zero and finite variance. Therefore, by a standard law of large numbers, $\overline{\tau(X)} - \tau \xrightarrow{p} 0$. Second, by Theorem 1, $B_M = O_p(N^{-1/k}) = o_p(1)$. Finally, because $\mathbb{E}[\varepsilon_i^2|\mathbf{X},\mathbf{W}] \leq \bar{\sigma}^2$ and $\mathbb{E}[\varepsilon_i\varepsilon_j|\mathbf{X},\mathbf{W}] = 0$ $(i \neq j)$, we obtain

$$\begin{aligned}\mathbb{E}\big[(\sqrt{N}E_M)^2\big] &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left(1+\frac{K_M(i)}{M}\right)^2\varepsilon_i^2\right] \\ &= \mathbb{E}\left[\left(1+\frac{K_M(i)}{M}\right)^2\sigma^2(X_i,W_i)\right] = O(1),\end{aligned}$$

where the last equality comes from Lemma 3. By Markov's inequality $E_M = O_p(N^{-1/2}) = o_p(1)$. *Q.E.D.*

PROOF OF THEOREM 4: We only prove the first assertion in the theorem because the second follows the same argument. We can write $\sqrt{N}(\widehat{\tau}_M - B_M - \tau) = \sqrt{N}(\overline{\tau(X)} - \tau) + \sqrt{N}E_M$. First, consider the contribution of $\sqrt{N}(\overline{\tau(X)} - \tau)$. By a standard central limit theorem,

$$\text{(A.3)} \qquad \sqrt{N}\big(\overline{\tau(X)} - \tau\big) \xrightarrow{d} \mathcal{N}(0, V^{\tau(X)}).$$

Second, consider the contribution of $\sqrt{N}E_M/\sqrt{V^E} = \sum_{i=1}^{N} E_{M,i}/\sqrt{NV^E}$. Conditional on $\mathbf{W}$ and $\mathbf{X}$ the unit-level terms $E_{M,i} = (2W_i - 1)(1 + K_M(i)/M)\varepsilon_i$ are independent with zero means and nonidentical distributions. The conditional variance of $E_{M,i}$ is $(1 + K_M(i)/M)^2\sigma^2(X_i, W_i)$. We will use a Lindeberg–

Feller central limit theorem for $\sqrt{N}E_M/\sqrt{V^E}$. For a given $\mathbf{X}, \mathbf{W}$, the Lindeberg–Feller condition requires that

$$\text{(A.4)} \qquad \frac{1}{NV^E}\sum_{i=1}^{N}\mathbb{E}\big[(E_{M,i})^2\mathbb{1}\big\{|E_{M,i}| \geq \eta\sqrt{NV^E}\big\}\big|\mathbf{X},\mathbf{W}\big] \to 0$$

for all $\eta > 0$. To prove that the (A.4) condition holds, notice that by Hölder's and Markov's inequalities we have

$$\begin{aligned}
&\mathbb{E}\big[(E_{M,i})^2\mathbb{1}\big\{|E_{M,i}| \geq \eta\sqrt{NV^E}\big\}\big|\mathbf{X},\mathbf{W}\big] \\
&\quad\leq \big(\mathbb{E}[(E_{M,i})^4|\mathbf{X},\mathbf{W}]\big)^{1/2}\big(\mathbb{E}\big[\mathbb{1}\big\{|E_{M,i}| \geq \eta\sqrt{NV^E}\big\}\big|\mathbf{X},\mathbf{W}\big]\big)^{1/2} \\
&\quad\leq \big(\mathbb{E}[(E_{M,i})^4|\mathbf{X},\mathbf{W}]\big)^{1/2}\big(\Pr\big(|E_{M,i}| \geq \eta\sqrt{NV^E}\,|\mathbf{X},\mathbf{W}\big)\big) \\
&\quad\leq \big(\mathbb{E}[(E_{M,i})^4|\mathbf{X},\mathbf{W}]\big)^{1/2}\frac{\mathbb{E}[(E_{M,i})^2|\mathbf{X},\mathbf{W}]}{\eta^2 NV^E}.
\end{aligned}$$

Let $\bar{\sigma}^2 = \sup_{w,x}\sigma^2(x,w) < \infty$, $\underline{\sigma}^2 = \inf_{w,x}\sigma^2(x,w) > 0$, and $\bar{C} = \sup_{w,x}\mathbb{E}[\varepsilon_i^4|X_i = x, W_i = w] < \infty$. Notice that $V^E \geq \underline{\sigma}^2$. Therefore,

$$\begin{aligned}
&\frac{1}{NV^E}\sum_{i=1}^{N}\mathbb{E}\big[(E_{M,i})^2\mathbb{1}\big\{|E_{M,i}| \geq \eta\sqrt{NV^E}\big\}\big|\mathbf{X},\mathbf{W}\big] \\
&\quad\leq \frac{1}{NV^E}\sum_{i=1}^{N}\left(\left(1+\frac{K_M(i)}{M}\right)^4\mathbb{E}[\varepsilon_i^4|\mathbf{X},\mathbf{W}]\right)^{1/2} \\
&\qquad\qquad\times\frac{(1+K_M(i)/M)^2\sigma^2(X_i,W_i)}{\eta^2 NV^E} \\
&\quad\leq \frac{\bar{\sigma}^2\bar{C}^{1/2}}{\eta^2\underline{\sigma}^4}\frac{1}{N}\left(\frac{1}{N}\sum_{i=1}^{N}\left(1+\frac{K_M(i)}{M}\right)^4\right).
\end{aligned}$$

Because $\mathbb{E}[(1+K_M(i)/M)^4]$ is uniformly bounded, by Markov's inequality, the factor in parentheses is bounded in probability. Hence, the Lindeberg–Feller condition is satisfied for almost all $\mathbf{X}$ and $\mathbf{W}$. As a result,

$$\frac{N^{1/2}\sum_{i=1}^{N}E_{M,i}}{\big(\sum_{i=1}^{N}(1+K_M(i)/M)^2\sigma^2(X_i,W_i)\big)^{1/2}} = \frac{N^{1/2}E_M}{\sqrt{V^E}} \xrightarrow{d} \mathcal{N}(0,1).$$

Finally, $\sqrt{N}E_M/\sqrt{V^E}$ and $\sqrt{N}(\overline{\tau(X)} - \tau)$ are asymptotically independent (the central limit theorem for $\sqrt{N}E_M/\sqrt{V^E}$ holds conditional on $\mathbf{X}$ and $\mathbf{W}$).

Thus, the fact that both converge to standard normal distributions, boundedness of V^E and $V^{\tau(X)}$, and boundedness away from zero of V^E imply that $(V^E + V^{\tau(X)})^{-1/2} N^{1/2} (\widehat{\tau}_M - B_M - \tau)$ converges to a standard normal distribution. *Q.E.D.*

The proofs of Theorems 5, 6, and 7 are available on the authors' webpages.

REFERENCES

ABADIE, A., D. DRUKKER, J. HERR, AND G. IMBENS (2004): "Implementing Matching Estimators for Average Treatment Effects in Stata," *The Stata Journal*, 4, 290–311.

ABADIE, A., AND G. IMBENS (2002): "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects," Technical Working Paper T0283, NBER.

——— (2005): "On the Failure of the Bootstrap for Matching Estimators," Mimeo, Kennedy School of Government, Harvard University.

BARNOW, B. S., G. G. CAIN, AND A. S. GOLDBERGER (1980): "Issues in the Analysis of Selectivity Bias," in *Evaluation Studies*, Vol. 5, ed. by E. Stromsdorfer and G. Farkas. San Francisco: Sage, 43–59.

COCHRAN, W., AND D. RUBIN (1973): "Controlling Bias in Observational Studies: A Review," *Sankhyā*, 35, 417–446.

DEHEJIA, R., AND S. WAHBA (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, 315–331.

HECKMAN, J., H. ICHIMURA, AND P. TODD (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294.

HECKMAN, J., AND R. ROBB (1984): "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, ed. by J. Heckman and B. Singer. Cambridge, U.K.: Cambridge University Press, 156–245.

HIRANO, K., G. IMBENS, AND G. RIDDER (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189.

IMBENS, G. (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Survey," *Review of Economics and Statistics*, 86, 4–30.

JOHNSON, N., S. KOTZ, AND A. KEMP (1992): *Univariate Discrete Distributions* (Second Ed.). New York: Wiley.

OKABE, A., B. BOOTS, K. SUGIHARA, AND S. NOK CHIU (2000): *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams* (Second Ed.). New York: Wiley.

OLVER, F. W. J. (1997): *Asymptotics and Special Functions* (Second Ed.). New York: Academic Press.

ROSENBAUM, P. (1995): *Observational Studies*. New York: Springer-Verlag.

ROSENBAUM, P., AND D. RUBIN (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

RUBIN, D. (1973): "Matching to Remove Bias in Observational Studies," *Biometrics*, 29, 159–183.

——— (1977): "Assignment to Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics*, 2, 1–26.

RUDIN, W. (1976): *Principles Mathematical Analysis* (Third Ed.). New York: McGraw-Hill.

STROOCK, D. W. (1994): *A Concise Introduction to the Theory of Integration*. Boston: Birkhäuser.