TECHNICAL WORKING PAPER SERIES

SIMPLE AND BIAS-CORRECTED MATCHING ESTIMATORS FOR AVERAGE TREATMENT EFFECTS

Alberto Abadie
Guido Imbens

Technical Working Paper 283
http://www.nber.org/papers/T0283

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2002

We wish to thank Gary Chamberlain, Geert Dhaene, Jin Hahn, Jim Heckman, Hide Ichimura, Whitney Newey, Jack Porter, Jim Powell, Paul Rosenbaum, Ed Vytlacil, and participants at seminars at Berkeley, Brown, Chicago, Harvard/MIT, McGill, Princeton, Yale, the 2001 EC2 conference in Louvain, and the 2002 conference on evaluation of social policies in Madrid for comments, and Don Rubin for many discussions on these topics. Financial support for this research was generously provided through NSF grants SBR-9818644 and SES 0136789 (Imbens). The views expressed in this paper are those of the authors and not necessarily those of the National Bureau of Economic Research.

© 2001 by Alberto Abadie and Guido Imbens. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Simple and Bias-Corrected Matching Estimators for Average Treatment Effects
Alberto Abadie and Guido Imbens
NBER Technical Working Paper No. 283
October 2002
JEL No. C100, C130, C140, J240, J310

ABSTRACT

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. In this article, we develop a new framework to analyze the properties of matching estimators and establish a number of new results. First, we show that matching estimators include a conditional bias term which may not vanish at a rate faster than root-N when more than one continuous variable is used for matching. As a result, matching estimators may not be root-N-consistent. Second, we show that even after removing the conditional bias, matching estimators with a fixed number of matches do not reach the semiparametric efficiency bound for average treatment effects, although the efficiency loss may be small. Third, we propose a bias-correction that removes the conditional bias asymptotically, making matching estimators root-N-consistent. Fourth, we provide a new estimator for the conditional variance that does not require consistent nonparametric estimation of unknown functions. We apply the bias-corrected matching estimators to the study of the effects of a labor market program previously analyzed by Lalonde (1986). We also carry out a small simulation study based on Lalonde's example where a simple implementation of the biascorrected matching estimator performs well compared to both simple matching estimators and to regression estimators in terms of bias and root-mean-squared-error. Software for implementing the proposed estimators in STATA and Matlab is available from the authors on the web.

Alberto Abadie
John F. Kennedy School of Government
Harvard University
79 John F. Kennedy Street
Cambridge, MA 02138
and NBER
alberto_abadie@harvard.edu

Guido Imbens
Department of Economics, and Department
of Agricultural and Resource Economics
University of California at Berkeley
661 Evans Hall #3880
Berkeley, CA 94720-3880
and NBER
imbens@econ.berkeley.edu

1. Introduction

Estimation of average treatment effects is an important goal of much evaluation research, both in academic studies (e.g, Ashenfelter and Card, 1985; Lalonde, 1986; Heckman, Ichimura, and Todd, 1997; Dehejia and Wahba, 1999; Blundell, Costa Dias, Meghir, and Van Reenen, 2001), as well as in government sponsored evaluations of social programs (e.g., Bloom, Michalopoulos, Hill, and Lei, 2002). Often, analyses are based on the assumption that assignment to treatment is unconfounded, that is, based on observable pretreatment variables only, and that there is sufficient overlap in the distributions of the pretreatment variables (Barnow, Cain and Goldberger, 1980; Heckman and Robb, 1984; Rubin 1977). Under these assumptions one can estimate the average effect within subpopulations defined by the pretreatment variables by differencing average treatment and control outcomes. The population average treatment effect can then be estimated by averaging these conditional average treatment effects over the appropriate distribution of the covariates. Methods implementing this in parametric forms have a long history. See for example Cochran and Rubin (1973), Barnow, Cain, and Goldberger (1980), Rosenbaum and Rubin (1985), Rosenbaum (1995), and Heckman and Robb (1984). 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, Ichimura, and Todd (1997) and Heckman, Ichimura, Smith, and Todd (1998) focus on the average effect on the treated and consider estimators based on local linear regression. Robins and Rotnitzky (1995) and Robins, Rotnitzky, and Zhao (1995), in the related setting of missing data problems, propose efficient estimators that combine weighting and regression adjustment. Hirano, Imbens, and Ridder (2000) 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 (1983a), leads to an efficient estimator. Ichimura and Linton (2001) consider higher order expansions of such estimators to analyze optimal bandwidth choices.

Alternatively, simple matching estimators are often used to estimate average treatment effects when assignment for treatment is believed to be unconfounded. These estimators match each treated unit to one or a small number of untreated units with similar values for the pretreatment variables. Then, the average effect of the treatment on the treated units is estimated by averaging within-match differences in the outcome variable between the treated and the untreated units (see, e.g., Rosenbaum, 1989, 1995; Gu and Rosenbaum, 1993; Rubin, 1973a,b; Dehejia and Wahba,

1999; Zhao, 2001; Becker and Ichino, 2002; Frölich, 2000). Typically, matching is done without replacement, so each control is used as a match only once and matches are independent. Matching estimators have great intuitive appeal, and are widely used in practice, as they do not require the researcher to set any smoothing parameters other than the number of matches. However, their formal large sample properties have not received much attention.

In this article, we propose a new framework to study simple matching estimators and establish a number of new results. In contrast with much of the previous literature, we allow each unit to be used as match more than once. Matching with replacement allows us to reduce biases, since it produces matches of higher quality than matching without replacement. This is important because we will show that matching estimators may have poor bias properties. In addition, matching with replacement enables us to consider estimators that match all units, treated as well as controls, so that the estimand is identical to the average treatment effect that is the focus of the Hahn (1998), Robins and Rotnitzky (1995), and Hirano, Imbens and Ridder (2000) studies.

Our results show that the large sample properties of simple matching estimators are not necessarily very attractive. First, we show that matching estimators include a conditional bias term which may not vanish at a rate faster than $N^{-1/2}$ when more than one continuous variable is used for matching. As a result, matching estimators may not be $N^{1/2}$-consistent. This crucial role for the dimension of the covariates also arises in nonparametric differencing methods for regression models (Honoré, 1992; Yatchew, 1999; Estes and Honoré, 2001). Second, even if the dimension of the covariates is low enough for the conditional bias term to vanish asymptotically, we show that the simple matching estimator with a fixed number of matches does not achieve the semiparametric efficiency bound as calculated by Hahn (1998). However, for the case when only one continuous covariate is used to match (as for matching on the propensity score), we show that the efficiency loss can be made arbitrarily close to zero by allowing a sufficiently large number of matches.

We also investigate estimators that combine matching with an additional bias correction based on a nonparametric extension of the regression adjustment proposed in Rubin (1973b) and Quade (1982). We show that the nonparametric bias correction removes the conditional bias asymptotically without affecting the variance, making matching estimators $N^{1/2}$-consistent. Compared to estimators based on regression adjustment without matching (e.g., Hahn, 1998; Heckman, Ichimura, and Todd, 1997; Heckman, Ichimura, Smith, and Todd, 1998) or estimators based on weighting by the inverse of the propensity score, (Hirano, Imbens, and Ridder, 2000)

the new estimators incorporate an additional layer of robustness, since the matching ensures consistency without accurate approximations to either the regression function or the propensity score.

Most of the evaluation literature has focused on estimation of the population average treatment effect. In some cases, however, it may be of interest to focus on the average treatment effect for the sample at hand. We show that matching estimators can also be interpreted as estimators of conditional average treatment effects for the sample, which can be estimated more precisely than the population average treatment effect. For this case, we propose an estimator of the variance of matching estimators that does not rely on consistent nonparametric estimation of unknown functions.

We apply the estimators to an example analyzed previously by Lalonde (1986), Heckman and Hotz (1989), Dehejia and Wahba (1999), Smith and Todd (2001) and Zhao (2002). For that example, we show that simple matching estimators are very sensitive to the choice for the number of matches, whereas a simple implementation of the bias correction considered in this article solves that problem. In a small simulation study based on a data generating process designed to mimic the data from Lalonde's application, we find that a simple implementation of the bias-corrected matching estimator performs well compared to both simple matching estimators and to regression estimators, in terms of bias and root-mean-squared-error.

In the next section we introduce the notation and define the estimators. In Section 3 we discuss the large sample properties of simple matching estimators. In Section 4 we analyze bias corrections. In Section 5 we propose a simple estimator for the conditional variance of matching estimators. In Section 6 we apply the estimators to Lalonde's example. In Section 7 we carry out a small simulation study to investigate the properties of the various estimators in a design modeled on the data from Section 6. Section 8 concludes. The appendix contains proofs.

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, for $i = 1, \ldots, N$, with all units exchangeable, let $(Y_i(0), Y_i(1))$ denote the two potential outcomes given the control treatment and given the active treatment respectively. The variable W_i, for $W_i \in \{0, 1\}$ 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 X_i. Estimands of interest are the population average treatment effect

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)],$$

and the average effect for the treated

$$\tau_t = \mathbb{E}[Y_i(1) - Y_i(0)|W_i = 1].$$

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

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

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

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

The dimension of X, denoted by k, will be seen to play an important role in the properties of the matching estimators. We assume that all covariates have continuous distributions.[1] The combination of the two conditions in Assumption 2 is referred to as strong ignorability (Rosenbaum and Rubin, 1983a). These conditions are strong, and in many cases may not be satisfied. In many studies, however, researchers have found it useful to consider estimators based on these or similar conditions. See, for example, Cochran (1968), Cochran and Rubin (1973), Rubin (1973a,b), Barnow, Cain, and Goldberger (1980), Heckman and Robb (1984), Rosenbaum and Rubin (1984), Ashenfelter and Card (1985), Lalonde (1986), Card and Sullivan (1988), Manski, Sandefur, McLanahan, and Powers (1992), Robins and Rotnitzky (1995), Robins, Rotnitzky, and Zhao (1995), Rosenbaum (1995), Heckman, Ichimura, and Todd (1997), Hahn (1998), Heckman,

[1] Discrete covariates can be easily dealt with by analyzing estimating average treatment effects within subsamples defined by their values. The number of discrete 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.

Ichimura, Smith, and Todd (1998), Angrist (1998), Lechner (1998), Dehejia and Wahba (1999), Becker and Ichino (2002), Blundell, Costa Dias, Meghir, and Van Reenen (2001), and Hotz, Imbens, and Mortimer (1999). If the first condition, unconfoundedness, is deemed implausible in a given application, methods allowing for selection on unobservables such as instrumental variable analyses (e.g., Heckman and Robb, 1984; Angrist, Imbens and Rubin, 1996; Abadie, 2002), sensitivity analyses (Rosenbaum and Rubin, 1983b), or bounds calculations (Manski, 1990, 1995) may be considered. See for general discussion of such issues the surveys in Heckman and Robb (1984), Angrist and Krueger (2000), Blundell and Costa Dias (2001), and Heckman, Lalonde, and Smith (2000). The importance of second part of the assumption, the restriction on the probability of assignment, has been discussed in Rubin (1977), Heckman, Ichimura, and Todd (1997), and Dehejia and Wahba (1999). Compactness and convexity of the support of the covariates are convenient regularity conditions.

Under Assumption 2 the average treatment effect for the subpopulation with pretreatment variables equal to $X = x$, $\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$, is identified from the distribution of (Y, W, X) because

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

To get the average effect of interest we average this conditional treatment effect over the marginal distribution of X:

$$\tau = \mathbb{E}[\tau(X)],$$

or over the conditional distribution to get the average effect for the treated:

$$\tau_t = \mathbb{E}[\tau(X)|W = 1].$$

Next we introduce some additional notation. For $x \in \mathbb{X}$ and $w \in \{0, 1\}$, let let $\mu_w(x) = \mathbb{E}[Y(w)|X = x]$ and $\sigma_w^2(x) = \mathbb{V}[Y(w)|X = x]$ be the conditional mean and variance respectively of $Y(w)$ given $X = x$, and let $\varepsilon_i = Y_i - \mu_{W_i}(X_i)$. By the unconfoundedness assumption

$$\mu_w(x) = \mathbb{E}[Y(w)|X = x] = \mathbb{E}[Y(w)|X = x, W = w] = \mathbb{E}[Y|X = x, W = w].$$

Similarly, $\sigma_w^2(x) = \mathbb{V}(Y|X = x, W = w)$. 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, 1983a). In our analysis, we adopt the following two assumptions.

ASSUMPTION 3: *(i) $\mu_w(x)$ and $\sigma_w^2(x)$ are continuous in x for all w, and (ii) the fourth moments of the conditional distribution of Y given $W = w$ and $X = x$ exist and are uniformly bounded.*

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

The numbers of control and treated units are $N_0 = \sum_i (1 - W_i)$ and $N_1 = \sum_i W_i$ respectively, with $N = N_0 + N_1$. Let $\|x\| = (x'x)^{1/2}$, for $x \in \mathbb{X}$ be the standard Euclidean vector norm.[2] Let $j_m(i)$ be the index j that solves

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

where $1\{\cdot\}$ is the indicator function, equal to one if the expression in brackets is true and zero otherwise. In other words, $j_m(i)$ is the index of the unit that is the m-th closest to unit i in terms of the distance measure based on the norm $\|\cdot\|$, among the units with the treatment opposite to that of unit i.[3] In particular, $j_1(i)$, sometimes for notational convenience denoted by $j(i)$, is the nearest match for unit i. For notational simplicity and since we only consider continuous covariates, we ignore the possibility of ties, which only happen with probability zero. Let $\mathcal{J}_M(i)$ denote the set of indices for the first M matches for unit i:

$$\mathcal{J}_M(i) = \{j_1(i), \ldots, j_M(i)\}.$$

Define the catchment area $A_M(i)$ as the subset of $\mathbb{X}$ such that each observation, j, with $W_j = 1 - W_i$ and $X_j \in A_M(i)$ is matched to i:

$$A_M(i) = \Big\{x \;\Big|\; \textstyle\sum_{j\,|\,W_j=W_i} 1\{\|X_j - x\| \leq \|X_i - x\|\} \leq M\Big\}.$$

Finally, let $K_M(i)$ denote the number of times unit i is used as a match given that M matches per unit are done:

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

In many matching methods (e.g., Rosenbaum, 1995), the matching is carried out without replacement, so that every unit is used as a match at most once, and $K_M(i) \leq 1$. However, when both treated and control units matched it is imperative that units can be used as matches more than once. We show below that the distribution of $K_M(i)$ is an important determinant of the variance of the estimators.

[2] 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, since $\|x\|_V = ((Px)'(Px))^{1/2}$ for P such that $P'P = V$.

[3] For this definition to make sense, we assume that $N_0 \geq m$ and $N_1 \geq m$.

2.2. Estimators

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. All estimators we consider impute the unobserved potential outcomes in some way. The first estimator, the simple matching estimator, uses the following estimates for the potential outcomes:

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

and

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

The simple matching estimator we shall study is

$$\hat{\tau}_M^{sm} = \frac{1}{N} \sum_{i=1}^{N} \Big(\hat{Y}_i(1) - \hat{Y}_i(0) \Big). \tag{1}$$

Consider the case with a single match ($M = 1$). The differences $\hat{Y}_i(1) - \hat{Y}_i(0)$ and $\hat{Y}_j(1) - \hat{Y}_j(0)$ are not necessarily independent, and in fact they will be identical if i is matched to l (that is, $j(i) = l$) and l is matched to i (that is, $j(l) = i$). This procedure differs from standard pairwise matching procedures where one constructs a number of distinct pairs, without replacement. Matching with replacement leads to a higher variance, but produces higher match quality, and thus typically a lower bias.

The computational ease of the simple matching estimator is illustrated in Table 1 for an example with four units. In this example unit 1 is matched to unit 3, units 2 and 3 are both matched to unit 1, and unit 4 is matched to unit 2. Hence unit 1 is used as a match twice, units 2 and 3 are used as a match once, and unit 4 is never used as a match. The estimated average treatment effect is $\sum_{i=1}^{4} \hat{\tau}_i/4 = (2 + 5 + 2 + 0)/4 = 9/4$.

The simple matching estimator can easily be modified to estimate the average treatment effect for the treated:

$$\tau_M^{smt} = \frac{\sum_{i=1}^{N} W_i(\hat{Y}_i(1) - \hat{Y}_i(0))}{\sum_{i=1}^{N} W_i} = \frac{1}{N_1} \sum_{W_i=1} \Big(Y_i - \hat{Y}_i(0) \Big), \tag{2}$$

because if $W_i = 1$, then $\hat{Y}_i(1) = Y_i$.

We shall compare the matching estimators to covariance-adjustment or regression imputation estimators. Let $\hat{\mu}_w(X_i)$ be a consistent estimator of $\mu_w(X_i)$. Let

$$\bar{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ \hat{\mu}_0(X_i) & \text{if } W_i = 1, \end{cases} \tag{3}$$

and

$$\bar{Y}_i(1) = \begin{cases} \hat{\mu}_1(X_i) & \text{if } W_i = 0, \\ Y_i & \text{if } W_i = 1. \end{cases} \tag{4}$$

The regression imputation estimator is defined by

$$\hat{\tau}^{reg} = \frac{1}{N} \sum_{i=1}^{N} \left(\bar{Y}_i(1) - \bar{Y}_i(0) \right). \tag{5}$$

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. However, the regression imputation and matching estimators differ in the way they change with the number of observations. We classify as matching estimators those estimators which use a finite and fixed number of matches. We classify as regression imputation estimators those for which $\hat{\mu}_w(x)$ is a consistent estimator for $\mu_w(x)$. The estimators considered by Hahn (1998) and some of the those considered by Heckman, Ichimura, and Todd (1997) and Heckman, Ichimura, Smith, and Todd (1998) are regression imputation estimators. Hahn shows that if nonparametric series estimation is used for $\mathbb{E}[YW|X]$, $\mathbb{E}[Y(1-W)|X]$, and $\mathbb{E}[W|X]$, and those are used to estimate $\mu_1(x)$ as $\hat{\mu}_1(x) = \hat{\mathbb{E}}[YW|X = x]/\hat{\mathbb{E}}[W|X = x]$ and $\mu_0(x)$ as $\hat{\mu}_0(x) = \hat{\mathbb{E}}[Y(1-W)|X = x]/\hat{\mathbb{E}}[1-W|X = x]$, then the regression imputation estimator is asymptotically efficient for τ.

In addition we consider a bias-corrected matching estimator where the difference within the matches is adjusted for the difference in covariate values:

$$\tilde{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ \dfrac{1}{M} \displaystyle\sum_{j \in \mathcal{J}_M(i)} \left(Y_j + \hat{\mu}_0(X_i) - \hat{\mu}_0(X_j) \right) & \text{if } W_i = 1, \end{cases} \tag{6}$$

and

$$\tilde{Y}_i(1) = \begin{cases} \dfrac{1}{M} \displaystyle\sum_{j \in \mathcal{J}_M(i)} \left(Y_j + \hat{\mu}_1(X_i) - \hat{\mu}_1(X_j) \right) & \text{if } W_i = 0, \\ Y_i & \text{if } W_i = 1, \end{cases} \tag{7}$$

with corresponding estimator

$$\hat{\tau}_M^{bcm} = \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{Y}_i(1) - \tilde{Y}_i(0)\right). \tag{8}$$

Rubin (1979) and Quade (1982) discusses such estimators in the context of matching without replacement and with linear covariance adjustment.

To set the stage for some of the discussion below, consider the bias of the simple matching estimator relative to the average effect in the sample. Conditional on $\{X_i, W_i\}_{i=1}^N$, the bias is, under Assumption 2:

$$E\left[\frac{1}{N}\sum_{i=1}^{N}\left(\left(\hat{Y}_i(1) - \hat{Y}_i(0)\right) - \left(Y_i(1) - Y_i(0)\right)\right)\middle|\, \{X_i, W_i\}_{i=1}^N\right]$$
$$= \frac{1}{N}\sum_{i=1}^{N}(2W_i - 1)\left\{\frac{1}{M}\sum_{m=1}^{M}\left(\mu_{W_{j_m(i)}}(X_i) - \mu_{W_{j_m(i)}}(X_{j_m(i)})\right)\right\}. \tag{9}$$

That is, the conditional bias consists of terms of the form $\mu_w(X_i) - \mu_w(X_{j_m(i)})$. These terms are small when $X_i \simeq X_{j_m(i)}$, as long as the regression functions are continuous. Similarly, the bias of the regression imputation estimator consists of terms of the form $\mu_w(X_i) - \mathbb{E}[\hat{\mu}_w(X_i)]$, which are small when $\mathbb{E}[\hat{\mu}_w(X_i)] \simeq \mu_w(X_i)$. On the other hand, the bias of the bias-corrected estimator consists of terms of the form $\mu_w(X_i) - \mu_w(X_{j_m(i)}) - \mathbb{E}[\hat{\mu}_w(X_i) - \hat{\mu}_w(X_{j_m(i)})]$, which are small if either $X_i \simeq X_{j_m(i)}$ or $\mathbb{E}[\hat{\mu}_w(X_i)] \simeq \mu_w(X_i)$. The bias-adjusted matching estimator combines some of the bias reductions from the matching, by comparing units with similar values of the covariates, and the bias-reduction from the regression. Compared to only regression imputation, the bias-corrected matching estimator relies less on the accuracy of the estimator of the regression function since it only needs to adjust for relatively small differences in the covariates.

We are interested in the properties of the simple and bias-corrected matching estimators in large samples, that is, as N increases, for fixed M.[4] The properties of interest include bias and variance. Of particular interest is the dependence of these results on the dimension of the covariates. Some of these properties will be considered conditional on the covariates. In particular, we will propose an estimator for the conditional variance of matching estimators given

[4]Of course, M could be specified as a function of the number observations. This would entail, however, the selection of a smoothing parameter as a function of the number of observations; something that simple matching methods allows one to avoid. The purpose of this article is to study the properties of a simple matching procedure which does not require the selection of smoothing parameters as functions of the sample size. In addition, we will show that simple matching estimators, with fixed M, may incorporate large biases created by poor match quality. Letting M increase with the sample size may only exacerbate this problem, since matches of lower quality would be made.

$X_1, \ldots, X_N, W_1, \ldots, W_N$, viewed as estimators of the sample average conditional treatment effect $\overline{\tau(X)} = \sum \tau(X_i)/N$, or its version for the treated $\overline{\tau(X)}_t = \sum W_i \cdot \tau(X_i)/N_1$. There are two reasons for focusing on the conditional distribution. First, in many cases one is interested in the average effect for the sample at hand, rather than for the hypothetical population this sample is drawn from, especially given that the former can typically be estimated more precisely. The second reason is that there exists an estimator for the conditional variance that, in the spirit of the matching estimator, does not rely on additional choices for smoothing parameters. The difference between the marginal variance and the conditional variance is the variance of $\tau(X)$, $V^{\tau(X)} = \mathbb{E}[(\tau(X) - \tau)^2]$, divided by the sample size. This variance represents the difference between the sample distribution of the covariates and the population. Therefore, estimating the unconditional variance requires estimating the variance of $\tau(X)$, which, in turn, as in Hirano, Imbens, and Ridder (2001), requires choices regarding the smoothing parameters in nonparametric estimation of the conditional means and variances.

3. Simple Matching Estimators

In this section we investigate the properties of the simple matching estimator $\hat{\tau}_M^{sm}$ defined in (1). Let $\mathbf{X}$ and $\mathbf{W}$ be the matrices with i-th row equal to X_i', and W_i, respectively. Define the two $N \times N$ matrices $\mathbf{A}_1(\mathbf{X}, \mathbf{W})$ and $\mathbf{A}_0(\mathbf{X}, \mathbf{W})$, with typical element

$$\mathbf{A}_{1,ij} = \begin{cases} 1 & \text{if } i = j,\ W_i = 1 \\ 1/M & \text{if } j \in \mathcal{J}_M(i),\ W_i = 0, \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

and

$$\mathbf{A}_{0,ij} = \begin{cases} 1 & \text{if } i = j,\ W_i = 0 \\ 1/M & \text{if } j \in \mathcal{J}_M(i),\ W_i = 1, \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

and define $\mathbf{A} = \mathbf{A}_1 - \mathbf{A}_0$. For the example in Table 1,

$$\mathbf{A}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{A}_0 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{A} = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}.$$

Notice that for any $N \times 1$ vector $\mathbf{V} = (V_1, \ldots, V_N)'$:

$$\iota_N' \mathbf{A} \mathbf{V} = \sum_{i=1}^{N} (2W_i - 1) \left(1 + \frac{K_M(i)}{M}\right) V_i = \sum_{i=1}^{N} (2W_i - 1) \frac{1}{M} \sum_{m=1}^{M} (V_i - V_{j_m(i)}), \tag{12}$$

where ι_N be the N-dimensional vector with all elements equal to one.

Let $\mathbf{Y}$, $\mathbf{Y}(0)$, $\mathbf{Y}(1)$, $\hat{\mathbf{Y}}(0)$, and $\hat{\mathbf{Y}}(1)$ be the matrices with i-th row equal to Y_i, $Y_i(0)$, $Y_i(1)$, $\hat{Y}_i(0)$, and $\hat{Y}_i(1)$, respectively. Furthermore, let $\boldsymbol{\mu}(\mathbf{X},\mathbf{W})$ and $\boldsymbol{\varepsilon}$ be the $N\times 1$ vectors with i-th element equal to $\mu_{W_i}(X_i)$ and ε_i, respectively, and let $\boldsymbol{\mu}_0(\mathbf{X})$ and $\boldsymbol{\mu}_1(\mathbf{X})$ are the $N\times 1$ vectors with i-th element equal to $\mu_0(X_i)$ and $\mu_1(X_i)$, respectively. Then

$$\hat{\mathbf{Y}}(1)=\mathbf{A}_1\mathbf{Y},\quad \text{and}\ \ \hat{\mathbf{Y}}(0)=\mathbf{A}_0\mathbf{Y}.$$

We can now write the estimator $\hat{\tau}_M^{sm}$ as

$$\hat{\tau}_M^{sm}=\iota_N'\Big(\hat{\mathbf{Y}}(1)-\hat{\mathbf{Y}}(0)\Big)\Big/N=\Big(\iota_N'\mathbf{A}_1\mathbf{Y}-\iota_N'\mathbf{A}_0\mathbf{Y}\Big)\Big/N=\iota_N'\mathbf{A}\mathbf{Y}/N$$

$$=\iota_N'\mathbf{A}\boldsymbol{\mu}(\mathbf{X},\mathbf{W})/N+\iota_N'\mathbf{A}\boldsymbol{\varepsilon}/N.$$

Using equation (12), we can also write this as:

$$\hat{\tau}_M^{sm}=\iota_N'\mathbf{A}\mathbf{Y}/N=\frac{1}{N}\sum_{i=1}^{N}(2W_i-1)\left(1+\frac{K_M(i)}{M}\right)Y_i. \tag{13}$$

Finally, using the fact that $\mathbf{A}_1\boldsymbol{\mu}(\mathbf{X},\mathbf{W})=\mathbf{A}_1\boldsymbol{\mu}_1(\mathbf{X})$ and $\mathbf{A}_0\boldsymbol{\mu}(\mathbf{X},\mathbf{W})=\mathbf{A}_0\boldsymbol{\mu}_0(\mathbf{X})$ we can write

$$\hat{\tau}_M^{sm}-\tau=\Big(\overline{\tau(X)}-\tau\Big)+E^{sm}+B^{sm}, \tag{14}$$

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

$$\overline{\tau(X)}=\iota_N'\Big(\boldsymbol{\mu}_1(\mathbf{X})-\boldsymbol{\mu}_0(\mathbf{X})\Big)/N, \tag{15}$$

E^{sm} is the contribution of the residuals:

$$E^{sm}=\iota_N'\mathbf{A}\boldsymbol{\varepsilon}/N=\frac{1}{N}\sum_{i=1}^{N}E_i^{sm},$$

where $E_i^{sm}=(2W_i-1)\cdot(1+K_M(i)/M)\cdot\varepsilon_i$, and B^{sm} is the bias relative to the average treatment effect for the sample, conditional on $\mathbf{X}$ and $\mathbf{W}$:

$$B^{sm}=\iota_N'(\mathbf{A}_1-I_N)\boldsymbol{\mu}_1(\mathbf{X})/N-\iota_N'(\mathbf{A}_0-I_N)\boldsymbol{\mu}_0(\mathbf{X})/N=\frac{1}{N}\sum_{i=1}^{N}B_i^{sm}, \tag{16}$$

where $B_i^{sm}=(2W_i-1)(1/M)\sum_{j\in\mathcal{J}_M(i)}(\mu_{1-W_i}(X_i)-\mu_{1-W_i}(X_j))$. We will refer to B^{sm} as the bias term, or the conditional bias, and to $\text{Bias}^{sm}=\mathbb{E}[B^{sm}]$ as the (unconditional) bias. If the

matching is exact, and $X_i = X_{j_m(i)}$ for all i, then the bias term is equal to zero. In general it is not and its properties will be analyzed in Section 3.1. The first two terms on the right-hand side of (14) are important for the large sample variance of the estimator. The first term depends only on the covariates $\mathbf{X}$ and has variance equal to the variance of the treatment effect, $V^{\tau(X)}/N = \mathbb{E}[(\tau(X)-\tau)^2]/N$. The variance of the second term is the conditional variance of the estimator. We will analyze these two terms in Section 3.2.

Similarly we can write the estimator for the average effect for the treated, (2), as

$$\hat{\tau}_M^{sm,t} - \tau_t = \left(\overline{\tau(X)}_t - \tau_t\right) + E^{sm,t} + B^{sm,t}, \tag{17}$$

where $\overline{\tau(X)}_t$ is the average conditional treatment effect over sample of treated:

$$\overline{\tau(X)}_t = \frac{1}{N_1} \sum_{i:W_i=1} \left(\mu_1(X_i) - \mu_0(X_i)\right), \tag{18}$$

$E^{sm,t}$ is the contribution of the residuals:

$$E^{sm,t} = \frac{1}{N_1} \sum_{i=1}^{N} E_i^{sm,t},$$

where $E_i^{sm,t} = (W_i - (1-W_i)\cdot K_M(i)/M)\cdot \varepsilon_i$, and $B^{sm,t}$ is the bias term:

$$B^{sm,t} = -\iota_N'(\mathbf{A}_0 - I_N)\boldsymbol{\mu}_0(\mathbf{X})/N = \frac{1}{N_1}\sum_{i=1}^{N} B_i^{sm,t}, \tag{19}$$

where $B_i^{sm,t} = (1-W_i)(1/M)\sum_m(\mu_0(X_i) - \mu_0(X_{j_m(i)}))$.

3.1. Bias

The conditional bias in equation (16) consists 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:

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

In order to study the components of the bias it is therefore useful to analyze the distribution of the matching discrepancy $X_{j_m(i)} - X_i$.

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 $X_1, ..., X_N$ from some distribution over the support

$\mathbb{X}$ (with density $f(x)$ and distribution function $F(x)$). Now, consider the closest match to z in the sample. Let

$$j_1 = \text{argmin}_{j=1,\dots,N} \|X_j - z\|,$$

and let $U_1 = X_{j_1} - z$ be the matching discrepancy. We are interested in the distribution of the difference U_1, which is a $k \times 1$ vector. More generally, we are interested in the distribution of the m-th closest match discrepancy, $U_m = X_{j_m} - z$, where X_{j_m} is the m-th closest match to z from the random sample of size N. The following lemma describes some key asymptotic properties of the matching discrepancy.

Lemma 1: (Matching Discrepancy – Asymptotic Properties)
Suppose that $f(z) > 0$ and that f is differentiable in a neighborhood of z. Then, $N^{1/k} \cdot U_m \xrightarrow{d} V_m$, where

$$f_{V_m}(v) = \frac{f(z)}{(m-1)!} \left(\|v\|^k \frac{f(z)}{k} \frac{2\pi^{k/2}}{\Gamma(k/2)} \right)^{m-1} \exp\left(-\|v\|^k \frac{f(z)}{k} \frac{2\pi^{k/2}}{\Gamma(k/2)} \right),$$

and $\Gamma(y) = \int_0^\infty e^{-t} t^{y-1} dt$ (for $y > 0$) is Euler's Gamma Function. Moreover, the first three moments of U_m are:

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

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

and

$$E[\|U_m\|^3] = O\left(N^{-3/k}\right),$$

where I_k is the identity matrix of size k.

(All proofs are given in the appendix.)
The lemma shows that the order of the matching discrepancy increases with the number of continuous covariates. Intuitively, as the number of covariates increases, it becomes more difficult to find close matches. 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.

Lemma 2: (Matching Discrepancy – Uniformly Bounded Moments)
If Assumption 1 holds, then all the moments of $N^{1/k} \cdot U_m$ are uniformly bounded in N and $z \in \mathbb{X}$.

These results allow us to calculate the bias and stochastic order of the bias term.

Theorem 1: (Bias for the Average Treatment Effect)
Under assumptions 1, 2 and 4, and if $\mu_0(x)$ and $\mu_1(x)$ are three times continuously differentiable with bounded third derivatives, and $f_0(x)$ and $f_1(x)$ are differentiable, then
(i) $B^{sm} = O_p(N^{-1/k})$, and
(ii) the bias of the simple matching estimator is

$$\begin{aligned}\text{Bias}^{sm} = \mathbb{E}[B^{sm}] = \left(\frac{1}{M}\sum_{m=1}^{M}\Gamma\left(\frac{mk+2}{k}\right)\frac{1}{(m-1)!k}\right)\frac{1}{N^{2/k}} \times \\ \left\{\frac{(1-p)}{p^{2/k}}\int\left(f_1(x)\frac{\pi^{k/2}}{\Gamma\left(1+k/2\right)}\right)^{-2/k}\left\{\frac{1}{f_1(x)}\frac{\partial f_1}{\partial x'}(x)\frac{\partial \mu_1}{\partial x}(x)+\frac{1}{2}\,\text{tr}\left(\frac{\partial^2\mu_1}{\partial x'\partial x}(x)\right)\right\}f_0(x)dx\right. \\ \left.-\frac{p}{(1-p)^{2/k}}\int\left(f_0(x)\frac{\pi^{k/2}}{\Gamma\left(1+k/2\right)}\right)^{-2/k}\left\{\frac{1}{f_0(x)}\frac{\partial f_0}{\partial x'}(x)\frac{\partial \mu_0}{\partial x}(x)+\frac{1}{2}\,\text{tr}\left(\frac{\partial^2\mu_0}{\partial x'\partial x}(x)\right)\right\}f_1(x)dx\right\} \\ +\ o\left(\frac{1}{N^{2/k}}\right).\end{aligned}$$

Consider the implications of this theorem for the asymptotic properties of the simple matching estimator. First note that $\sqrt{N}(\overline{\tau(X)} - \tau) = O_p(1)$ with a normal limiting distribution, by a standard central limit theorem. Also, as will be shown later, $\sqrt{N}E^{sm} = O_p(1)$, with again a normal limiting distribution. Now, suppose the covariate is scalar ($k = 1$). In that case $B^{sm} = O_p(N^{-1})$. Hence the asymptotic properties of the simple matching estimator will be dominated by those of $\overline{\tau(X)} - \tau$ and E^{sm}, and $\sqrt{N}(\hat{\tau}^{sm} - \tau)$ will be asymptotically normal.

Next, consider the case with $k = 2$. In that case $B^{sm} = O_p(N^{-1/2})$, and the asymptotic properties will be determined by all three terms. Note that there is no asymptotic bias as $\text{Bias}^{sm} = O(N^{-1})$. However, it is unclear whether the estimator in this case is normally distributed as we have no asymptotic distribution theory for $\sqrt{N} \cdot B^{sm}$ for this case.

Next, consider the case with $k \geq 3$. Now the order of B^{sm} is $O_p(N^{-1/k})$, so that the normalization factor for $\hat{\tau}^{sm} - \tau$ is $N^{1/k}$. In this case the asymptotic distribution is dominated by the bias term. Note that the asymptotic bias itself is still zero as $\text{Bias}^{sm} = O(N^{-2/k})$. Note also that experimental data does not reduce the order of Bias^{sm}. If the data comes from a randomized

experiment, then $f_0(x)$ and $f_1(x)$ coincide. However, this is not enough in general to reduce the order of the bias if a matching procedure is adopted.

The bias for the average treatment effect for the treated follows directly from the earlier result:

COROLLARY 1: (BIAS FOR THE AVERAGE TREATMENT EFFECT ON THE TREATED)
Under assumptions 1, 2 and 4, if $\mu_0(x)$ has bounded third derivatives, and $f_0(x)$ is differentiable, then
(i) $B^{sm,t} = O_p(N_0^{-1/k})$, and
(ii)

$$\text{Bias}^{sm,t} = \mathbb{E}[B^{sm,t}] = -\left(\frac{1}{M}\sum_{m=1}^{M}\Gamma\left(\frac{mk+2}{k}\right)\frac{1}{(m-1)!k}\right)\frac{1}{N_0^{2/k}}\times$$
$$\frac{p}{(1-p)^{2/k}}\int\left(f_0(x)\frac{\pi^{k/2}}{\Gamma\left(1+k/2\right)}\right)^{-2/k}\left\{\frac{1}{f_0(x)}\frac{\partial f_0}{\partial x'}(x)\frac{\partial \mu_0}{\partial x}(x)+\frac{1}{2}\,\text{tr}\left(\frac{\partial^2\mu_0}{\partial x'\partial x}(x)\right)\right\}f_1(x)dx$$
$$+\ o\left(\frac{1}{N_0^{2/k}}\right),$$

This case is particularly relevant since often matching estimators have been used to estimate the average effect for the treated. Generally in those cases the bias is ignored. This is justified if there is only a single continuous covariate. It is also justified using an asymptotic arguments if the number of controls is very large relative to the number of treated. Suppose that the two sample sizes go to infinity at different rates, $N_1 = O(N_0^s)$. Then $B^{sm,t} = O_p(N_0^{-1/k}) = O_p(N_1^{1/(sk)})$. Hence if $s < 2/k$, it follows that $B^{sm,t} = o_p(N_1^{-1/2})$, and the bias term will get dominated in the large sample distribution by the two other terms, $\overline{\tau(X)}_t - \tau$ and $E^{sm,t}$, which are $O_p(N_1^{-1/2})$.

3.2. CONDITIONAL VARIANCE

In this section we investigate the conditional variance of the simple matching estimator $\hat{\tau}_M^{sm}$. Consider the representation of the estimator in (14). Only the second term contributes to the conditional variance. Conditional on $\mathbf{X}$ and $\mathbf{W}$, the variance of $\hat{\tau}$ is

$$\mathbb{V}(\hat{\tau}_M^{sm}|\mathbf{X},\mathbf{W}) = \mathbb{V}(E^{sm}|\mathbf{X},\mathbf{W}) = \mathbb{V}(\iota_N'\mathbf{A}\boldsymbol{\varepsilon}/N|\mathbf{X},\mathbf{W}) = \frac{1}{N^2}\iota_N'\mathbf{A}\Omega\mathbf{A}'\iota_N, \tag{20}$$

where

$$\Omega = E\Big[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X},\mathbf{W}\Big],$$

is a diagonal matrix with the ith diagonal element equal to $\sigma^2_{W_i}(X_i)$, the conditional variance of Y_i given X_i and W_i. Note that (20) gives the exact variance, not relying on any large sample approximations. Using the representation of the simple matching estimator in equation (13), we can write this as:

$$\mathbb{V}(\hat{\tau}^{sm}_M|\mathbf{X},\mathbf{W}) = \frac{1}{N^2}\sum_{i=1}^{N}\left(1+\frac{K_M(i)}{M}\right)^2\sigma^2_{W_i}(X_i). \tag{21}$$

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

Lemma 3: *Suppose Assumptions 1 to 4 hold. Then*
(i) $K_M(i) = O_p(1)$*, and its moments are bounded uniformly in* N*, and*
(ii)

$$V^E = \lim_{N\to\infty}\mathbb{E}\left[\left(1+\frac{K_M(i)}{M}\right)^2\sigma^2_{W_i}(X_i)\right],$$

is finite.

3.3. Consistency and Asymptotic Normality

In this section we show that the simple matching estimator is consistent for the average treatment effect and that, without the bias term, is $N^{1/2}$-consistent and asymptotically normal.

Theorem 2: (Consistency of the Simple Matching Estimator)
Suppose Assumptions 1, 2, and 4 hold. If in addition $\mu_1(x)$ *and* $\mu_0(x)$ *are continuous, then*

$$\hat{\tau}^{sm} - \tau \xrightarrow{p} 0.$$

Note that the consistency result does not require restrictions on the dimension of the covariates. The conditions are largely smoothness of the regression functions, which implies that $\mu_w(X_i) - \mu_w(X_{j_m(i)})$ converges to zero. This convergence is uniform by the restrictions on the two conditional densities $f_w(x)$, which in turn follows from the fact that the propensity score is bounded away from zero and one, and from the compact support of the covariates.

Next, we state the formal result for asymptotic normality. The first result gives an asymptotic normality result for the estimator $\hat{\tau}^{sm}$ after subtracting the bias B^{sm}.

Theorem 3: (Asymptotic Normality for the Simple Matching Estimator)
Suppose Assumptions 1 to 4 hold, and that $\mu_1(x)$ *and* $\mu_0(x)$ *have bounded third derivatives. Then*

$$\sqrt{N}(\hat{\tau}^{sm} - B^{sm} - \tau) \xrightarrow{d} \mathcal{N}\left(0, V^E + V^{\tau(X)}\right).$$

In the scalar covariate case there is no need to remove the bias:

COROLLARY 2: (ASYMPTOTIC NORMALITY FOR SIMPLE MATCHING ESTIMATOR WITH SCALAR COVARIATE)
Suppose Assumptions 1 to 4 hold, and that $\mu_1(x)$ and $\mu_0(x)$ have bounded third derivatives. Suppose in addition that the covariate is a scalar ($k = 1$). Then

$$\sqrt{N}(\hat{\tau}^{sm} - \tau) \xrightarrow{d} \mathcal{N}\left(0, V^E + V^{\tau(X)}\right).$$

If we focus on $\hat{\tau}^{sm}$ as an estimator for the conditional average treatment effect $\overline{\tau(X)}$, we obtain the following result:

COROLLARY 3: (ASYMPTOTIC NORMALITY FOR THE SIMPLE MATCHING ESTIMATOR AS AN ESTIMATOR OF $\overline{\tau(X)}$)
Suppose Assumptions 1 to 4, and that $\mu_1(x)$ and $\mu_0(x)$ have bounded third derivatives. Then

$$\sqrt{N}(\hat{\tau}^{sm} - B^{sm} - \overline{\tau(X)}) \xrightarrow{d} \mathcal{N}\left(0, V^E\right).$$

3.4. EFFICIENCY

To compare the efficiency of the estimator considered here to previously proposed estimators and in particularly to the efficiency bound calculated by Hahn (1998), it is useful to go beyond the conditional variance and compute the unconditional variance. In general the key to the efficiency properties of the matching estimators is the distribution of $K_M(i)$, the number of times each unit is used as a match. It is difficult to work out the limiting distribution of this variable for the general case.[5] Here we investigate the form of the variance for the special case with a scalar covariate and a general M.

THEOREM 4: *Suppose $k = 1$. If Assumptions 1 to 4 hold, then*

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

[5]The key is the second moment of the volume of the "catchment area" $A_M(i)$, defined as the subset of $\mathbb{X}$ such that each observation, j, with $W_j = 1 - W_i$ and $X_j \in A_M(i)$ is matched to i. In the single match case with $M = 1$ these objects are studied in stochastic geometry where they are known as Poisson-Voronoi tesselations (Moller, 1994; Okabe, Boots, Sugihara and Nok Chiu, 2000; Stoyan, Kendall, and Mecke, 1995). The variance of the volume of such objects under uniform $f_0(x)$ and $f_1(x)$, normalized by the mean, has been worked out numerically for the one, two, and three dimensional cases.

Since the semiparametric efficiency bound for this problem is, as established by Hahn (1998),

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

the matching estimator is not efficient in general. However, the efficiency loss can be bounded in percentage terms in this case:

$$\frac{N \cdot \mathbb{V}(\hat{\tau}_M^{sm}) - \mathbb{V}^{eff}}{\mathbb{V}^{eff}} \leq \frac{1}{2M}.$$

The efficiency loss quickly disappears 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. Bias Corrected Matching

In this section we analyze the properties of the bias corrected matching estimator $\hat{\tau}_M^{bcm}$, defined in equation (8). The bias correction presented in equation (8) requires the estimation of the regression functions $\mu_0(x)$ and $\mu_1(x)$. In order to establish the asymptotic behavior of the bias-corrected estimator, in this section, we consider a nonparametric series estimator for the two regression functions with $K(N)$ terms in the series, where $K(N)$ increases with N. This type of nonparametric estimation relies however on selecting smoothing parameters as functions of the sample size, something that matching estimator allows to avoid. For this reason, in sections 6 and 7 we consider a simple implementation of the bias correction which uses linear regression to estimate $\mu_0(x)$ and $\mu_1(x)$.

Let $\lambda = (\lambda_1, ..., \lambda_k)$ be a multi-index of dimension k, that is, a k-dimensional vector of non-negative integers, with $|\lambda| = \sum_{i=1}^k \lambda_i$, and let $x^\lambda = x_1^{\lambda_1} \dots x_k^{\lambda_k}$. Consider a series $\{\lambda(r)\}_{r=1}^\infty$ containing all distinct such vectors and such that $|\lambda(r)|$ is nondecreasing. Let $p_r(x) = x^{\lambda(r)}$, where $p^K(x) = (p_1(x), ..., p_K(x))'$. Following Newey (1995), the nonparametric series estimator of the regression function $\mu_w(x)$ is given by:

$$\hat{\mu}_w(x) = p^{K(N)}(x)' \left(\sum_{i:W_i=w} p^{K(N)}(X_i) p^{K(N)}(X_i)' \right)^- \sum_{i:W_i=w} p^{K(N)}(X_i) Y_i,$$

where $(\cdot)^-$ denotes a generalized inverse. Given the estimated regression function, let $\hat{B}^{sm}$ be the

estimated bias term:

$$\hat{B}^{sm} = \frac{1}{N}\sum_{i=1}^{N}\left\{ W_i \cdot \left(\frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} (\hat{\mu}_0(X_i) - \hat{\mu}_0(X_j)) \right) \right.$$

$$\left. -(1 - W_i) \cdot \left(\frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} (\hat{\mu}_1(X_i) - \hat{\mu}_1(X_j)) \right) \right\},$$

so that $\hat{\tau}^{bcm} = \hat{\tau}^{sm} - \hat{B}^{sm}$.

The following theorem shows that the bias correction removes the bias without affecting the asymptotic variance.

Theorem 5: (Bias Corrected Matching Estimator)
Suppose that Assumptions 1 to 4 hold. Assume also:
(i) The support of X, $\mathbb{X} \subset \mathbb{R}^k$, is a Cartesian product of compact intervals,
(ii) $K(N) = N^\nu$, with $0 < \nu < 2/(3k + 4k^2)$,
(iii) There is a C such that for each multi-index λ the λ-th partial derivative of $\mu_w(x)$ exists for $w = 0, 1$ and is bounded by $C^{|\lambda|}$. Then,

$$\sqrt{N}\left(B^{sm} - \hat{B}^{sm}\right) \xrightarrow{d} 0,$$

and

$$\sqrt{N}(\hat{\tau}^{bcm} - \tau) \xrightarrow{d} \mathcal{N}\left(0, V^E + V^{\tau(X)}\right).$$

Thus, the bias corrected matching estimator has the same normalized variance as the simple matching estimator.

5. Estimating the Conditional Variance

Estimating the conditional variance $V^E = \mathbb{E}[\iota_N' \mathbf{A}\Omega\mathbf{A}'\iota_N/N]$ is complicated by the fact that it involves the conditional outcome variances $\sigma_w^2(x)$. In principle, one can estimate the conditional variances $\sigma_w^2(x)$ consistently, first using nonparametric regression to obtain $\mu_w(x)$, and then using nonparametric regression again to obtain $\sigma_w^2(x)$. Although this leads to a consistent estimator for the conditional variance, it would require exactly the type of nonparametric regression that the simple matching estimator allows one to avoid. For this reason, we propose a new estimator of the conditional variance of the simple matching estimator which does not require consistent nonparametric estimation of $\sigma_w^2(x)$.

The conditional variance of the average treatment effect estimator depends on the unit-level variances $\sigma^2_w(x)$ only through an average. To estimate these unit-level variances we use a matching approach. Our method can be interpreted as a nonparametric estimator for $\sigma^2_w(x)$ with a fixed bandwidth, where instead of the original matching of treated to control units, we now match treated units with treated units and control units with control units. This leads to an approximately unbiased estimate of $\sigma^2_w(x)$, although not a consistent one. However, the average of these inconsistent variance estimators is consistent for the average of the variances. Suppose we have two pairs i and j with the same covariates, $X_i = X_j = x$ and the same treatment $W_i = w$, and consider the squared difference between the within-pair differences:

$$E\Big[\Big(Y_i - Y_j\Big)^2\Big|X_i = X_j = x, W_i = w\Big] = 2 \cdot \sigma^2_w(x).$$

In that case we can estimate the variance $\sigma^2_{W_i}(X_i)$ as $\hat{\sigma}^2_{W_i}(X_i) = (Y_i - Y_j)^2/2$. This estimator is unbiased, but it is not consistent as its variance does not go to zero with the sample size. However, this is not necessary for the estimator for the normalized variance of $\hat{\tau}^{sm}_M$ to be consistent.

In practice, it may not be possible to find different pairs with the same value of the covariates. Hence let us consider the nearest pair to pair i by solving

$$l(i) = \text{argmin}_{l:l\neq i, w_i = w_l}||X_i - X_l||,$$

and let

$$\widehat{\sigma}^2_{W_i}(X_i) = \frac{1}{2}\Big(Y_i - Y_{l(i)}\Big)^2, \tag{22}$$

be an estimator for the conditional variance $\sigma^2_{W_i}(X_i)$.[6] The next theorem establishes consistency of an estimator of the conditional variance based on the estimators of $\sigma^2_{W_i}(X_i)$ defined in equation (22).

THEOREM 6: *Suppose that Assumptions 1 to 4 hold, and let $\widehat{\sigma}^2_{W_i}(X_i)$ be as in equation (22). Then*

$$\iota'_N \mathbf{A}\hat{\Omega}\mathbf{A}'\iota_N/N = \frac{1}{N}\sum_{i=1}^{N}\left(1 + \frac{K_M(i)}{M}\right)^2 \widehat{\sigma}^2_{W_i}(X_i) \xrightarrow{p} V^E.$$

[6]More generally one can use a number of nearest neighbors to estimate the local variances with the same result. Such estimates would have slightly higher bias but also lower variances.

6. An Application to the the Evaluation of a Labor Market Program

In this section we apply the estimators studied in this article to data from an evaluation of a job training program first analyzed by Lalonde (1986) and subsequently by Heckman and Hotz (1989), Dehejia and Wahba (1999) and Smith and Todd (2001).[7] We use experimental data from a randomized evaluation of the job training program and also a nonexperimental sample from the Panel Study of Income Dynamics (PSID). Using the experimental data we obtain an unbiased estimate of the average effect of the training. We then see how well the non-experimental matching estimates compare using the experimental trainees and the nonexperimental controls from the PSID. Given the size of the experimental and the PSID samples, and in line with previous studies using these data, we focus on the average effect for the treated and therefore only match the treated units.

Table 2 presents summary statistics for the three groups. The first two columns present the summary statistics for the experimental trainees. The second pair of columns presents summary statistics for the experimental controls. The third pair of columns presents summary statistics for the non-experimental control group constructed from the PSID. The last two columns present t-statistics for the hypothesis that the population averages for the trainees and the experimental controls, and for the trainees and the PSID controls, respectively, are zero. Panel A contains the results for pretreatment variables and Panel B for outcomes. Note the large differences in background characteristics between the trainees and the PSID sample. This is what makes drawing causal inferences from comparisons between the PSID sample and the trainee group a tenuous task. From Panel B, we can obtain an unbiased estimate of the effect of the training on earnings in 1978 by comparing the averages for the trainees and the experimental controls, $6.35 - 4.55 = 1.80$ with a standard error of 0.67 (earnings are measured in thousand dollars). Using a normal approximation to the limiting distribution of the effect of the training on earnings in 1978, we obtain a 95% confidence interval, which is [0.49, 3.10].

Table 3 presents estimates of the causal effect of training on earnings using various matching and regression adjustment estimators. Panel A reports estimates for the experimental data (experimental trainees and experimental controls). Panel B reports estimates based on the experimental trainees and the PSID controls. The first set of rows in each case reports matching estimates, based on a number of matches including 1, 4, 16, 64 and 2490. The matching estimates

[7] Programs for implementing the matching estimators in Matlab and STATA is available from the authors on the web at http://elsa.berkeley.edu/users/imbens/.

include simple matching with no bias adjustment, and bias-adjusted matching. All matching estimators use the Euclidean norm to measure the distance between different values for the covariates, after normalizing the covariates to have zero mean and unit variance. For the bias adjustment the regression uses all nine coe higher order terms. The bias correction is estimated using only the matched control units. Note that since we only match the treated units, there is no need to estimate the regression function for the trainees. The last three rows of each panel report estimates based on linear regression with no controls, all covariates linearly and all covariates with quadratic terms and a full set of interactions.

The experimental estimates range from 1.17 (bias corrected matching with one match) to 2.27 (quadratic regression). The non-experimental estimates have a much wider range, from -15.20 (simple difference) to 3.26 (quadratic regression). For the non-experimental sample, using a single match, there is little difference between the simple matching estimator and its bias-corrected version, 2.09 and 2.45 respectively. However, simple matching, without bias-correction, produces radically different estimates when the number of matches changes, a troubling result for the empirical implementation of these estimators. With $M \geq 16$ the simple matching estimator produces results outside the experimental 95% confidence interval. In contrast, the bias-corrected matching estimator shows a much more robust behavior when the number of matches changes: only with $M = 2490$ (that is, when all controls are matched to each treated) the bias-corrected estimate deteriorates to 0.84, still inside the experimental 95% confidence interval.

To see how well the simple matching estimator performs in terms of balancing the covariates, Table 4 reports average differences within the matched pairs. First, all the covariates are normalized to have zero mean and unit variance. The first two columns report the averages of the normalized covariates for the PSID controls and the experimental trainees. Before matching, the averages for some of the variables are more than one standard deviation apart, e.g., the earnings and employment variables. The next pair of columns reports the within-matched-pairs average difference and the standard deviation of this within-pair difference. For all the indicator variables the matching is exact: every trainee is matched to someone with the same ethnicity, marital status and employment history for the years 1974 and 1975. The other, more continuously distributed variables are not matched exactly, but the quality of the matches appears very high: the average difference within the pairs is very small compared to the average difference between trainees and controls before the matching, and it is also small compared to the standard deviations of these differences. If we increase the number of matches the quality of the matches goes down, with even

the indicator variables no longer matched exactly, but in most cases the average difference is still far smaller than the standard deviation till we get to 16 or more matches. As expected, matching quality deteriorates when the number of matches increases. This explains that, as shown in Table 3, the bias-correction matters more for larger M. The last row reports matching differences for logistic estimates of the propensity score. Although the matching is not directly on the propensity score, with single matches the average difference in the propensity score is only 0.21, whereas without matching the difference between trainees and controls is 8.16, 40 times higher.

7. A Monte Carlo Study

In this section, we discuss some simulations designed to assess the performance of the various matching estimators. To mimic as closely as possible the behavior of matching estimators in real applications, we simulated data sets that closely resemble the Lalonde data set analyzed in the previous section.

In the simulation we have nine regressors, designed to match the following variables in the Lalonde data set: age, education, black, hispanic, married, earnings1974, unemployed1974, earnings1975, unemployed1975. For each simulated data set we sampled with replacement 185 observations from the empirical covariate distribution of the trainees, and 2490 observations from the empirical covariate distribution of the PSID controls. This gives us the joint distribution of covariates and treatment indicators. For the conditional distribution of the outcome given covariates, we estimated a two-part model on the PSID controls, where the probability of zero earnings is a logistic function of the covariates with a full set of quadratic terms and interactions. Conditional on being positive, the log of earnings is a function of the covariates with again a full set of quadratic terms and interactions. We then assume a constant treatment effect of 2.0.

For each data set simulated in this way we report results for the same set of estimators. For each estimator we report the mean and median bias, the root-mean-squared-error (rmse), the median-absolute-error (mae), the standard deviation, the average estimated standard error, and the coverage rates for nominal 95% and 90% confidence intervals. The results are reported in Table 5.

In terms of rmse and mae, the bias-adjusted matching estimator is best with 4 or 16 matches. The simple matching estimator does not perform as well neither in terms of bias or rmse. The pure regression adjustment estimators do not perform very well. They have high rmse and substantial bias. Bias-corrected estimator also perform better in terms of coverage rates. Non-corrected

matching estimators and regression estimators have lower than nominal coverage rates for any value of M.

8. Conclusion

In this paper we derive large sample properties of simple matching estimators that are widely used in applied evaluation research. The formal large sample properties turn out to be surprisingly poor. We show out that simple matching estimators may include biases which do not disappear in large samples, under the standard $N^{1/2}$ normalization. We also show that matching estimators with a fixed number of matches are not efficient. We suggest a nonparametric bias-adjustment that renders matching estimators $N^{1/2}$-consistent. In simulations based on realistic settings for nonexperimental program evaluations, a simple implementation of this estimator where the bias-adjustment is based on linear regression appears to perform well compared to both matching estimators without bias-adjustment and regression-based estimators.

Appendix

Before proving Lemma 1, we collect some results on integration using polar coordinates that will be useful. See for example Stroock (1999). Let $S_k = \{\omega \in \mathbb{R}^k : \|\omega\| = 1\}$ be the unit k-sphere, and λ_{S_k} be its surface measure. Then, the area of the unit k-sphere is:

$$\int_{S_k} \lambda_{S_k}(d\omega) = \frac{2\pi^{k/2}}{\Gamma(k/2)}.$$

The volume of the unit k-ball is:

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

In addition,

$$\int_{S_k} \omega\, \lambda_{S_k}(d\omega) = 0,$$

and

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

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

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

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. Let $b(0) = 0$, $b(r) > 0$ for $r > 0$, and that for every $\tilde{r} > 0$ the infimum of $b(r)$ over $r \geq \tilde{r}$ is positive. Suppose that there exist positive real numbers a_0, b_0, α, 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 $\int_0^\infty a(r) \exp(-Nb(r))\, dr$ converges absolutely throughout its range for all sufficiently large N. Then, for $N \to \infty$

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

Proof: It follows from Theorem 7.1 in Olver (1997).

Proof of Lemma 1: First consider the conditional probability of unit i being the m-th closest match to z, given $X_i = x$:

$$\Pr(j_m = i | X_i = x)$$

$$= \left(\begin{array}{c} N-1 \\ m-1 \end{array} \right) (\Pr(\|X-z\| > \|x-z\|))^{N-m} (\Pr(\|X-z\| \leq \|x-z\|))^{m-1}$$

$$= \left(\begin{array}{c} N-1 \\ m-1 \end{array} \right) (1-\Pr(\|X-z\| \leq \|x-z\|))^{N-m} (\Pr(\|X-z\| \leq \|x-z\|))^{m-1}.$$

Since the marginal probability of unit i being the m-th closest match to z is $\Pr(j_m = i) = 1/N$, and the marginal density is $f(x)$, the distribution of X_i, conditional on it being the m-th closest match, is:

$$f_{X_i|j_m=i}(x) = Nf(x) \cdot \Pr(j_m = i|X_i = x)$$

$$= Nf(x) \left(\begin{array}{c} N-1 \\ m-1 \end{array} \right) (1-\Pr(\|X-z\| \leq \|x-z\|))^{N-m} (\Pr(\|X-z\| \leq \|x-z\|))^{m-1},$$

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

$$f_{U_m}(u) = N \left(\begin{array}{c} N-1 \\ m-1 \end{array} \right) f(z+u) (1-\Pr(\|X-z\| \leq \|u\|))^{N-m}$$

$$\times (\Pr(\|X-z\| \leq \|u\|))^{m-1}. \tag{A.1}$$

Transform to $V_m = N^{1/k} \cdot U_m$ with Jacobian N^{-1} to get:

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

Note that $\Pr(\|X-z\| \leq \|v\|N^{-1/k})$ is

$$\int_0^{\|v\|/N^{1/k}} r^{k-1} \left(\int_{S_k} f(z+r\omega) \lambda_{S_k}(d\omega) \right) dr,$$

where $S_k = \{\omega \in \mathbb{R}^k : \|\omega\| = 1\}$ is the unit k -sphere, and λ_{S_k} is its surface measure. The derivative w.r.t. N is

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

Therefore,

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

In addition, it is easy to check that

$$N^{1-m} \left(\begin{array}{c} N-1 \\ m-1 \end{array} \right) = \frac{1}{(m-1)!} + o(1).$$

Therefore,

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

The previous equation shows that the density of V_m converges pointwise to a non-negative function which is rotation invariant with respect to the origin. To check that this function defines a proper distribution, transform to polar coordinates and integrate:

$$\int_0^\infty \frac{f(z)}{(m-1)!}\, r^{k-1}\left(\int_{S_k}\left(r^k\frac{f(z)}{k}\int_{S_k}\lambda(d\omega)\right)^{m-1}\exp\left(-r^k\frac{f(z)}{k}\int_{S_k}\lambda(d\omega)\right)\lambda(d\omega)\right)dr$$
$$= \int_0^\infty \frac{kr^{mk-1}}{(m-1)!}\left(\frac{f(z)}{k}\int_{S_k}\lambda(d\omega)\right)^m\exp\left(-r^k\frac{f(z)}{k}\int_{S_k}\lambda(d\omega)\right)dr.$$

Transform $t = r^k$ to get

$$\int_0^\infty \frac{t^{m-1}}{(m-1)!}\left(\frac{f(z)}{k}\int_{S_k}\lambda(d\omega)\right)^m\exp\left(-t\,\frac{f(z)}{k}\int_{S_k}\lambda(d\omega)\right)dt,$$

which is equal to one because is the integral of the density of a gamma random variable with parameters $(m, k(f(z)\int_{S_k}\lambda(d\omega))^{-1})$ over its support. 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),

$$EU_m = N\begin{pmatrix} N-1 \\ m-1 \end{pmatrix} A_m,$$

where

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

Changing variables to polar coordinates gives:

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

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

$$\int_{\mathbb{R}^k} f(x)1\{\|x-z\|\le r\}dx = \int_{\mathbb{R}^k} f(z+v)1\{\|v\|\le r\}dv$$

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

and substituting this into the expression for A_m gives:

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

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

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

where

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

and

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

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

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

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

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

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

which is also continuous. Using L'Hospital's rule:

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

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

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

Therefore,

$$\lim_{r\to 0} q(r)r^{-(k+1)} = \lim_{r\to 0} \frac{dc}{dr}(r) = \frac{1}{k} \frac{\partial f}{\partial x}(z) \int_{S_k} \lambda_{S_k}(d\omega).$$

Similar calculations yield

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

Therefore

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

Now, it is clear that

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

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

$$\begin{aligned}
a_0 &= \left(\frac{1}{k}f(z)\int_{S_k}\lambda_{S_k}(d\omega)\right)^{m}\frac{1}{f(z)}\frac{\partial f}{\partial x}(z) \\
b_0 &= \frac{1}{k}f(z)\int_{S_k}\lambda_{S_k}(d\omega).
\end{aligned}$$

Applying Lemma A.1, we get

$$A_m = \Gamma\left(\frac{mk+2}{k}\right)\frac{a_0}{kb_0^{(mk+2)/k}}\frac{1}{N^{(mk+2)/k}} + o\left(\frac{1}{N^{(mk+2)/k}}\right)$$

$$= \Gamma\left(\frac{mk+2}{k}\right)\frac{1}{k}\left(\frac{f(z)}{k}\int_{S_k}\lambda_{S_k}(d\omega)\right)^{-2/k}\frac{1}{f(z)}\frac{df}{dx}(z)\frac{1}{N^{(mk+2)/k}} + o\left(\frac{1}{N^{(mk+2)/k}}\right).$$

$$= \Gamma\left(\frac{mk+2}{k}\right)\frac{1}{k}\left(f(z)\frac{\pi^{k/2}}{\Gamma\left(1+\frac{k}{2}\right)}\right)^{-2/k}\frac{1}{f(z)}\frac{df}{dx}(z)\frac{1}{N^{(mk+2)/k}} + o\left(\frac{1}{N^{(mk+2)/k}}\right).$$

Now, since

$$\lim_{N\to\infty}\frac{N^m/(m-1)!}{N\begin{pmatrix} N-1 \\ m-1 \end{pmatrix}} = 1,$$

we have that

$$E[U_m] = N\begin{pmatrix} N-1 \\ m-1 \end{pmatrix}A_m$$

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

which finishes the proof for the second result of the theorem.
To get the result for $E[U_m U_m']$, notice that

$$\mathbb{E}[U_m U_m'] = N \begin{pmatrix} N-1 \\ m-1 \end{pmatrix} B_m,$$

where

$$B_m = \int_{\mathbb{R}^k} uu'\, f(z+u)\,(1-\Pr(\|X-z\| \le \|u\|))^{N-m}\,(\Pr(\|X-z\| \le \|u\|))^{m-1}\, du.$$

Transforming to polar coordinates again leads to

$$B_m = \int_0^\infty r^{k-1} \left(\int_{S_k} r^2 \omega\omega'\, f(z+r\omega)\lambda_{S_k}(d\omega) \right) \left(1 - \int_0^r s^{k-1} \left(\int_{S_k} f(z+s\omega)\lambda_{S_k}(d\omega) \right) ds \right)^{N-m}$$

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

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

where, as before

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

and

$$\tilde{a}(r) = r^{k+1} \cdot \left(\int_{S_k} \omega\omega'\, f(z+r\omega)\lambda_{S_k}(d\omega) \right) \frac{\left(\int_0^r s^{k-1} \left(\int_{S_k} f(z+s\omega)\lambda_{S_k}(d\omega) \right) ds \right)^{m-1}}{\left(1 - \int_0^r s^{k-1} \left(\int_{S_k} f(z+s\omega)\lambda_{S_k}(d\omega) \right) ds \right)^{m}}.$$

That is, $\tilde{a}(r) = \tilde{q}(r)p(r)$, $\tilde{q}(r) = r^{k+1}\tilde{c}(r)$, and, as before, $p(r) = (g(r))^{m-1}$, where

$$\tilde{c}(r) = \frac{\displaystyle\int_{S_k} \omega\omega'\, f(z+r\omega)\lambda_{S_k}(d\omega)}{\displaystyle 1 - \int_0^r s^{k-1} \left(\int_{S_k} f(z+s\omega)\lambda_{S_k}(d\omega) \right) ds},$$

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

Clearly,

$$\lim_{r\to 0} \tilde{q}(r) r^{-(k+1)} = \lim_{r\to 0} \tilde{c}(r) = \frac{1}{k} f(z) \int_{S_k} \lambda_{S_k}(d\omega) I_k.$$

Hence,

$$\begin{aligned}\lim_{r\to 0}\tilde{a}(r)r^{-(mk+1)} &= \left(\lim_{r\to 0}p(r)r^{-(m-1)k}\right)\left(\lim_{r\to 0}\tilde{q}(r)r^{-(k+1)}\right)\\ &= \left(\frac{1}{k}f(z)\int_{S_k}\lambda_{S_k}(d\omega)\right)^{m-1}\frac{1}{k}f(z)\int_{S_k}\lambda_{S_k}(d\omega)I_k\\ &= \left(\frac{1}{k}f(z)\int_{S_k}\lambda_{S_k}(d\omega)\right)^{m}I_k.\end{aligned}$$

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

$$\begin{aligned}a_0 &= \left(\frac{1}{k}f(z)\int_{S_k}\lambda_{S_k}(d\omega)\right)^{m}I_k\\ b_0 &= \frac{1}{k}f(z)\int_{S_k}\lambda_{S_k}(d\omega).\end{aligned}$$

Applying Lemma A.1, we get

$$B_m = \Gamma\left(\frac{mk+2}{k}\right)\frac{a_0}{kb_0^{(mk+2)/k}}\frac{1}{N^{(mk+2)/k}} + o\left(\frac{1}{N^{(mk+2)/k}}\right)$$

$$= \Gamma\left(\frac{mk+2}{k}\right)\frac{1}{k}\left(\frac{f(z)}{k}\int_{S_k}\lambda_{S_k}(d\omega)\right)^{-2/k}\frac{1}{N^{(mk+2)/k}}\cdot I_k + o\left(\frac{1}{N^{(mk+2)/k}}\right).$$

$$= \Gamma\left(\frac{mk+2}{k}\right)\frac{1}{k}\left(f(z)\frac{\pi^{k/2}}{\Gamma\left(1+k/2\right)}\right)^{-2/k}\frac{1}{N^{(mk+2)/k}}\cdot I_k + o\left(\frac{1}{N^{(mk+2)/k}}\right).$$

Hence, using the fact that

$$\lim_{N\to\infty}\frac{N^m/(m-1)!}{N\begin{pmatrix} N-1 \\ m-1 \end{pmatrix}} = 1,$$

we have that

$$\mathbb{E}[U_m U_m'] = N\begin{pmatrix} N-1 \\ m-1 \end{pmatrix}\cdot B_m$$

$$= \Gamma\left(\frac{mk+2}{k}\right)\frac{1}{(m-1)!k}\left(f(z)\frac{\pi^{k/2}}{\Gamma\left(1+\frac{k}{2}\right)}\right)^{-2/k}\frac{1}{N^{2/k}}\cdot I + o\left(\frac{1}{N^{2/k}}\right).$$

Using the same techniques as for the first two moments,

$$E\|U_m\|^3 = N\begin{pmatrix} N-1 \\ m-1 \end{pmatrix}C_m,$$

where

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

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

and

$$\bar{a}(r) = r^{k+2}\cdot\left(\int_{S_k} f(z+r\omega)\lambda_{S_k}(d\omega)\right)\frac{\left(\int_0^r s^{k-1}\left(\int_{S_k} f(z+s\omega)\lambda_{S_k}(d\omega)\right)ds\right)^{m-1}}{\left(1-\int_0^r s^{k-1}\left(\int_{S_k} f(z+s\omega)\lambda_{S_k}(d\omega)\right)ds\right)^{m}}.$$

That is, $\bar{a}(r) = \bar{q}(r)p(r)$, $\bar{q}(r) = r^{k+2}\bar{c}(r)$, and $p(r) = (g(r))^{m-1}$, where

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

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

Now,

$$\lim_{r\to 0}\tilde{q}(r)r^{-(k+2)} = \lim_{r\to 0}\tilde{c}(r) = f(z)\int_{S_k}\lambda_{S_k}(d\omega).$$

Hence,

$$\begin{aligned}\lim_{r\to 0}\bar{a}(r)r^{-(mk+2)} &= \left(\lim_{r\to 0}p(r)r^{-(m-1)k}\right)\left(\lim_{r\to 0}\bar{q}(r)r^{-(k+2)}\right)\\ &= \left(\frac{1}{k}f(z)\int_{S_k}\lambda_{S_k}(d\omega)\right)^{m-1} f(z)\int_{S_k}\lambda_{S_k}(d\omega)\\ &= \left(\frac{1}{k}f(z)\int_{S_k}\lambda_{S_k}(d\omega)\right)^{m} k.\end{aligned}$$

Therefore, the conditions of Lemma A.1 hold for $\alpha = mk+3$, $\beta = k$

$$\begin{aligned}a_0 &= \left(\frac{1}{k}f(z)\int_{S_k}\lambda_{S_k}(d\omega)\right)^{m} k\\ b_0 &= \frac{1}{k}f(z)\int_{S_k}\lambda_{S_k}(d\omega).\end{aligned}$$

Applying Lemma A.1, we get

$$C_m = \Gamma\left(\frac{mk+3}{k}\right)\frac{a_0}{kb_0^{(mk+3)/k}}\frac{1}{N^{(mk+3)/k}} + o\left(\frac{1}{N^{(mk+3)/k}}\right)$$

$$= \Gamma\left(\frac{mk+3}{k}\right)\left(\frac{f(z)}{k}\int_{S_k}\lambda_{S_k}(d\omega)\right)^{-3/k}\frac{1}{N^{(mk+3)/k}} + o\left(\frac{1}{N^{(mk+3)/k}}\right).$$

$$= \Gamma\left(\frac{mk+3}{k}\right)\left(f(z)\frac{\pi^{k/2}}{\Gamma\left(1+k/2\right)}\right)^{-3/k}\frac{1}{N^{(mk+3)/k}} + o\left(\frac{1}{N^{(mk+3)/k}}\right).$$

Hence, using the fact that

$$\lim_{N\to\infty}\frac{N^m/(m-1)!}{N\left(\begin{array}{c} N-1 \\ m-1 \end{array}\right)} = 1,$$

we have that

$$E[\|U_m\|^3] = N\left(\begin{array}{c} N-1 \\ m-1 \end{array}\right)\cdot C_m$$

$$= \Gamma\left(\frac{mk+3}{k}\right)\frac{1}{(m-1)!}\left(f(z)\frac{\pi^{k/2}}{\Gamma\left(1+k/2\right)}\right)^{-3/k}\frac{1}{N^{3/k}} + o\left(\frac{1}{N^{3/k}}\right).$$

Therefore

$$E\|U_m\|^3 = O\left(\frac{1}{N^{3/k}}\right).$$

□

Proof of Lemma 2: The proof consists of showing that the density of $V_m = N^{1/k}\cdot U_m$, denoted by $f_{V_m}(v)$, is bounded by $\bar{f}_{V_m}(v)$ followed by a proof that $\int\|v\|^L\bar{f}_{V_m}(v)dv$ is uniformly bounded in N, for any $L>0$. It is enough to show the result for $N>m$ (bounded support guarantees finite moments of V_m for any given N, and in particular for $N=m$.) Recall from the proof of Lemma 1 that

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

Define $\underline{f} = \inf_{x\in\mathbb{X}} f(x)$ and $\bar{f} = \sup_{x\in\mathbb{X}} f(x)$. By assumption, $\underline{f}>0$ and $\bar{f}$ is finite. Let $\bar{u}$ be the diameter of $\mathbb{X}$ ($\bar{u} = \sup_{x,y\in\mathbb{X}}\|x-y\|$). Consider all the balls $B(x,\bar{u})$ with centers $x\in\mathbb{X}$ and radius $\bar{u}$. Let c 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 $0<c<1$, and that, since $\mathbb{X}$ is convex, this proportion can only increase for a smaller radius. Let $x\in\mathbb{X}$ and $\|v\|\le N^{1/k}\bar{u}$.

$$\begin{aligned} \Pr\left(\|X-z\|>\frac{\|v\|}{N^{1/k}}\right) &= 1-\int_0^{\|v\|N^{-1/k}} r^{k-1}\int_{S_k} f(z+r\omega)\lambda_{S_k}(d\omega)dr \\ &\le 1-c\underline{f}\int_0^{\|v\|N^{-1/k}} r^{k-1}\int_{S_k}\lambda_{S_k}(d\omega)dr \\ &= 1-c\frac{\|v\|^k}{N}\underline{f}\frac{\pi^{k/2}}{\Gamma(1+k/2)}. \end{aligned}$$

Note also that

$$0\le c\frac{\|v\|^k}{N}\underline{f}\frac{\pi^{k/2}}{\Gamma(1+k/2)} \le c\bar{u}^k\underline{f}\frac{\pi^{k/2}}{\Gamma(1+k/2)} \le 1.$$

Similarly,

$$\Pr\left(\|X-z\| \leq \frac{\|v\|}{N^{1/k}}\right) \leq \frac{\|v\|^k}{N}\bar{f}\frac{\pi^{k/2}}{\Gamma(1+k/2)}.$$

Hence, using the fact that for positive a, $\log(a) \leq a-1$ and thus for all $0 < b < N$ we have $(1-b/N)^{(N-m)} \leq \exp(-b(N-m)/N) \leq \exp(-b/(m+1))$, and that

$$N^{1-m}\begin{pmatrix} N-1 \\ m-1 \end{pmatrix} \leq \frac{1}{(m-1)!},$$

it follows that

$$f_{V_m}(v) \leq \frac{1}{(m-1)!}\bar{f}\exp\left(-\frac{\|v\|^k}{(m+1)}\underline{f}\frac{2\pi^{k/2}}{\Gamma(k/2)}\right)\left(\|v\|^k\bar{f}\frac{2\pi^{k/2}}{\Gamma(k/2)}\right)^{m-1}$$

$$= C_1 \cdot \|v\|^{k(m-1)} \cdot \exp\left(-C_2 \cdot \|v\|^k\right),$$

with C_1 and C_2 positive. This inequality holds trivially for $\|v\| > N^{1/k}\bar{u}$. This establishes an exponential bound that does not depend on N or z. Hence for all N and z, $\int \|v\|^L f_{V_m}(v)dv$ is finite and thus all moments of $N^{1/k} \cdot U_m$ are uniformly bounded in N and z. □

Proof of Theorem 1:
For part (i) of the theorem, define the unit-level matching discrepancy $U_{m,i} = X_i - X_{j_m(i)}$, and tm from the m-th match:

$$B^{sm}_{m,i} = W_i \cdot \left(\mu_0(X_i) - \mu_0(X_{j_m(i)})\right) - (1-W_i) \cdot \left(\mu_1(X_i) - \mu_1(X_{j_m(i)})\right)$$

$$= W_i \cdot (\mu_0(X_i) - \mu_0(X_i + U_{m,i})) - (1-W_i) \cdot (\mu_1(X_i) - \mu_1(X_i + U_{m,i}))\,.$$

Hence $|B^{sm}_{m,i}| \leq C \cdot \|U_{m,i}\|$, where $C = \sup_x \|\partial\mu_0(x)/\partial x\| + \sup_x \|\partial\mu_1(x)/\partial x\|$ which is finite by assumption. The bias term is

$$B^{sm}_M = \frac{1}{N \cdot M}\sum_{i=1}^{N}\sum_{m=1}^{M} B^{sm}_{m,i}.$$

Now consider

$$\mathbb{E}[N^{2/k}(B^{sm})^2] = N^{2/k} \cdot \mathbb{E}\left[\frac{1}{N^2 \cdot M^2}\sum_{i,j}\sum_{l,m} B^{sm}_{m,i} \cdot B^{sm}_{l,j}\right]$$

$$\leq N^{2/k} \cdot \max_{m,i}\mathbb{E}\left[(B^{sm}_{m,i})^2\right] \leq C^2 \cdot \max_{m,i}\mathbb{E}\left[(N^{1/k}\|U_{m,i}\|)^2\right].$$

By Lemma 1, for any given m, the second moment of $N^{1/k}U_{m,i}$ is finite for all N and i. Since m only takes on M values, $N^{1/k}B^{sm}$ has a finite second moment, and B^{sm} is $O_p(N^{-1/k})$, proving the first part of the theorem.
Next consider the second part of the theorem. The bias is

$$\text{Bias}^{sm} = \mathbb{E}[B^{sm}] = \mathbb{E}[\widehat{Y}_i(1) - \widehat{Y}_i(0)] - \mathbb{E}[Y_i(1) - Y_i(0)]$$

$$= \int E\left[(\widehat{Y}_i(1) - \widehat{Y}_i(0)) - (Y_i(1) - Y_i(0))\Big| X_i = x\right] f(x)\, dx$$

$$= (1-p) \int E\left[(1/M)\textstyle\sum_{j\in\mathcal{J}_M(i)} \mu_1(X_j) - \mu_1(X_i)\Big| X_i = x, W_i = 0\right] f_0(x)\, dx$$

$$- \ p \int E\left[(1/M)\textstyle\sum_{j\in\mathcal{J}_M(i)} \mu_0(X_j) - \mu_0(X_i)\Big| X_i = x, W_i = 1\right] f_1(x)\, dx. \tag{A.2}$$

Applying a second order Taylor expansion

$$\mathbb{E}[\mu_1(X_{j_m(i)}) - \mu_1(X_i)|X_i = x, W_i = 0, \iota_N'\mathbf{W} = N_1]$$

$$= \mathbb{E}[(X_{j_m(i)} - x)'|X_i = x, W_i = 0, \iota_N'\mathbf{W} = N_1]\frac{\partial \mu_1}{\partial X}(x)$$

$$+\frac{1}{2}\operatorname{tr}\left(\frac{\partial^2 \mu_1}{\partial X'\partial X}(x)\mathbb{E}[(X_{j_m(i)} - x)(X_{j_m(i)} - x)'|X_i = x, W_i = 0, \iota_N'\mathbf{W} = N_1]\right) + R(x),$$

where $|R(x)| = O\left(\mathbb{E}[\|X_{j_m(i)} - x\|^3|X_i = x, W_i = 0, \iota_N'\mathbf{W} = N_1]\right)$. Applying Lemma 1, we get

$$\mathbb{E}[\mu_1(X_{j_m(i)}) - \mu_1(X_i)|X_i = x, W_i = 0, \iota_N'\mathbf{W} = N_1]$$

$$= \Gamma\left(\frac{mk+2}{k}\right)\frac{1}{(m-1)!k}\left(f_1(x)\frac{\pi^{k/2}}{\Gamma\left(1+k/2\right)}\right)^{-2/k}$$

$$\times\left\{\frac{1}{f_1(x)}\frac{\partial f_1}{\partial X'}(x)\frac{\partial \mu_1}{\partial X}(x) + \frac{1}{2}\operatorname{tr}\left(\frac{\partial^2 \mu_1}{\partial X'\partial X}(x)\right)\right\}\cdot\frac{1}{N_1^{2/k}} + o\left(\frac{1}{N_1^{2/k}}\right)$$

Note that

$$\sum_{N_1=M}^{N-M}\frac{1}{N_1^{2/k}}\cdot\Pr\left(\iota_N'\mathbf{W} = N_1|X_i = x, W_i = 0\right)$$

$$= \sum_{N_1=M}^{N-M}\frac{1}{N_1^{2/k}}\cdot\binom{N}{N_1}p^{N_1}(1-p)^{N-N_1}$$

$$= \frac{1}{p^{2/k}N^{2/k}}\sum_{N_1=M}^{N-M}\frac{p^{2/k}}{(N_1/N)^{2/k}}\cdot\binom{N}{N_1}p^{N_1}(1-p)^{N-N_1} = \frac{1}{p^{2/k}N^{2/k}} + o\left(\frac{1}{N^{2/k}}\right),$$

since $N_1/N = p + o_p(1)$. Therefore,

$$E\left[(1/M)\textstyle\sum_{j\in\mathcal{J}_M(i)} \mu_1(X_j) - \mu_1(X_i)|X_i = x, W_i = 0\right]$$

$$= \left(\frac{1}{M} \sum_{m=1}^{M} \Gamma\left(\frac{mk+2}{k}\right) \frac{1}{(m-1)!k} \right) \left(f_1(x) \frac{\pi^{k/2}}{\Gamma\left(1+k/2\right)} \right)^{-2/k} \frac{1}{p^{2/k}}$$

$$\times \left\{ \frac{1}{f_1(x)} \frac{\partial f_1}{\partial X'}(x) \frac{\partial \mu_1}{\partial X}(x) + \frac{1}{2} \operatorname{tr}\left(\frac{\partial^2 \mu_1}{\partial X' \partial X}(x) \right) \right\} \cdot \frac{1}{N^{2/k}} + o\left(\frac{1}{N^{2/k}} \right). \tag{A.3}$$

Similarly,

$$E\left[(1/M) \textstyle\sum_{j \in \mathcal{J}_M(i)} \mu_0(X_j) - \mu_0(X_i) | X_i = x, W_i = 1 \right]$$

$$= \left(\frac{1}{M} \sum_{m=1}^{M} \Gamma\left(\frac{mk+2}{k}\right) \frac{1}{(m-1)!k} \right) \left(f_0(x) \frac{\pi^{k/2}}{\Gamma\left(1+k/2\right)} \right)^{-2/k} \frac{1}{(1-p)^{2/k}}$$

$$\times \left\{ \frac{1}{f_0(x)} \frac{\partial f_0}{\partial X'}(x) \frac{\partial \mu_0}{\partial X}(x) + \frac{1}{2} \operatorname{tr}\left(\frac{\partial^2 \mu_0}{\partial X' \partial X}(x) \right) \right\} \cdot \frac{1}{N^{2/k}} + o\left(\frac{1}{N^{2/k}} \right). \tag{A.4}$$

Combine equations (A.2), (A.3), and (A.4) to obtain the result. □

Corollary 1 follows directly from Theorem 1, and its proof is therefore omitted.

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 $\mathbb{X}$ be a compact and convex set of dimension equal to k and $\bar{u} = \sup_{x,y \in \mathbb{X}} \|x - y\|$. Consider all the balls $B(x,u)$ with centers $x \in \mathbb{X}$ 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, since $\mathbb{X}$ is convex, this proportion nondecreasing in u, so let $c = c(\bar{u})$, and $c(u) \geq c$ for $u \leq \bar{u}$.
The proof consists of three parts. First we derive an exponential bound 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 $A_M(i)$, the subset of $\mathbb{X}$ such that each observation, j, with $W_j = 1 - W_i$ and $X_j \in A_M(i)$ is matched to i:

$$A_M(i) = \left\{ x \;\middle|\; \textstyle\sum_{j \,|\, W_j = W_i} 1\{\|X_j - x\| \leq \|X_i - x\|\} \leq M \right\}.$$

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,

$$\Pr\left(\|X_j - X_i\| > u \cdot N_{1-W_i}^{-1/k} \,\middle|\, W_1, \ldots, W_N, W_j = 1 - W_i, X_i = x \right)$$

$$= 1 - \int_0^{u N_{1-W_i}^{-1/k}} r^{k-1} \int_{S_k} f_{1-W_i}(x + r\omega) \lambda_{S_k}(d\omega) dr$$

$$\leq 1 - c\underline{f} \int_0^{u N_{1-W_i}^{-1/k}} r^{k-1} \int_{S_k} \lambda_{S_k}(d\omega) dr$$

$$= 1 - c\underline{f} u^k N_{1-W_i}^{-1} \pi^{k/2} / \Gamma(1 + k/2).$$

Similarly

$$\Pr\left(\|X_j - X_i\| \le u \cdot N_{1-W_i}^{-1/k} \middle| W_1, \ldots, W_N, W_j = 1 - W_i, X_i = x\right) \le \bar{f} u^k N_{1-W_i}^{-1} \pi^{k/2} / \Gamma(1 + k/2).$$

Thus

$$\Pr\left(\|X_j - X_i\| > u \cdot N_{1-W_i}^{-1/k} \middle| W_1, \ldots, W_N, X_i = x, j \in \mathcal{J}_M(i)\right)$$

$$\le \Pr\left(\|X_j - X_i\| > u \cdot N_{1-W_i}^{-1/k} \middle| W_1, \ldots, W_N, X_i = x, j = j_M(i)\right)$$

$$= \sum_{m=0}^{M-1} \binom{N_{1-W_i}}{m} \Pr\left(\|X_j - X_i\| > u \cdot N_{1-W_i}^{-1/k} \middle| W_1, \ldots, W_N, W_j = 1 - W_i, X_i = x\right)^{N_{1-W_i}-m}$$
$$\cdot \Pr\left(\|X_j - X_i\| \le u \cdot N_{1-W_i}^{-1/k} \middle| W_1, \ldots, W_N, W_j = 1 - W_i, X_i = x\right)^{m}.$$

Notice that

$$\binom{N_{1-W_i}}{m} \Pr\left(\|X_j - X_i\| \le u \cdot N_{1-W_i}^{-1/k} \middle| W_1, \ldots, W_N, W_j = 1 - W_i, X_i = x\right)^{m}$$

$$\le \frac{1}{m!} \left(u^k \bar{f} \frac{\pi^{k/2}}{\Gamma(1 + k/2)}\right)^{m}.$$

Therefore,

$$\Pr\left(\|X_j - X_i\| > u \cdot N_{1-W_i}^{-1/k} \middle| W_1, \ldots, W_N, X_i = x, j \in \mathcal{J}_M(i)\right)$$

$$\le \sum_{m=0}^{M-1} \frac{1}{m!} \left(u^k \bar{f} \frac{\pi^{k/2}}{\Gamma(1 + k/2)}\right)^{m} \left(1 - u^k c\underline{f} \frac{\pi^{k/2}}{\Gamma(1 + k/2)} \cdot \frac{1}{N_{1-W_i}}\right)^{N_{1-W_i}-m}.$$

Then, for some constant $C_1 > 0$,

$$\Pr\left(\|X_j - X_i\| > u \cdot N_{1-W_i}^{-1/k} \middle| W_1, \ldots, W_N, X_i = x, j \in \mathcal{J}_M(i)\right)$$

$$\le C_1 \max\{1, u^{k(M-1)}\} \sum_{m=0}^{M-1} \left(1 - u^k c\underline{f} \frac{\pi^{k/2}}{\Gamma(1 + k/2)} \cdot \frac{1}{N_{1-W_i}}\right)^{N_{1-W_i}-m}$$

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

(The last inequality holds because for $a > 0$, $\log a \le a - 1$.) Note that this bound also holds for $u \ge N_{1-W_i}^{1/k} \bar{u}$, since in that case the probability that $\|X_{j_m(i)} - X_i\| > u \cdot N_{1-W_i}^{-1/k}$ is zero. Since the bound does not depend on x, this inequality also holds without conditioning on x.

Next, we consider for unit i, the volume $B_M(i)$ of the catchment area $A_M(i)$, defined as:

$$B_{M(i)} = \int_{A_{M(i)}} dx,$$

Conditional on $W_1, \ldots, W_N$, the match $j \in \mathcal{J}_M(i)$, and $A_M(i)$ the distribution of X_j is proportional to $f_{1-W_i}(x) \cdot 1\{x \in A_M(i)\}$. Note that a ball with radius $(b/2)^{1/k}/(\pi^{k/2}/\Gamma(1+k/2))^{1/k}$ has volume $b/2$. Therefore

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

As a result, if

$$\Pr\left(\|X_j - X_i\| > \frac{(b/2)^{1/k}}{(\pi^{k/2}/\Gamma(1+k/2))^{1/k}} \,\middle|\, W_1, \ldots, W_N, j \in \mathcal{J}_M(i) \right) \leq \delta \, \frac{\underline{f}}{2\bar{f}}, \tag{A.5}$$

then it must be the case that $\Pr(B_M(i) \geq b \,|\, W_1, \ldots, W_N, j \in \mathcal{J}_M(i)) \leq \delta$. In fact, the inequality in equation (A.5) has been established above for

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

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

$$\Pr(N_{1-W_i} \, B_M(i) \geq t \,|\, W_1, \ldots, W_N, j \in \mathcal{J}_M(i)) \leq C_2 \max\{1, C_3 t^{M-1}\} \exp(-C_4 t),$$

for some positive constants, C_2, C_3, and C_4. This establishes an uniform exponential bound, so all the moments of $N_{1-W_i} B_M(i)$ exist conditional on $W_1, \ldots, W_N, j \in \mathcal{J}_M(i)$ (uniformly in N). Since conditioning on $j \in \mathcal{J}_M(i)$ only increases the moments of $B_M(i)$ we conclude that all the moments of $N_{1-W_i} B_M(i)$ are uniformly bounded 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. Conditional on the catchment area, $A_M(i)$, and on $W_1, \ldots, W_N$, the distribution is binomial with parameters N_{1-W_i} and $P_M(i)$, where the probability of a catch is the integral of the density over the catchment area:

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

Therefore, conditional on $A_M(i)$, and $W_1, \ldots, W_N$, the r-th moment of $K_M(i)$ is

$$\mathbb{E}[K_M^r(i) | A_M(i), W_1, \ldots, W_N,] = \sum_{n=1}^{r} \frac{S(r,n) N_{1-W_i}! \, P_M(i)^n}{(N_{1-W_i} - n)!} \leq \bar{f} \sum_{n=1}^{r} S(r,n) (N_{1-W_i} B_M(i))^n,$$

where $S(r,n)$ are Stirling numbers of the second kind. Then,

$$\mathbb{E}[K_M^r(i)] \leq \bar{f} \sum_{n=1}^{r} S(r,n) \cdot \mathbb{E}\left[\left(\frac{N_{1-W_i}}{N_{W_i}} \right)^n (N_{W_i} B_M(i))^n \right]$$

is uniformly bounded in N (by the Law of Iterated Expectations and Hölder's Inequality). This proves the first part of the Lemma.

Next, consider part (ii) of Lemma 3. Because the moments of $K_M(i)$ are bounded uniformly in N, and because the variance $\sigma_w^2(x)$ is bounded by $\bar{\sigma}^2 = \sup_{w,x} \sigma_w^2(x)$, finite by Assumption 3, the expectation of

$(1+K_M/M)^2\sigma_w^2(x)$ is bounded by $\bar{\sigma}^2\mathbb{E}[(1+K_M/M)^2]$, and hence the expectation of $(1+K_M/M)^2\sigma_W^2(X)$ is finite.

□

Proof of Theorem 2:
Consider the three terms. First, by a standard law of large numbers $\bar{\tau}(X) \xrightarrow{p} \tau$. Second, by Theorem 1, $B^{sm} = O_p(N^{-1/k}) = o_p(1)$. Third, $N \cdot \mathbb{E}[(E^{sm})^2] = V^E$, with V^E finite, so that $E^{sm} = O_p(N^{-1}) = o_p(1)$.

Proof of Theorem 3:
First, consider the contribution of $\sqrt{N}(\overline{\tau(X)} - \tau)$. By a standard central limit theorem

$$\sqrt{N} \cdot (\overline{\tau(X)} - \tau) \xrightarrow{d} \mathcal{N}(0, V^{\tau(X)}). \tag{A.6}$$

Second, consider the contribution of $\sqrt{N} \cdot E^{sm}$:

$$\sqrt{N} \cdot E^{sm} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} E_i^{sm}.$$

Conditional on $\mathbf{W}$ and $\mathbf{X}$ the unit-level terms E_i^{sm} are independent with non-identical distributions. All conditional means are zero. The conditional variances are $(1 + K_M(i)/M)^2 \cdot \sigma_{W_i}^2(X_i)$. We will use a Lindeberg-Feller central limit theorem. Define

$$\Omega_N^2 = \sum_{i=1}^{N} (1 + K_M(i)/M)^2 \sigma_{W_i}^2(X_i),$$

as the sum of the variances. We will show that the Lindeberg-Feller condition that for all $\varepsilon > 0$

$$\frac{1}{\Omega_N^2} \sum_{i=1}^{N} \mathbb{E}\left[(E_i^{sm})^2 1\{|E_i^{sm}| \geq \varepsilon \Omega_N\}\right] \to 0, \tag{A.7}$$

is satisfied almost surely. First note that if the Lth moment of ε_i is finite, then the Lth moment of E_i^{sm} is finite. Hence, by Assumption 1, and Lemma 3, $\mathbb{E}[(E_i^{sm})^4]$ is finite. To prove that (A.7) condition holds, note that by Hölder's inequality we have

$$\mathbb{E}\left[(E_i^{sm})^2 1\{|E_i^{sm}| \geq \varepsilon \Omega_N\}\right] \leq \left(\mathbb{E}\left[(E_i^{sm})^4\right]\right)^{1/2} \left(\mathbb{E}\left[1\{|E_i^{sm}| \geq \varepsilon \Omega_N\}\right]\right)^{1/2}$$

$$\leq \left(\mathbb{E}\left[(E_i^{sm})^4\right]\right)^{1/2} \left(Pr\left(|E_i^{sm}| \geq \varepsilon \Omega_N\right)\right)$$

$$\leq \left(\mathbb{E}\left[(E_i^{sm})^4\right]\right)^{1/2} \Pr\left(\max_{j=1,\ldots,N} (E_j^{sm})^2 \geq \varepsilon^2 \Omega_N^2\right).$$

Hence

$$\frac{1}{\Omega_N^2} \sum_{i=1}^{N} \mathbb{E}\left[(E_i^{sm})^2 1\{|E_i^{sm}| \geq \varepsilon \Omega_N\}\right]$$

$$\leq \frac{1}{\Omega_N^2} \sum_{i=1}^{N} \left(\mathbb{E}\left[(E_i^{sm})^4\right]\right)^{1/2} \left(\Pr\left(\max_{j=1,\ldots,N} (E_j^{sm})^2 \geq \varepsilon^2 \Omega_N^2\right)\right).$$

$$\leq \frac{1}{\Omega_N^2/N} \left(\mathbb{E}\left[(E_i^{sm})^4\right]\right)^{1/2} \Pr\left(\max_{j=1,\dots,N}(E_j^{sm})^2 \geq \varepsilon^2\Omega_N^2\right).$$

Since $\Omega_N/N > \inf_{w,x}\sigma_w^2(x) > 0$, this is bounded by

$$C \cdot \Pr\left(\max_{j=1,\dots,N}(E_j^{sm})^2 \geq \varepsilon^2\Omega_N^2\right).$$

Note that the second factor converges to zero as $\max_i(E_i^{sm})^2$ is of order $o_p(N^{1/2})$ since the second moment of E_i^{sm} exists by assumption. Hence the Lindeberg-Feller condition (A.7) is satisfied.
Because (A.7) holds, we have

$$\frac{N^{1/2}\sum_{i=1}^N E_i^{sm}}{\sum_{i=1}^N(1+K_M(i)/M)^2\sigma^2_{W_i}(X_i)} = \frac{N^{3/2}\cdot E^{sm}}{\Omega_N} \xrightarrow{d} \mathcal{N}(0,1), \tag{A.8}$$

by the Lindeberg-Feller central limit theorem.
Next, we show

$$\Omega_N^2/N \longrightarrow V^E = \mathbb{E}[(1+K_M(i)/M)^2\sigma^2_{W_i}(X_i)].$$

First note that the expectation of Ω_N^2/N is equal to $\mathbb{E}[(1+K_M(i))^2\sigma^2_{W_i}(X_i)]$. Second, consider the variance

$$\frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N \mathbb{E}\left[\left((1+K_M(i)/M)^2\sigma^2_{W_i}(X_i) - \mathbb{E}[(1+K_M(i)/M)^2\sigma^2_{W_i}(X_i)]\right)\right.$$

$$\left.\times\left((1+K_M(j)/M)^2\sigma^2_{W_j}(X_j) - \mathbb{E}[(1+K_M(i)/M)^2\sigma^2_{W_i}(X_i)]\right)\right].$$

Now the volumes of the catchment areas B_i satisfy $Pr(B_i \leq b | B_j \geq c, W_i = W_j) \geq Pr(B_i \leq b | W_i = W_j)$. To see this note that the B_i all have the same distribution. Hence given the adding up condition (the catchment areas partition the covariate space), it must be that conditional on B_j being larger than c, the distribution function of all others must increase. This makes the volumes B_i and B_j negatively correlated. Hence the counts $K_M(i)$ and $K_M(j)$ are negatively correlated, and thus the covariances are negative, implying that the sum is less than the sum of the terms with $i = j$, so that the variance is less than or equal to

$$\frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left((1+K_M(i)/M)^2\sigma^2_{W_i}(X_i) - \mathbb{E}[(1+K_M(i))^2\sigma^2_{W_i}(X_i)]\right)^2\right].$$

$$= \frac{1}{N}\mathbb{E}\left[\left((1+K_M(i)/M)^2\sigma^2_{W_i}(X_i) - \mathbb{E}[(1+K_M(i))^2\sigma^2_{W_i}(X_i)]\right)^2\right].$$

The expectation is finite because $K_M(i)$ has finite fourth moment, so the variance goes to zero.
Hence $N^{3/2}\cdot E^{sm}/\Omega_N \xrightarrow{d} N^{1/2}E^{sm}/V^E$, and thus

$$\sqrt{N}\cdot E^{sm} \xrightarrow{d} \mathcal{N}(0, V^E). \tag{A.9}$$

Finally, E^{sm} and $\overline{\tau(X)} - \tau$ are uncorrelated (take expectations conditional on $\mathbf{X}$ and $\mathbf{W}$). Thus, combining (A.6), (A.9) and the zero correlation gives the result in the theorem. □

Corollaries 2 and 3 follow directly from Theorem 3 and their proofs are therefore omitted.

Before proving Theorem 4, we give some preliminary results. The exact conditional distribution of $K_M(i)$ is,

$$K_M(i) \;\Big|\; \mathbf{W}, \{X_j\}_{W_j=1}, W_i = 1 \;\sim\; \text{Binomial}\left(N_0, \int_{A_M(i)} f_0(z)dz\right),$$

and

$$K_M(i) \;\Big|\; \mathbf{W}, \{X_j\}_{W_j=0}, W_i = 0 \;\sim\; \text{Binomial}\left(N_1, \int_{A_M(i)} f_1(z)dz\right).$$

Let us describe the set $A_M(i)$ in more detail for the special case in which X is a scalar. First, let $\bar{r}_w(x)$ be the number of units with $W_i = w$ and $X_i \geq x$. Then, define $X_{(i,k)} = X_j$ if $\bar{r}_{W_i}(X_i) - \bar{r}_{W_i}(X_j) = k$, and $\bar{r}_{W_i}(X_i) - \lim_{x \uparrow X_j} \bar{r}_{W_i}(x) = k - 1$. Then the set $A_M(i)$ is equal to the interval

$$A_M(i) = (X_i/2 + X_{(i,-M)}/2, X_i/2 + X_{(i,M)})/2),$$

with width $(X_{(i,M)} - X_{(i,-M)})/2$.

LEMMA A.2: *Given $X_i = x$, and $W_i = 1$*

$$2N_1 \cdot \frac{f_1(x)}{f_0(x)} \cdot \int_{A_M(i)} f_0(z)dz \xrightarrow{d} \text{Gamma } (2M, 1),$$

and given $X_i = x$ and $W_i = 0$,

$$2N_0 \cdot \frac{f_0(x)}{f_1(x)} \cdot \int_{A_M(i)} f_1(z)dz \xrightarrow{d} \text{Gamma } (2M, 1).$$

PROOF: We only prove the first part of the lemma. The second part follows exactly the same proof. First we establish that

$$2N_1 f_1(x) \cdot \int_{A_M(i)} dz = N_1 f_1(x) \left(X_{(i,M)} - X_{(i,-M)}\right) + o_p(1) \xrightarrow{d} \text{Gamma } (2M, 1).$$

Let $F_1(x)$ be the distribution function of X given $W = 1$. Then $D = F_1(X_{(i,+M)}) - F_1(X_{(i,-M)})$ is the difference in order statistics of the uniform distribution, $2M$ orders apart. Hence the exact distribution of D is Beta with parameters $2M$ and N_1. For large N_1, the distribution of $N_1 D$ is then Gamma with parameters $2M$ and 1. Now approximate $N_1 D$ as

$$N_1 D = N_1 \cdot \left(F_1(X_{(i,M)}) - F_1(X_{(i,-M)})\right) = N_1 f_1(\tilde{X}_i) \cdot \left(X_{(i,M)} - X_{(i,-M)}\right).$$

For $\tilde{X}_i \in (X_{(i,-M)}, X_{(i,M)})$. The first claims follows because $\tilde{X}_i$ converges almost surely to x. Second, we show that

$$2N_1 \frac{f_1(x)}{f_0(x)} \cdot \int_{A_M(i)} f_0(z)dz - 2N_1 f_1(x) \cdot \int_{A_M(i)} dz = o_p(1).$$

This difference can be written as

$$2N_1 \frac{f_1(x)}{f_0(x)} \cdot \int_{A_M(i)} f_0(z)dz - 2N_1 \frac{f_1(x)}{f_0(x)} \cdot \int_{A_M(i)} f_0(x)dz$$

$$= 2N_1 \frac{f_1(x)}{f_0(x)} \cdot \left(\int_{A_M(i)} (f_0(z) - f_0(x))dz\right).$$

Notice that

$$\left|\left(\int_{A_M(i)}(f_0(z)-f_0(x))dz\right)\Big/\left(\int_{A_M(i)}dz\right)\right| \leq$$
$$\sup\left|\frac{\partial f_0}{\partial z}\right|\left(\int_{A_M(i)}|z-x|\,dz\right)\Big/\left(\int_{A_M(i)}dz\right) \leq \sup\left|\frac{\partial f_0}{\partial z}\right|\left(\int_{A_M(i)}dz\right) = o_p(1).$$

because $|\partial f_0/\partial z|$ is bounded and $A_M(i)$ vanishes asymptotically. Thus,

$$\left|2N_1\frac{f_1(x)}{f_0(x)}\cdot\left(\int_{A_M(i)}(f_0(z)-f_0(x))dz\right)\right| \leq$$
$$\left|2N_1\frac{f_1(x)}{f_0(x)}\cdot\int_{A_M(i)}dz\right|\cdot\left|\left(\int_{A_M(i)}(f_0(z)-f_0(x))dz\right)\Big/\left(\int_{A_M(i)}dz\right)\right| = o_p(1).$$

□

Proof of Theorem 4:

Consider

$$\begin{aligned}E\left[\left(1+\frac{K_M(i)}{M}\right)^2\sigma^2_{W_i}(X_i)\right] &= E\left[\left(1+\frac{K_M(i)}{M}\right)^2\sigma_1^2(X_i)\,\middle|\,W_i=1\right]p \\ &+ E\left[\left(1+\frac{K_M(i)}{M}\right)^2\sigma_0^2(X_i)\,\middle|\,W_i=0\right](1-p).\end{aligned}$$

Define $P_M(i) = W_i\cdot\int_{A_M(i)}f_0(z)dz + (1-W_i)\cdot\int_{A_M(i)}f_1(z)dz$. We know that

$$\mathbb{E}[K_M(i)|\mathbf{W},\{X_j\}_{W_j=1},W_i=1] = N_0P_M(i),$$

and

$$\mathbb{E}[K_M^2(i)|\mathbf{W},\{X_j\}_{W_j=1},W_i=1] = N_0P_M(i)+N_0(N_0-1)P_M(i)^2.$$

Therefore,

$$\mathbb{E}\left[\left(1+\frac{K_M(i)}{M}\right)^2\sigma_1^2(X_i)\,\middle|\,W_i=1\right]$$
$$= \mathbb{E}\left[\left(1+\frac{1}{M^2}\left\{N_0P_M(i)+N_0(N_0-1)P_M(i)^2\right\}+\frac{2}{M}N_0P_M(i)\right)\sigma_1^2(X_i)\,\middle|\,W_i=1\right].$$

From the previous results, this expectation is equal to

$$\mathbb{E}\left[\left(1+\frac{1}{M}\left\{\frac{(1-p)}{p}\frac{f_0(X_i)}{f_1(X_i)}+\frac{(1-p)^2}{2p^2}\frac{f_0(X_i)^2}{f_1(X_i)^2}(2M+1)\right\}+\frac{2(1-p)}{p}\frac{f_0(X_i)}{f_1(X_i)}\right)\sigma_1^2(X_i)\,\middle|\,W_i=1\right]+o(1).$$

Rearranging terms, we get:

$$\mathbb{E}\left[\left(1+\frac{K_M(i)}{M}\right)^2\sigma_1^2(X_i)\,\middle|\,W_i=1\right] = \mathbb{E}\left[\left(1+\frac{(1-p)}{p}\frac{f_0(X_i)}{f_1(X_i)}\right)^2\sigma_1^2(X_i)\,\middle|\,W_i=1\right]$$
$$+\frac{1}{M}\mathbb{E}\left[\left(\frac{(1-p)}{p}\frac{f_0(X_i)}{f_1(X_i)}+\frac{(1-p)^2}{2p^2}\frac{f_0(X_i)^2}{f_1(X_i)^2}\right)\sigma_1^2(X_i)\,\middle|\,W_i=1\right]+o(1).$$

Notice that,

$$\mathbb{E}\left[\left(1+\frac{(1-p)}{p}\frac{f_0(X_i)}{f_1(X_i)}\right)^2\sigma_1^2(X_i)\,\middle|\,W_i=1\right]p$$
$$=\mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)^2}\,\middle|\,W_i=1\right]p=\int\frac{\sigma_1^2(x)}{e(x)}\frac{pf_1(x)}{e(x)f(x)}f(x)dx=\mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)}\right].$$

In addition,

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{(1-p)}{p}\frac{f_0(X_i)}{f_1(X_i)}+\frac{(1-p)^2}{2p^2}\frac{f_0(X_i)^2}{f_1(X_i)^2}\right)\sigma_1^2(X_i)\,\middle|\,W_i=1\right]p\\&=\int\left((1-p)f_0(x)+\frac{(1-p)^2}{2p}\frac{f_0(x)^2}{f_1(x)}\right)\sigma_1^2(x)dx\\&=\frac{1}{2}\int(1-p)f_0(x)\left(2+\frac{1-p}{p}\frac{f_0(x)}{f_1(x)}\right)\sigma_1^2(x)dx=\frac{1}{2}\int(1-p)f_0(x)\left(1+\frac{1}{e(x)}\right)\sigma_1^2(x)dx\\&=\mathbb{E}\left[(1-e(X_i))\left(1+\frac{1}{e(X_i)}\right)\sigma_1^2(X_i)\right]=\mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)}-e(X_i)\sigma_1^2(x)\right].\end{aligned}$$

Therefore,

$$\mathbb{E}\left[\left(1+\frac{K_M(i)}{M}\right)^2\sigma_1^2(X_i)\,\middle|\,W_i=1\right]p=\left(1+\frac{1}{2M}\right)\mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)}\right]-\frac{1}{2M}\mathbb{E}[e(X_i)\sigma_1^2(X_i)]+o(1).$$

The analogous result holds conditioning on $W_i=0$, therefore

$$\begin{aligned}\mathbb{E}\left[\left(1+\frac{K_M(i)}{M}\right)^2\sigma_{W_i}^2(X_i)\right]&=\left(1+\frac{1}{2M}\right)\mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)}+\frac{\sigma_0^2(X_i)}{1-e(X_i)}\right]\\&-\frac{1}{2M}\mathbb{E}[e(X_i)\sigma_1^2(X_i)+(1-e(X_i))\sigma_0^2(X_i)]+o(1).\end{aligned}$$

As a result,

$$N\mathrm{Var}(\hat{\tau}_M^{sm})=\left(1+\frac{1}{2M}\right)\mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)}+\frac{\sigma_0^2(X_i)}{1-e(X_i)}\right]+\mathrm{Var}(\tau(X_i))-\frac{1}{2M}\mathrm{Var}(\varepsilon)+o(1).$$

□

Before proving Theorem 5 we state two auxiliary lemmas. Let λ be a multi-index of dimension k, that is, an k-dimensional vector of non-negative integers, with $|\lambda|=\sum_{i=1}^k\lambda_i$, and let Λ_l be the set of λ such that $|\lambda|=l$. Furthermore, let $x^\lambda=x_1^{\lambda_1}\dots x_k^{\lambda_k}$, and let $\partial^\lambda g(x)=\partial^{|\lambda|}g(x)/\partial x_1^{\lambda_1}\dots\partial x_k^{\lambda_k}$. Define $|g(\cdot)|_d=\max_{|\lambda\leq d|}\sup_x|\partial^\lambda g(x)|$.

Lemma A.3: (Uniform Convergence of Series Estimators of Regression Functions, Newey 1995)
Suppose the conditions in Theorem 5 hold. Then for any $\alpha>0$ and non-negative integer d,

$$|\hat{\mu}_w(\cdot)-\mu_w(\cdot)|_d=O_p\left(K^{1+2k}\left((K/N)^{1/2}+K^{-\alpha}\right)\right),$$

for $w=0,1$.

Proof: Assumptions 3.1, 4.1, 4.2 and 4.3 in Newey (1995) are satisfied (with $\mu_w(x)$ infinitely often differentiable), implying that Newey's Theorem 4.4 applies. □

LEMMA A.4: (UNIT-LEVEL BIAS CORRECTION)
Suppose the conditions in Theorem 5 hold. Then

$$\max_i |\hat{\mu}_w(X_i) - \hat{\mu}_w(X_{j_m(i)}) - \big(\mu_w(X_i) - \mu_w(X_{j_m(i)})\big)| = o_p(N^{-1/2}),$$

for $w = 0, 1$.

PROOF:
Fix the non-negative integer $L > (k-2)/2$. Let $U_{m,i} = X_{j_m(i)} - X_i$, with jth element $U_{m,i,j}$. Use a Taylor series expansion around X_i to write

$$\hat{\mu}_w(X_{j_m(i)}) = \hat{\mu}_w(X_i) + \sum_{l=1}^{L} \sum_{\lambda \in \Lambda_l} \partial^\lambda \hat{\mu}_w(X_i) U_{m,i}^\lambda + \sum_{\lambda \in \Lambda_{L+1}} \partial^\lambda \hat{\mu}(\tilde{x}, w) U_{m,i}^\lambda.$$

First consider the last sum, $\sum_{\lambda \in \Lambda_{L+1}} \partial^\lambda \hat{\mu}(\tilde{x}, w) U_{m,i}^\lambda$. By the assumptions in Theorem 5, the first factor in each term is bounded by $C^{|\lambda|} = C^{L+1}$. The second factor in each term is of the form $\prod_{j=1}^k U_{m,i,j}^{\lambda_j}$. The factor $U_{m,i}^{\lambda_j}$ is of order $O_p(N^{-\lambda_j/k})$, so that the product is of the order $O_p(N^{-\sum_{j=1}^k \lambda_j/k}) = O_p(N^{-(L+1)/k})$. All moments of $N^{1/k} U_{m,i}$ are finite, hence with $\partial \hat{\mu}_w(x)$ bounded for $|\lambda| \leq L+1$, all moments of $N^{(L+1)/k} \sum_{\lambda \in \Lambda_{L+1}} \partial^\lambda \hat{\mu}(\tilde{x}, w) U_{m,i}^\lambda$ are finite. Therefore for any $\varepsilon > 0$

$$\max_{i=1,\ldots,N} \sum_{\lambda \in \Lambda_{L+1}} \partial^\lambda \hat{\mu}(\tilde{x}, w) U_{m,i}^\lambda = o_p(N^{-(L+1)/k+\varepsilon}).$$

Because $L > (k-2)/2$, we can choose $\varepsilon > 0$ such that this is $o_p(N^{-1/2}$. Hence,

$$\max_{i=1,\ldots,N} \left(\hat{\mu}_w(X_{j_m(i)}) - \hat{\mu}_w(X_i) - \sum_{l=1}^{L} \sum_{\lambda \in \Lambda_l} \partial^\lambda \hat{\mu}_w(X_i) U_m(X_i)^\lambda \right) = o_p(N^{-1/2}). \tag{A.10}$$

Using the same argument as we used for the estimated regression function $\hat{\mu}_w(x)$ we have for the true regression function $\mu_w(x)$,

$$\max_{i=1\ldots,N} \left(\mu_w(X_{j_m(i)}) - \mu_w(X_i) = \sum_{l=1}^{L} \sum_{\lambda \in \Lambda_l} \partial^\lambda \mu_w(X_i) U_m(X_i)^\lambda \right) = o_p(N^{-1/2}). \tag{A.11}$$

Now consider the difference:

$$\hat{\mu}_w(X_{j_m(i)}) - \hat{\mu}_w(X_i) - \big(\mu_w(X_{j_m(i)}) - \mu_w(X_i)\big)$$

$$= \sum_{l=1}^{L} \sum_{\lambda \in \Lambda_l} \big(\partial^\lambda \hat{\mu}_w(X_i) - \partial^\lambda \mu_w(X_i)\big) \cdot U_m(X_i)^\lambda + o_p(N^{-1/2}).$$

Consider for a particular $\lambda \in \Lambda_l$ the term $\big(\partial^\lambda \hat{\mu}_w(X_i) - \partial^\lambda \mu_w(X_i)\big) \cdot U_m(X_i)^\lambda$. The second factor is, using the same argument as before, of order $O_p(N^{-l/k})$, Since $l \geq 1$, the second factor is at most $O_p(N^{-1/k})$, and because all the relevant moments exist $\max_i U_{m,i}^\lambda = o_p(N^{-1/k+\varepsilon})$ for any $\varepsilon > 0$. Now consider the first factor. By Lemma A.3, $|\sup(\partial^\lambda \hat{\mu}_w(x) - \partial^\lambda \mu_w(x))|$ is of order $O_p\left(K^{1+2k}\left((K/N)^{1/2} + K^{-\alpha}\right)\right)$. With $K = N^\nu$, this is $O_p\left(N^{\nu(3/2+2k)-1/2} + N^{-\alpha\nu(1+2k)}\right)$. We can choose α large enough so that for any given ν the first term dominates. Hence the order of the product is $O_p\left(N^{\nu(3/2+2k)-1/2}\right) \cdot O_p(N^{-1/k})$. By the

assumptions in Theorem 5 we have $\nu < 2/(3k+4k^2)$. Hence, for ε small enough we have $\nu(3/2+2k)-1/2 < 1/k - 1/2 + \varepsilon$, and therefore the order is $o_p(N^{-1/2})$. □

Proof of Theorem 5:
The difference $|\hat{B}^{sm} - B^{sm}|$ can be written as

$$\left|\hat{B}^{sm} - B^{sm}\right| = \left|\frac{1}{N}\left(\sum_{i|w_i=0} \hat{\mu}_1(X_i) - \hat{\mu}_1(X_{j_m(i)}) - \big(\mu_1(X_i) - \mu_1(X_{j_m(i)})\big)\right.\right.$$

$$\left.\left. + \sum_{i|w_i=1} \hat{\mu}_0(X_i) - \hat{\mu}_0(X_{j_m(i)}) - \big(\mu_0(X_i) - \mu_0(X_{j_m(i)})\big)\right)\right|$$

$$\leq \frac{1}{N}\sum_{i=1}^{N}\sum_{w=0,1} \left|\hat{\mu}_w(X_i) - \hat{\mu}_w(X_{j_m(i)}) - \big(\mu_w(X_i) - \mu_w(X_{j_m(i)})\big)\right|$$

$$\leq \max_{i=1,\dots,N} \sum_{w=0,1} \left|\hat{\mu}_w(X_i) - \hat{\mu}_w(X_{j_m(i)}) - \big(\mu_w(X_i) - \mu_w(X_{j_m(i)})\big)\right| = o_p(N^{-1/2}),$$

by Lemma A.4. □

Proof of Theorem 6:
By the triangle inequality

$$\left|\iota_N'\mathbf{A}\hat{\Omega}\mathbf{A}'\iota_N/N - V^E\right| \leq \left|\iota_N'\mathbf{A}\hat{\Omega}\mathbf{A}'\iota_N/N - \iota_N'\mathbf{A}\Omega\mathbf{A}'\iota_N/N\right| + \left|\iota_N'\mathbf{A}\Omega\mathbf{A}'\iota_N/N - V^E\right|.$$

Lemma 3 shows that

$$\iota_N'\mathbf{A}\Omega\mathbf{A}'\iota_N/N - V^E = o_p(1),$$

so we only need to show that

$$\iota_N'\mathbf{A}\hat{\Omega}\mathbf{A}'\iota_N/N - \iota_N'\mathbf{A}\Omega\mathbf{A}'\iota_N/N = o_p(1). \tag{A.12}$$

First we make two preliminary observations. The first uses the fact that matching units of one type to the nearest units of the same type is slightly different from matching to nearest units of the opposite type. One implication is that $L(i) = \sum_j 1\{l(j) = i\}$ is bounded from above (by a function of the dimension). For example, with $k = 1$, $L(i) \leq 2$: no unit can be the closest to more than two other units.
Second, the supremum of the matching discrepancies goes to zero

$$\sup_{x,w} \|X_i - X_{l(i)}\| \xrightarrow{d} 0.$$

This follows from the compactness of the covariate spaces and the fact that the densities are bounded away from zero. To see this, fix $\varepsilon > 0$. Then we can partition the covariate space into $\bar{N}$ subsets $\mathbb{X}_n$ such that $\max_{n\leq\bar{N}} \sup_{x,y\in\mathbb{X}_n} \|x - y\| < \varepsilon$. With probability approaching one, the number of observations in each subset is at least two. Hence the distance to the nearest match is less than ε.
The implication of the second observation is that

$$\sup_i \|\mathbb{E}[\hat{\sigma}^2_{W_i}(X_i)|\mathbf{X},\mathbf{W}] - \sigma^2_{W_i}(X_i)\| \xrightarrow{p} 0.$$

To see this note that the expectation can be written as

$$\mathbb{E}[\hat{\sigma}^2_{W_i}(X_i)|\mathbf{X},\mathbf{W}] = \frac{1}{2}\left(\sigma^2_{W_i}(X_i) + \sigma^2_{W_{l(i)}}(X_{l(i)}) + \left(\mu_{W_i}(X_i) - \mu_{W_{l(i)}}(X_{l(i)})\right)^2\right),$$

which, by continuity of $\sigma^2_w(x)$ and $\mu_w(x)$ in x, goes to $\sigma^2_{W_i}(X_i)$ if $\sup_{w,x}\|X_i - X_{l(i)}\|$ does.
Now, to prove (A.12), we write, using the representation in (21),

$$\left|\iota_N'\mathbf{A}\hat{\Omega}\mathbf{A}'\iota_N/N - \iota_N'\mathbf{A}\Omega\mathbf{A}'\iota_N/N\right| = \left|\frac{1}{N}\sum_{i=1}^N\left(1+\frac{K_M(i)}{M}\right)^2\cdot\left(\sigma^2_{W_i}(X_i) - \hat{\sigma}^2_{W_i}(X_i)\right)\right|$$

$$\leq \left|\frac{1}{N}\sum_{i=1}^N\left(1+\frac{K_M(i)}{M}\right)^2\cdot\left(\sigma^2_{W_i}(X_i) - \mathbb{E}[\hat{\sigma}^2_{W_i}(X_i)|\mathbf{X},\mathbf{W}]\right)\right| \tag{A.13}$$

$$+\left|\frac{1}{N}\sum_{i=1}^N\left(1+\frac{K_M(i)}{M}\right)^2\cdot\left(\mathbb{E}[\hat{\sigma}^2_{W_i}(X_i)|\mathbf{X},\mathbf{W}] - \hat{\sigma}^2_{W_i}(X_i)\right)\right| \tag{A.14}$$

For (A.13) we have:

$$\left|\frac{1}{N}\sum_{i=1}^N\left(1+\frac{K_M(i)}{M}\right)^2\cdot\left(\sigma^2_{W_i}(X_i) - \mathbb{E}[\hat{\sigma}^2_{W_i}(X_i)|\mathbf{X},\mathbf{W}]\right)\right|$$

$$\leq \max_i\left(\sigma^2_{W_i}(X_i) - \mathbb{E}[\hat{\sigma}^2_{W_i}(X_i)|\mathbf{X},\mathbf{W}]\right)\cdot\left|\frac{1}{N}\sum_{i=1}^N\left(1+\frac{K_M(i)}{M}\right)^2\right|.$$

The second factor satisfies a law of large numbers by Lemma 3(i). The first factor converges to zero, and thus the entire expression converges to zero.
To show that (A.14) converges to zero, first decompose

$$\hat{\sigma}^2_{W_i}(X_i) - \mathbb{E}[\hat{\sigma}^2_{W_i}(X_i)|\mathbf{X},\mathbf{W}]$$

$$= \frac{1}{2}\left(\varepsilon_i^2 - \sigma^2_{W_i}(X_i) + \varepsilon^2_{l(i)} - \sigma^2_{W_{l(i)}}(X_{l(i)}) - 2\varepsilon_i\cdot\varepsilon_{l(i)} + (\varepsilon_i - \varepsilon_{l(i)})\cdot(\mu_{W_i}(X_i) - \mu_{W_{l(i)}}(X_{l(i)}))\right).$$

Ignoring the terms involving $\varepsilon_i\cdot\varepsilon_{l(i)}$, and the terms linear in ε_i, we can write (A.14) as $\frac{1}{N}\sum_{i=1}^N C_i(\varepsilon_i^2 - \sigma^2_{W_i})$, where

$$C_i = \left(1+\frac{K_M(i)}{M}\right)^2 + \sum_{j|i=l(j)}\left(1+\frac{K_M(j)}{M}\right)^2 \leq \bar{L}\cdot\max_i\left(1+\frac{K_M(i)}{M}\right)^2,$$

where $\bar{L}$ is the upper bound on $L(i)$. With all moments of $K_M(i)$ existing, $N^{-\alpha}\max_{i=1}^N\left(1+\frac{K_M(i)}{M}\right)^2$ is $o_p(1)$ for all $\alpha > 0$, and therefore

$$\frac{1}{N^2}\sum_{i=1}^N C_i^2 \xrightarrow{p} 0,$$

and so the sum $\frac{1}{N}\sum_{i=1}^{N} C_i(\varepsilon_i^2 - \sigma_{W_i}^2)$ converges to zero. Similarly we can write the terms linear in ε_i as $\sum_i C_i \cdot \varepsilon_i$, with $|C_i| \leq \bar{L} \cdot \max_i \left(1 + \frac{K_M(i)}{M}\right)^2 \cdot \sup_{x,w} |\mu_w(x)|$, which therefore again satisfies $N^{-2}\sum_{i=1}^{N} C_i \xrightarrow{p} 0$, which shows convergence of the sum of these terms. Finally consider the terms involving $\varepsilon_i \cdot \varepsilon_{l(i)}$. They sum up to

$$\frac{1}{N}\sum_{i=1}^{N}\left(1 + \frac{K_M(i)}{M}\right)^2 \cdot \varepsilon_i \cdot \varepsilon_{l(i)}.$$

Take the expectation of the square of this expression. There are only $2N$ terms with non-zero expectations, and hence the sum converges to zero. □

References

Abadie, A. (2002), "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics* (forthcoming).

Angrist, J. (1998), "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica*, 66, 249-289.

Angrist, J.D., G.W. Imbens and D.B. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444-472.

Angrist, J. D. and A. B. Krueger (2000), "Empirical Strategies in Labor Economics," in A. Ashenfelter and D. Card eds. Handbook of Labor Economics, vol. 3. New York: Elsevier Science.

Ashenfelter, O., and D. Card, (1985), "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs", *Review of Economics and Statistics*, 67, 648-660.

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.

Becker, S., and A. Ichino, (2002), "Estimation of Average Treatment Effects Based on Propensity Scores," forthcoming, *The Stata Journal*

Bloom, H., C. Michalopoulos, C. Hill, and Y. Lei, (2002) "Can Nonexperimental Comparison Group Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs," Manpower Demonstration Research Corporation, June 2002.

Blundell, R., and M. Costa Dias (2002), "Alternative Approaches to Evaluation in Empirical Microeconomics," Institute for Fiscal Studies, Cemmap working paper cwp10/02.

Blundell, R., M. Costa Dias, C. Meghir., and J. van Reenen, (2001), "Evaluating the Employment Impact of a Mandatory Job Search Assistance Program", IFS Working Paper WP01/20.

Card, D., and Sullivan, (1988), "Measuring the Effect of Subsidized Training Programs on Movements In and Out of Employment", *Econometrica*, vol. 56, no. 3 497-530.

Cochran, W., (1968) "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies", *Biometrics* 24, 295-314.

Cochran, W., and D. Rubin (1973) "Controlling Bias in Observational Studies: A Review" *Sankhya*, 35, 417-46.

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.

Estes, E.M., and B.E. Honoré, (2001), "Partially Linear Regression Using One Nearest Neighbor," unpublished manuscript, Princeton University.

Frölich, M. (2000), "Treatment Evaluation: Matching versus Local Polynomial Regression," Discussion paper 2000-17, Department of Economics, University of St. Gallen.

Gu, X., and P. Rosenbaum, (1993), "Comparison of Multivariate Matching Methods: Structures, Distances and Algorithms", *Journal of Computational and Graphical Statistics*, 2, 405-20.

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., and J. Hotz, (1989), "Alternative Methods for Evaluating the Impact of Training Programs", (with discussion), *Journal of the American Statistical Association.*

Heckman, J., and R. Robb, (1984), "Alternative Methods for Evaluating the Impact of Interventions," in Heckman and Singer (eds.), *Longitudinal Analysis of Labor Market Data*, Cambridge, Cambridge University Press.

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

Heckman, J., H. Ichimura, J. Smith, and P. Todd, (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica* 66, 1017-1098.

Heckman, J.J., R.J. Lalonde, and J.A. Smith (2000), "The Economics and Econometrics of Active Labor Markets Programs," in A. Ashenfelter and D. Card eds. Handbook of Labor Economics, vol. 3. New York: Elsevier Science.

Hirano, K., G. Imbens, and G. Ridder, (2000), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," NBER Working Paper.

Hotz J., G. Imbens, and J. Mortimer (1999), "Predicting the Efficacy of Future Training Programs Using Past Experiences" NBER Working Paper.

Ichimura, H., and O. Linton, (2001), "Trick or Treat: Asymptotic Expansions for some Semiparametric Program Evaluation Estimators." unpublished manuscript, London School of Economics.

Lalonde, R.J., (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76, 604-620.

Lechner, M, (1998), "Earnings and Employment Effects of Continuous Off-the-job Training in East Germany After Unification," *Journal of Business and Economic Statistics.*

Manski, C., (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings*, 80, 319-323.

Manski, C., (1995): *Identification Problems in the Social Sciences*, Harvard University Press, Cambridge, MA.

Manski, C., G. Sandefur, S. McLanahan, and D. Powers (1992), "Alternative Estimates of the Effect of Family Structure During Adolescence on High School," *Journal of the American Statistical Association*, 87(417):25-37.

Moller, J., (1994), *Lectures on Random Voronoi Tessellations*, Springer Verlag, New York.

Newey, W.K., (1995) "Convergence Rates for Series Estimators," in G.S. Maddala, P.C.B. Phillips and T.N. Srinavasan eds. Statistical Methods of Economics and Quantitative Economics: Essays in Honor of C.R. Rao. Cambridge: Blackwell.

Okabe, A., B. Boots, K. Sugihara, and S. Nok Chiu, (2000), *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, 2nd Edition, Wiley, New York.

Olver, F.W.J., (1974), *Asymptotics and Special Functions.* Academic Press, New York.

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

Robins, J.M., and A. Rotnitzky, (1995), "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association*, 90, 122-129.

Robins, J.M., Rotnitzky, A., Zhao, L-P. (1995), "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90, 106-121.

Rosenbaum, P., (1984), "Conditional Permutation Tests and the Propensity Score in Observational Studies," *Journal of the American Statistical Association*, 79, 565-574.

ROSENBAUM, P., (1989), "Optimal Matching in Observational Studies", *Journal of the American Statistical Association*, 84, 1024-1032.

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

ROSENBAUM, P., (2000), "Covariance Adjustment in Randomized Experiments and Observational Studies," forthcoming, *Statistical Science*.

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

ROSENBAUM, P., AND D. RUBIN, (1983b), "Assessing the Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome," *Journal of the Royal Statistical Society*, Ser. B, 45, 212-218.

ROSENBAUM, P., AND D. RUBIN, (1984), "Reducing the Bias in Observational Studies Using Subclassification on the Propensity Score", *Journal of the American Statistical Association*, 79, 516-524.

ROSENBAUM, P., AND D. RUBIN, (1985), "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score", *American Statistician*, 39, 33-38.

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

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

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

RUBIN, D., (1979), "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies", *Journal of the American Statistical Association*, 74, 318-328.

SMITH, J. A. AND P. E. TODD, (2001), "Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods," *American Economic Review*, Papers and Proceedings, 91:112-118.

STOYAN, D., W. KENDALL, AND J. MECKE, (1995), *Stochastic Geometry and its Applications*, 2nd Edition, Wiley, New York.

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

YATCHEW, A., (1999), "Differencing Methods in Nonparametric Regression: Simple Techniques for the Applied Econometrician", Working Paper, Department of Economics, University of Toronto.

ZHAO, (2002) "Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics and an Application," unpublished manuscript, department of economics, Johns Hopkins University.

TABLE 1
A MATCHING ESTIMATOR WITH FOUR OBSERVATIONS

i	W_i	X_i	Y_i	$j(i)$	$K_1(i)$	$\hat{Y}_i(0)$	$\hat{Y}_i(1)$	$\hat{\tau}_i$
1	1	6	10	3	2	8	10	2
2	0	4	5	1	1	5	10	5
3	0	7	8	1	1	8	10	2
4	1	1	5	2	0	5	5	0
								$\hat{\tau} = 9/4$

TABLE 2
SUMMARY STATISTICS

	Experimental Data				PSID		T-statistic	
	Trainees (N=185)		Controls (N=260)		Controls (N=2490)		Train/	Train/
	mean	(s.d.)	mean	(s.d.)	mean	(s.d.)	Contr	PSID
Panel A: Pretreatment Variables								
Age	25.8	(7.16)	25.05	(7.06)	34.85	(10.44)	[1.1]	[-16.0]
Education	10.4	(2.01)	10.09	(1.61)	12.12	(3.08)	[1.4]	[-11.1]
Black	0.84	(0.36)	0.83	(0.38)	0.25	(0.43)	[0.5]	[21.0]
Hispanic	0.06	(0.24)	0.11	(0.31)	0.03	(0.18)	[-1.9]	[1.5]
Married	0.19	(0.39)	0.15	(0.36)	0.87	(0.34)	[1.0]	[-22.8]
Earnings '74	2.10	(4.89)	2.11	(5.69)	19.43	(13.41)	[-0.0]	[-38.6]
Unempl. '74	0.71	(0.46)	0.75	(0.43)	0.09	(0.28)	[-1.0]	[18.3]
Earnings '75	1.53	(3.22)	1.27	(3.10)	19.06	(13.60)	[0.9]	[-48.6]
Unempl. '75	0.60	(0.49)	0.68	(0.47)	0.10	(0.30)	[-1.8]	[13.8]
Panel B: Outcomes								
Earnings '78	6.35	(7.87)	4.55	(5.48)	21.55	(15.56)	[2.7]	[-23.1]
Unempl. '78	0.24	(0.43)	0.35	(0.48)	0.11	(0.32)	[-2.7]	[4.0]

Note: The first two columns give the average and standard deviation of the 185 trainees from the experimental data set. The second pair of columns give the average and standard deviation of the 260 controls from the experimental data set. The third pair of columns give the averages and standard deviations of the 2490 controls from the nonexperimental PSID sample. The seventh column gives t-statistics for the difference between the averages for the experimental trainees and controls. The last column gives the t-statistics for the differences between the averages for the experimental trainees and the PSID controls. The last two variables, earnings '78 and unemployed '78 are post-training. All the others are pre-training variables. Earnings data are in thousands dollars.

TABLE 3
EXPERIMENTAL AND NON-EXPERIMENTAL ESTIMATES
OF AVERAGE TREATMENT EFFECTS FOR LALONDE DATA

	$M=1$		$M=4$		$M=16$		$M=64$		All Controls	
	mean	(s.d.)	est	(s.e.)	est	(s.e.)	est	(s.e.)	est	(s.e.)
Panel A: Experimental Estimates										
simple matching	1.23	(0.89)	1.99	(0.79)	1.76	(0.80)	2.20	(0.75)	1.79	(0.75)
bias-adjusted matching	1.17	(0.89)	1.84	(0.79)	1.55	(0.80)	1.74	(0.75)	1.72	(0.75)
Regression Estimates										
dif	1.79	(0.67)								
linear	1.72	(0.65)								
quadratic	2.27	(0.73)								
Panel B: Non-experimental Estimates										
simple matching	2.09	(1.00)	1.62	(0.88)	0.47	(0.86)	-0.11	(0.75)	-15.20	(0.62)
bias-adjusted matching	2.45	(1.00)	2.51	(0.88)	2.48	(0.86)	2.26	(0.75)	0.84	(0.62)
Regression Estimates										
dif	-15.20	(0.66)								
linear	0.84	(0.86)								
quadratic	3.26	(0.98)								

Note: Panel A reports the results for the experimental data (experimental controls and trainees), and Panel B the results for the nonexperimental data (PSID controls with experimental trainees). In each panel the top part reports results for the matching estimators, with the number of matches equal to 1, 4, 16, 64 and 2490 (all controls). The second part reports results for three regression adjustment estimates, based on no covariates, all covariates entering linearly and all covariates entering with a fully set of quadratic terms and interactions. The outcome is earnings in 1978 in thousands dollars.

TABLE 4
MEAN COVARIATE DIFFERENCES IN MATCHED GROUPS

	Average		$M = 1$		$M = 4$		$M = 16$		$M = 64$		$M = 2490$	
	PSID	Trainees	dif	(s.d.)	dif	(s.d.)	dif	(s.d.)	dif	(s.d.)	dif	(s.d.)
Age	0.06	-0.80	-0.02	(0.65)	-0.06	(0.60)	-0.30	(0.41)	-0.57	(0.57)	-0.86	(0.68)
Education	0.04	-0.54	-0.10	(0.44)	-0.20	(0.48)	-0.25	(0.39)	-0.24	(0.42)	-0.58	(0.66)
Black	-0.09	1.21	-0.00	(0.00)	0.09	(0.32)	0.35	(0.47)	0.70	(0.66)	1.30	(0.80)
Hispanic	-0.01	0.14	-0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.01	(0.03)	0.15	(1.30)
Married	0.12	-1.64	0.00	(0.00)	-0.06	(0.30)	-0.33	(0.46)	-0.90	(0.85)	-1.76	(1.02)
Earnings '74	0.09	-1.18	-0.01	(0.10)	-0.01	(0.12)	-0.05	(0.17)	-0.15	(0.30)	-1.26	(0.36)
Unempl '74	-0.13	1.72	0.00	(0.00)	0.02	(0.17)	0.24	(0.40)	0.41	(0.72)	1.85	(1.36)
Earnings '75	0.09	-1.18	-0.04	(0.17)	-0.07	(0.15)	-0.11	(0.19)	-0.19	(0.26)	-1.26	(0.23)
Unempl '75	-0.10	1.36	0.00	(0.00)	0.00	(0.05)	0.03	(0.28)	0.10	(0.41)	1.46	(1.44)
Log Odds Prop Score	-7.08	1.08	0.21	(0.99)	0.56	(1.13)	1.70	(1.14)	3.20	(1.49)	8.16	(2.13)

Note: In this table all covariates have been normalized to have mean zero and unit variance. The first two columns present the averages for the experimental trainees and the PSID controls. The remaining pairs of columns present the average difference within the matched pairs and the standard deviation of this difference for matching based on 1, 4, 16, 64 and 2490 matches. For the last variable the the logarithm of the odds ratio of the propensity score is used. This log odds ratio has mean -6.52 and standard deviation 3.30 in the sample.

TABLE 5
SIMULATION RESULTS

M	Estimator	mean bias	median bias	rmse	mae	s.d.	mean s.e.	coverage (nom. 95%)	coverage (nom. 90%)
1	simple matching	-0.49	-0.45	0.87	0.55	0.72	0.84	0.93	0.88
	bias-adjusted	0.04	0.06	0.74	0.47	0.74	0.84	0.96	0.92
4	simple matching	-0.85	-0.84	1.03	0.84	0.58	0.59	0.72	0.60
	bias-adjusted	0.05	0.06	0.60	0.39	0.60	0.59	0.94	0.89
16	simple matching	-1.80	-1.78	1.89	1.78	0.57	0.52	0.07	0.04
	bias-adjusted	0.17	0.16	0.62	0.40	0.60	0.52	0.90	0.83
64	simple matching	-3.27	-3.25	3.32	3.25	0.59	0.52	0.00	0.00
	bias-adjusted	0.15	0.16	0.65	0.43	0.63	0.52	0.87	0.81
All (2490)	simple matching	-19.06	-19.06	19.07	19.06	0.61	0.43	0.00	0.00
	bias-adjusted	-2.04	-2.04	2.28	2.04	1.00	0.37	0.09	0.07
difference		-19.06	-19.06	19.07	19.06	0.61	0.61	0.00	0.00
linear regression		-2.04	-2.04	2.28	2.04	1.00	0.98	0.44	0.33
quadratic regression		2.70	2.64	3.02	2.64	1.34	1.24	0.40	0.27

Note: For each estimator summary statistics are provided for 10,000 replications of the data set. Results are reported for five values of the number of matches ($M = 1, 4, 16, 64, 2490$), and for two estimators: the simple matching estimator, the bias-adjusted matching estimator, based on the regression of only the matched treated and controls. The last three rows report results for the simple average treatment-control difference, the ordinary least squares estimator, and the ordinary least square estimator using a full set of quadratic terms and interactions. For each estimator we report the mean and median bias, the root-mean-squared-error, the median-absolute-error, the standard deviation of the estimators, the average estimate of the standard error, and the coverage rate of the nominal 95% and 90% confidence intervals.